Chapter 1

INTRODUCTION

The *Handbook of Psychological Assessment* is designed to develop a high level of practitioner competence by providing relevant, practical research, and theoretical information. It can serve as both a reference and an instructional guide. As a reference book, it aids in test selection and the development of a large number and variety of interpretive hypotheses. As an instructional text, it provides students with the basic tools for conducting an integrated psychological assessment. The significant and overriding emphasis in this book is on assessing areas that are of practical use in evaluating individuals in a clinical context. It is applied in its orientation, and, for the most part, theoretical discussion has been kept to a minimum. Many books written on psychological testing and the courses organized around these books focus primarily on test theory, with a brief overview of a large number of tests. In contrast, the intent of this book is to focus on the actual processes that practitioners go through during assessment. We begin with such issues as role clarification and evaluation of the referral question and end with treatment planning and the actual preparation of the report itself.

One of the crucial skills that we hope readers of this text will develop, or at least have enhanced, is a realistic appreciation of the assets and limitations of assessment. This includes an appraisal of psychological assessment as a general strategy as well as an awareness of the assets and limitations of specific instruments and procedures. A primary limitation of assessment lies in the incorrect handling of the data, which are not integrated in the context of other sources of information (behavioral observations, history, other test scores). Also, the results are not presented in a way that helps solve the unique problems clients or referral sources are confronting. To counter these limitations, the text continually provides practitioners with guidelines for integrating and presenting the data in as useful a manner as possible. The text is thus not only a book on test interpretation (although this is an important component) but on test integration within the wider context of assessment. As a result, psychologists should be able to create reports that are accurate, effective, concise, and highly valued by the persons who receive them.

ORGANIZATION OF THE HANDBOOK

The central organizational plan for the *Handbook of Psychological Assessment* replicates the sequence practitioners follow when performing an evaluation. They are initially concerned with clarifying their roles, ensuring that they understand all the implications of the referral question, deciding which procedures would be most appropriate for the assessment, and reminding themselves of the potential problems associated with clinical judgment (this chapter). They also need to understand the context in which they will conduct the assessment. This understanding includes appreciating the issues, concerns, terminology, and likely roles of the persons from these contexts. Practitioners also must follow clear ethical guidelines, know how to work with persons from diverse backgrounds, and recognize issues related to computer-assisted assessment and the ways that the preceding factors might influence their selection of procedures (see Chapter 2).

Once practitioners have fully understood the preliminary issues discussed in this chapter and Chapter 2, they must select different strategies of assessment. The three major strategies are interviewing, observing behavior, and psychological testing. An interview is likely to occur during the initial phases of assessment and is also essential in interpreting test scores and understanding behavioral observations (see Chapter 3). The assessment of actual behaviors might also be undertaken (see Chapter 4). Behavioral assessment might be either an end in itself or an adjunct to testing. It might involve a variety of strategies, such as the measurement of overt behaviors, cognitions, alterations in physiology, or relevant measures from self-report inventories.

The middle part of the book (Chapters 5 through 13) provides a general overview of the most frequently used tests. Each chapter begins with an introduction to the test in the form of a discussion of its history and development, current evaluation, and procedures for administration, as well as use with diverse populations. The main portions of these chapters provide a guide for interpretation, which includes such areas as the meaning of different scales, significant relations between scales, frequent trends, and the meaning of unusually high or low scores. When appropriate, there are additional subsections. For example, Chapter 5, "Wechsler Intelligence Scales," includes additional sections on the meaning of IQ scores, estimating premorbid IQ, and assessing special populations. Likewise, several chapters include alternative procedures for using the tests, such as Chapter 7, "Minnesota Multiphasic Personality Inventory," which includes procedures for both the MMPI-2 and the MMPI-2-RF, and Chapter 11, "The Rorschach," which includes both the Comprehensive System and the R-PAS versions of the Rorschach. Chapter 12, "Screening for Neuropsychological Impairment," varies somewhat from the preceding format in that it is more a compendium and interpretive guide to some of the most frequently used short neuropsychological tests. It also includes a section on special considerations in conducting a neuropsychological interview. This organization reflects the current emphasis on and strategies for assessing patients with possible neuropsychological dysfunction.

Several of the chapters on psychological tests are quite long, particularly those for the Wechsler intelligence scales, the Minnesota Multiphasic Personality Inventory, and the Rorschach. These chapters include extensive summaries of a wide variety of interpretive hypotheses intended for reference purposes when practitioners must generate interpretive hypotheses based on specific test scores. To gain initial familiarity with the tests, we recommend that practitioners or students carefully read the initial sections (history and development, psychometric properties, etc.) and then skim through the interpretation sections more quickly. Doing this provides the reader with a basic familiarity with the procedures and types of data obtainable from the tests. As practical test work progresses, clinicians can then study the interpretive hypotheses in greater depth and gradually develop more extensive knowledge of the scales and their interpretation.

Based primarily on current frequency of use, these tests are covered in this text: Wechsler intelligence scales (WAIS-IV/WISC-V), Wechsler Memory Scales (WMS-IV), Minnesota Multiphasic Personality Inventory (MMPI-2 and MMPI-2-RF), Millon Clinical Multiaxial Inventory (MCMI-IV), Personality Assessment Inventory (PAI), NEO Personality Inventory–3 (NEO-PI-3), Bender Visual Motor Gestalt Test–II, Repeatable Battery for the Assessment of Neuropsychological Status (RBANS), and the Rorschach (Comprehensive System and R-PAS; Camara, Nathan, & Puente, 2000; C. Piotrowski & Zalewski, 1993; Rabin, Barr, & Burton, 2005; Watkins, 1991; Watkins, Campbell, Nieberding, & Hallmark, 1995). The NEO-PI-3 was selected because of the importance of including a broad-based inventory of normal functioning, along with its excellent technical development and relatively large research base. We have also included Chapter 13 focusing on the most frequently used brief, symptom-focused inventories because of the increasing importance of monitoring treatment progress and outcome in a cost- and time-efficient managed care environment (Eisman et al., 2000; C. Piotrowski, 1999). The preceding instruments represent the core assessment devices used by most practitioners.

Finally, the clinician must generate relevant treatment recommendations and integrate the assessment results into a psychological report. Chapter 14 provides a systematic approach for working with assessment results to develop practical, evidence-based treatment recommendations. Chapter 15 presents guidelines for report writing, a report format, and four sample reports representative of the four most common types of referral settings: medical setting, legal context, educational context, and psychological clinic. Thus, the chapters follow a logical sequence and provide useful, concise, and practical knowledge.

ROLE OF THE CLINICIAN

The central role of clinicians conducting assessments should be to answer specific questions and make clear, specific, and reasonable recommendations to help improve functioning. To fulfill this role, clinicians must integrate a wide range of data and bring into focus diverse areas of knowledge. Thus, they are not merely administering and scoring tests. A useful distinction to highlight this point is the contrast between a psychometrist and a clinician conducting psychological assessment (Maloney & Ward, 1976; Matarazzo, 1990). Psychometrists tend to use tests merely to obtain data, and their task is often perceived as emphasizing the clerical and technical aspects of testing. Their approach is primarily data oriented, and the end product is often a series of traits or ability descriptions. These descriptions are typically unrelated to the person's overall context and do not address unique problems the person may be facing. In contrast, psychological assessment attempts to evaluate an individual in a problem situation so that the information derived from the assessment can somehow help with the problem. Tests are only one method of gathering data, and the test scores are not end products but merely means of generating hypotheses. Psychological assessment, then, places data in a wide perspective, with its focus being problem solving and decision making.

The distinction between psychometric testing and psychological assessment can be better understood and the ideal role of the clinician more clearly defined by briefly elaborating on the historical and methodological reasons for the development of the psychometric approach. When psychological tests were originally developed, group measurements of intelligence met with early and noteworthy success, especially in military and industrial settings where individual interviewing and case histories were too expensive and time consuming. An advantage of the data-oriented intelligence tests was that they appeared to be objective, which would reduce possible interviewer bias. More important, they were quite successful in producing a relatively high number of true positives when used for classification purposes. Their predictions were generally accurate and usable. However, these facts created the early expectation that all assessments could be performed using the same method and would provide a similar level of accuracy and usefulness. Later assessment strategies often tried to imitate the methods of earlier intelligence tests for variables such as personality and psychiatric diagnosis.

A further development consistent with the psychometric approach was the strategy of using a "test battery." It was reasoned that if a single test could produce accurate descriptions of an ability or trait, administering a series of tests could create a total picture of the person. The goal, then, was to develop a global yet definitive description for the person using purely objective methods. This goal encouraged the idea that the tool (psychological test) was the best process for achieving the goal, rather than being merely one technique in the overall assessment procedure. Behind this approach were the concepts of *individual differences* and *trait psychology*. These concepts assume that one of the best ways to describe the differences among individuals is to measure their strengths and weaknesses with respect to various traits. Thus, the clearest approach to the study of personality involved developing a relevant taxonomy of traits and then creating tests to measure those traits. Again, there was an emphasis on the tools as primary, with a deemphasis on the input of the clinician. These trends created a bias toward administration and clerical skills. In this context, the psychometrist requires little, if any, clinical expertise other than administering, scoring, and interpreting tests. According to such a view, the most preferred tests would be highly standardized and ideally machine-scored so that the normed scores, rather than the psychometrist, provide the interpretation.

The objective psychometric approach is most appropriately applicable to ability tests such as those measuring intelligence or mechanical skills. Its usefulness decreases, however, when users attempt to assess personality traits such as dependence, authoritarianism, or anxiety. Personality variables are far more complex and, therefore, need to be validated in the context of history, behavioral observations, and interpersonal relationships. For example, a moderately elevated score on a scale measuring high energy level takes on an entirely different meaning for a high-functioning physician than for an individual with a history of mood disorders and associated work and interpersonal difficulties. When the purely objective psychometric approach is used for the evaluation of problems in living (coping more effectively, resolving interpersonal relationships, etc.), its usefulness is questionable. Scores need to be connected to each other and to the context in which they emerge.

Psychological assessment is most useful in the understanding and evaluation of personality and in elucidating the likely underlying causes of problems in living. These issues involve a particular problem situation having to do with a specific individual. The central role of the clinician performing psychological assessment is that of an expert in human behavior who must deal with complex processes and understand test scores in the context of a person's life. The clinician must have knowledge concerning problem areas and, on the basis of this knowledge, form a general idea regarding behaviors to observe and areas in which to collect relevant data. Doing this involves an awareness and appreciation of multiple causation, interactional influences, and multiple relationships. As Woody (1980) stated, "Clinical assessment is individually oriented, but it always considers social existence; the objective is usually to help the person solve problems."

In addition to an awareness of the role suggested by psychological assessment, clinicians should be familiar with core knowledge related to measurement and clinical practice. This includes descriptive statistics, reliability (and measurement error), validity (and the meaning of test scores), normative interpretation, selection of appropriate tests, administration procedures, variables related to diversity (ethnicity, race, age, gender, culture, etc.), testing individuals with disabilities, and an appropriate amount of supervised experience (Turner, DeMers, Fox, & Reed, 2001). Persons performing psychological assessment should also have basic knowledge related to the demands, types of referral questions, and expectations of various contexts—particularly employment, education, vocational/career, health care (psychological, psychiatric, medical), and forensic. Furthermore, clinicians should know the main interpretive hypotheses in psychological testing and be able to identify, sift through, and evaluate a series of hypotheses to determine which are most relevant and accurate. Rather than merely knowing the labels and definitions for various types of anxiety or thought disorders, for example, clinicians should also have in-depth operational criteria for them. As another example, the concept of intelligence, as represented by the IQ score, can sometimes appear misleadingly straightforward. Intelligence test scores can be complex, though, involving a variety of cognitive abilities, the influence of cultural factors, varying performance under different conditions, and issues related to the nature of intelligence. Unless clinicians are familiar with these areas, they are not adequately prepared to handle IQ data.

The above knowledge should be integrated with relevant general coursework, including abnormal psychology, the psychology of adjustment, theories of personality, clinical neuropsychology, psychotherapy, and basic case management. A problem in many training programs is that, although students frequently have knowledge of abnormal psychology, personality theory, and test construction, they usually have insufficient training to integrate their knowledge into the interpretation of test results. Their training focuses on developing competency in administration and scoring rather than on knowledge relating to what they are testing.

The approach in this book is consistent with that of psychological assessment: Clinicians should be not only knowledgeable about traditional content areas in psychology and the various contexts of assessment but also able to integrate the test data into a relevant description of the person. This description, although focusing on the individual, should take into account the complexity of his or her social environment, personal history, and behavioral observations. Yet the goal is not merely to describe the person but rather to develop relevant answers to specific questions and present clear, specific, and reasonable recommendations that aid in problem solving and facilitate decision making.

PATTERNS OF TEST USAGE IN CLINICAL ASSESSMENT

Psychological assessment is crucial to the definition, training, and practice of professional psychology. Although the data are old, Watkins et al. (1995) found that fully 91% of all practicing psychologists engage in assessment, and 64% of all nonacademic advertisements listed assessment as an important prerequisite (Kinder, 1994). Assessment skills are also strong prerequisites for internships and postdoctoral training. The theory and instruments of assessment can be considered the very foundation of clinical investigation, applied research, and program evaluation. In many ways, psychological assessment is professional psychology's unique contribution to the wider arena of clinical practice. The early professional psychologists even defined themselves largely in the context of their role as psychological testers. Practicing psychologists spend 10% to 25% of their time conducting psychological assessment (Camara et al., 2000; Watkins, 1991; Watkins et al., 1995).

Although assessment has always been a core, defining feature of professional psychology, the patterns of use and relative importance of assessment have changed with time. During the 1940s and 1950s, psychological testing was frequently the single most important activity of professional psychologists. In contrast, the past 60 years have seen psychologists become involved in a far wider diversity of activities. Lubin and his colleagues (Lubin, Larsen, & Matarazzo, 1984; Lubin, Larsen, Matarazzo, & Seever, 1985, 1986) found that the average time spent performing assessment across five treatment settings was 44% in 1959, 29% in 1969, and only 22% in 1982. The average time spent in 1982 performing assessments in the five different settings ranged from 14% in counseling centers to 31% in psychiatric hospitals (Lubin et al., 1984, 1985, 1986). Camara et al. (2000) found that the vast majority of professional psychologists (81%) spend 0 to 4 hours a week conducting formal assessment, 15% spend 5 to 20 hours a week, and 4% spend more than 20 hours. It is expected that over the last 20 years, the time spent doing assessment has likely decreased even further. The gradual decrease in the total time spent in assessment is due in part to the widening role of psychologists. Whereas in the 1940s and 1950s a practicing psychologist was almost synonymous with a tester, professional psychologists currently are increasingly involved in administration, consultation, organizational development, and many areas of direct treatment (Bamgbose, Smith, Jesse, & Groth-Marnat, 1980; Groth-Marnat, 1988; Groth-Marnat & Edkins, 1996). Decline in testing has also been attributed to disillusionment with the testing process based on criticisms about the reliability and validity of many assessment devices (Garb, Wood, Nezworski, Grove, & Stejskal, 2001; Wood, Lilienfeld, Garb, & Nezworski, 2000; Ziskin & Faust, 2008) and reductions in reimbursement (Cashel, 2002). In addition, psychological assessment has come to include a wide variety of activities beyond merely the administration and interpretation of traditional tests. These include conducting structured and unstructured interviews, behavioral observations in natural settings, observations of interpersonal interactions, neuropsychological assessment, behavioral assessment, and using assessment findings as part of the overall therapeutic process (Finn, 2007; Garb, 2007).

The relative popularity of different traditional psychological tests has been surveyed since 1935 in many settings, such as academic institutions, psychiatric hospitals, counseling centers, Veterans Administration centers, institutions for those with developmental disabilities, private practice, and various memberships and professional organizations. Surveys (somewhat dated) of test usage have usually found that the 10 most frequently used tests are the Wechsler intelligence scales, Minnesota Multiphasic Personality Inventory, Rorschach, Bender Visual Motor Gestalt Test, Thematic Apperception Test, projective drawings (Human Figure Drawing, House-Tree-Person), Wechsler Memory Scale, Beck Depression Inventory, Millon Clinical Multiaxial Inventories, and California Psychological Inventory (Camara et al., 2000; Kamphaus, Petoskey, & Rowe, 2000; Lubin et al., 1985; C. Piotrowski & Zalewski, 1993; Watkins, 1991; Watkins et al., 1995). The pattern for the 10 most popular tests has remained quite stable since 1969, except that the ranking of Human Figure Drawings dropped (Camara et al., 2000). It is expected that some newer measures, especially the Personality Assessment Inventory, would be ranked quite highly in use. However, no recent surveys of test usage have been published. The pattern of test usage varies somewhat across different studies and varies considerably from setting to setting. Schools and centers for those with intellectual disabilities emphasize tests of intellectual abilities, such as the WISC-V and behavior rating scales; counseling centers are more likely to use vocational interest inventories; and psychiatric settings emphasize tests assessing level of pathology, such as the MMPI or MCMI.

One clear change in testing practices has been a relative decrease in the use and status of projective techniques (Groth-Marnat, 2000b; C. Piotrowski, 1999). Criticisms have been wide ranging but have centered on overly complex scoring systems, questionable norms, subjectivity of scoring, poor predictive utility, and inadequate or even nonexistent validity (Garb, 2005a; Garb et al., 2001; D. N. Miller, 2007; Pruitt, Smith, Thelen, & Lubin, 1985; D. Smith & Dumont, 1995). Further criticisms include the extensive time required to effectively learn the techniques, heavy reliance of projective techniques on psychoanalytic theory, and the greater time and cost efficiency of alternative objective tests. These criticisms have usually occurred from within the academic community, where the techniques are used less and less for research purposes (C. Piotrowski, 1999; C. Piotrowski & Zalewski, 1993; Watkins, 1991). As a result of these criticisms, there has been a slight but still noteworthy reduction in the use of the standard projective tests in professional practice (Archer, Buffington-Vollum, Stredny, & Handel, 2006; Camara et al., 2000; Kamphaus et al., 2000; C. Piotrowski, 1999). Although there has been a reduction, the Rorschach and Thematic Apperception Test (TAT) continue to have a strong foothold in clinical practice. This can be attributed to lack of time available for practitioners to learn new techniques, expectations that students in internships know how to use them, unavailability of other practical alternatives, and the fact that practitioners usually give more weight to clinical experience than to empirical evidence. This suggests distance between the quantitative, theoretical world of the academic and the practical, problem-oriented world of the practitioner. In fact, assessment practices in many professional settings seem to have little relationship to the number of research studies done on assessment tools, attitudes by academic faculty, or the psychometric quality of the test (Garb, Wood, Lilienfeld, & Nezworski, 2002). In contrast to the continued use of projective instruments in adult clinical settings, psychologists in child settings are likely to rely more on behavior rating scales (e.g., Child Behavior Checklist) than projective tests (Cashel, 2002; Kamphaus et al., 2000; D. N. Miller, 2007).

The earliest form of assessment was through clinical interview. Clinicians like Freud, Jung, and Adler used unstructured interaction to obtain information regarding history, diagnosis, and underlying structure of personality. Later clinicians organized interviews using outlines of the areas that should be discussed. During the 1960s and 1970s, much criticism was directed toward the interview, leading many psychologists to perceive interviews as unreliable and lacking empirical validation. Tests, in many ways, were designed to counter the subjectivity and bias of interview techniques. During the 1980s and 1990s, a wide variety of structured interview techniques gained popularity and have often been found to be reliable and valid indicators of a client's level of functioning. Structured interviews such as the Diagnostic Interview Schedule (DIS; Robins, Helzer, Cottler, & Goldring, 1989), Structured Clinical Interview for the DSM (SCID; Spitzer, Williams, & Gibbon, 1987), and Renard Diagnostic Interview (Helzer, Robins, Croughan, & Welner, 1981) are often given preference over psychological tests. These interviews, however, are very different from the traditional unstructured approaches. They have the advantage of being psychometrically sound even though they might lack important elements of rapport, idiographic richness, and flexibility that characterize less structured interactions (Garb, 2007; R. Rogers, 2001).

A further trend has been the development of neuropsychological assessment (see Groth-Marnat, 2000a; Lezak, Howieson, Bigler, & Tranel, 2012). The discipline is a synthesis between behavioral neurology and psychometrics and was created from a need to answer questions such as the nature of a person's organic deficits, severity of deficits, localization, and differentiating between functional versus organic impairment. The pathognomonic sign approach and the psychometric approaches are two clear traditions that have developed in the discipline. Clinicians relying primarily on a pathognomonic sign approach are more likely to interpret specific behaviors such as perseverations or weaknesses on one side of the body, which are highly indicative of the presence and nature of organic impairments. These clinicians tend to rely on the tradition of assessment associated with Luria (Bauer, 2000; Luria, 1973) and base their interview design and tests on a flexible method of testing possible hypotheses for different types of impairment. In contrast, the more quantitative tradition represented by Reitan and his colleagues (Reitan & Wolfson, 1993; Russell, 2000) is more likely to rely on critical cutoff scores, which distinguish between normal persons and those with brain damage. Reitan and Wolfson (1985, 1993) have recommended using an impairment index, which is the proportion of brain-sensitive tests that fall into the brain-damaged range. In actual practice, most clinical neuropsychologists are more likely to combine the psychometric and pathognomonic sign approaches (Rabin, Barr, & Burton, 2005). The two major neuropsychological test batteries are the Luria-Nebraska Neuropsychological Battery (Golden, Purisch, & Hammeke, 1985) and the Halstead Reitan Neuropsychological Test Battery (Reitan & Wolfson, 1993). A typical neuropsychological battery might include tests specifically designed to assess organic impairment along with tests such as the MMPI, Wechsler intelligence scales, and the Wide Range Achievement Test (WRAT-4). As a result, extensive research over the past 15 to 20 years has been directed toward developing a greater understanding of how the older and more traditional tests relate to different types and levels of cerebral dysfunction.

During the 1960s and 1970s, behavior therapy was increasingly used and accepted. Initially, behavior therapists were concerned with an idiographic approach to the functional analysis of behavior. As their techniques became more sophisticated, formalized methods of behavioral assessment began to arise. These techniques arose in part from dissatisfaction with the methods of diagnosis of the second edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-II*; American Psychiatric Association, 1968), as well as from a need to have assessment relate more directly to treatment and its outcomes. There was also a desire to be more accountable for documenting behavior change over time. For example, if behaviors related to anxiety decreased after therapy, the therapist should be able to demonstrate that the treatment had been successful. Behavioral assessment could involve measurements of movements (behavioral checklists, behavioral analysis), physiological responses (galvanic skin response [GSR], electromyograph [EMG]) or self-reports (self-monitoring, Symptom Checklist-90-R, assertiveness scales). Whereas the early behavioral assessment techniques showed little concern with the psychometric properties of their instruments, there has been an increasing push to have them meet adequate levels of reliability and validity (First, Frances, Widiger, Pincus, & Davis, 1992; Follette & Hayes, 1992). Despite the many formalized techniques of behavioral assessment, many behavior therapists feel that an unstructured, idiographic approach is most appropriate.

Traditional means of assessment, then, have decreased because of an overall increase in other activities of psychologists and an expansion in the definition of assessment. Currently, a psychologist doing assessment might include such techniques as interviewing, administering, and interpreting traditional psychological tests (MMPI-2/MMPI-A/MMPI-2-RF, WAIS-IV, etc.), naturalistic observations, neuropsychological assessment, and behavioral assessment. In addition, professional psychologists might be required to assess areas that were not given much emphasis before the 1980s: personality disorders (borderline personality, narcissism), stress and coping (life changes, burnout, existing coping resources), hypnotic responsiveness, psychological health, adaptation to new cultures, changes associated with increasing modernization, and strengths (related to positive psychology movements). Additional areas might include family systems interactions, relation between a person and his or her environment (social climate, social supports), cognitive processes related to behavior disorders, and level of personal control and self-efficacy. All these require clinicians to be continually aware of new and more specific assessment devices and to maintain flexibility in the approaches they take.

The future of psychological assessment will probably be most influenced by the trends toward computerized assessment, adaptation to managed health care, and distance health care delivery (Groth-Marnat, 2000b, 2009; Kay, 2007). Computerized assessment is likely to enhance efficiency through rapid scoring, complex decision rules, reduction in client–practitioner contact, novel presentation of stimuli (i.e., virtual reality), and generation of interpretive hypotheses (Lichtenberger, 2006). Future assessments are also likely to tailor the presentation of items based on the client's previous responses (Forbey & Ben-Porath, 2007). Unnecessary items will not be given, with one result being that a larger amount of information will be obtained through the presentation of relatively fewer items. This time efficiency is in part stimulated by the cost-savings policies of managed care, which require psychologists to demonstrate the cost-effectiveness of their services (Groth-Marnat, 1999; Groth-Marnat & Edkins, 1996). In assessment, this means linking assessment with treatment planning. Thus, psychological reports of the future are likely to need to link client dynamics directly to recommendations and treatment options. Whereas considerable evidence supports the cost-effectiveness of using psychological tests in organizational contexts, health care needs to demonstrate that assessment can increase the speed of treatment as well as optimize treatment outcome (Blount et al., 2007; Groth-Marnat, 1999; Groth-Marnat, Roberts, & Beutler, 2001; Lambert & Hawkins, 2004; Yates & Taub, 2003).

A further challenge and area for development is the role distance health will play in assessment (Leigh & Zaylor, 2000; M. J. Murphy, Levant, Hall, & Glueckauf, 2007). Distance assessment as a means in and of itself is likely to become important. Professional psychologists may be required to change their traditional face-to-face role to one of developing and monitoring new applications as well as consulting/collaborating with clients regarding the results of assessments derived from the computer.

EVALUATING PSYCHOLOGICAL TESTS

Before using a psychological test, clinicians should investigate and understand the theoretical orientation of the test, practical considerations, the appropriateness of the standardization sample, and the adequacy of its psychometric properties (reliability and validity). Often, helpful descriptions and reviews that relate to these issues can be found in the test manuals as well as past and future editions of the *Mental Measurements Yearbook* (Carlson, Geisinger, & Jonson, 2014); *Tests in Print* (L. L. Murphy, Geisinger, Carlson, & Spies, 2011); *Tests: A Comprehensive Reference for Assessment in Psychology, Education, and Business* (Maddox, 2003); and *Measures for Clinical Practice: A Sourcebook* (Fischer & Corcoran, 2007). Reviews can also be found in assessment-related journals, such as the *Journal of Personality Assessment,* the *Journal of Psychoeducational Assessment*, and *Educational and Psychological Measurement*. Table 1.1 outlines the more important questions that should be answered. Each issue outlined in this table is discussed further. The discussion reflects a practical focus on problems that clinicians using psychological tests are likely to confront. It is not intended to provide a comprehensive coverage of test theory and construction; if a more detailed treatment is required, the reader is referred to one of the many texts on psychological testing (e.g., Aiken & Groth-Marnat, 2006; R. M. Kaplan & Saccuzzo, 2005).

Theoretical Orientation

Before clinicians can effectively evaluate whether a test is appropriate, they must understand its theoretical orientation. Clinicians should research the construct that the test is supposed to measure and then examine how the test approaches this construct. This information can usually be found in the test manual. If for any reason the information in the manual is insufficient, clinicians should seek it elsewhere. Clinicians can often obtain additional useful information regarding the construct being measured by carefully studying the individual test items. Usually the manual provides an individual analysis of the items, which can help the potential test user evaluate whether they are relevant to the trait being measured.

Practical Considerations

A number of practical issues relate more to the context and manner in which the test is used than to its construction. First, tests vary in terms of the level of education

Table 1.1 Evaluating a Psychological Test

Theoretical Orientation

1. Do you adequately understand the theoretical construct the test is supposed to be measuring?
2. Do the test items correspond to the theoretical description of the construct?

Practical Considerations

1. If reading is required by the examinee, does his or her ability match the level required by the test?
2. How appropriate is the length of the test?

Standardization

1. Is the population to be tested similar to the population the test was standardized on?
2. Was the size of the standardization sample adequate?
3. Have specialized subgroup norms been established?
4. How adequately do the instructions permit standardized administration?

Reliability

1. Are reliability estimates sufficiently high (generally around .90 for clinical decision making and around .70 for research purposes)?
2. What implications do the relative stability of the trait, the method of estimating reliability, and the test format have on reliability?

Validity

1. What criteria and procedures were used to validate the test?
2. Will the test produce accurate measurements in the context and for the purpose for which you would like to use it?

(especially reading skill) that examinees must have to understand them adequately. The examinee must be able to read, comprehend, and respond appropriately to the test. Second, some tests are too long, which can lead to a loss of rapport with or extensive frustration on the part of the examinee. Administering short forms of the test may reduce these problems, provided these forms have been properly developed and are interpreted with appropriate caution. Finally, clinicians have to assess the extent to which they need training to administer and interpret the instrument. If further training is necessary, a plan must be developed for acquiring this training.

Standardization

Another central issue relates to the adequacy of norms (see Cicchetti, 1994). Each test has norms that reflect the distribution of scores by a standardization sample. The basis on which individual test scores have meaning relates directly to the similarity between the individual being tested and the sample. If a similarity exists between the group or individual being tested and the standardization sample, adequate comparisons can be made. For example, if the test was standardized on white American college students between the ages of 18 and 22, useful comparisons can be made for college students in that racial and age bracket (if we assume that the test is otherwise sufficiently reliable and valid). The more dissimilar the person is from this standardization group (e.g., different national group, over 70 years of age), the less useful the test is for evaluation. The examiner may need to consult the literature to determine whether research that followed the publication of the test manual has developed norms for different groups. This is particularly important for tests such as the MMPI and the Rorschach, for which norms for various cross-national populations have been published.

Three major questions that relate to the adequacy of norms must be answered. The first is whether the standardization group includes representation from the population on which the examiner would like to use the test. The test manual should include sufficient information to determine the representativeness of the standardization sample. If this information is insufficient or in any way incomplete, it greatly reduces the degree of confidence with which clinicians can use the test. The ideal and current practice is to use stratified random sampling. However, because this can be an extremely costly and time-consuming procedure, many tests do not meet this standard. The second question is whether the standardization group is large enough. If the group is too small, the results may not give stable estimates because of too much random fluctuation. Finally, a test may have specialized subgroup norms as well as broad national norms. Knowledge relating to subgroup norms gives examiners greater flexibility and confidence if they are using the test with similar subgroup populations (see Dana, 2005). This is particularly important when subgroups produce sets of scores that are significantly different from the normal standardization group. These subgroups can be based on factors such as ethnicity, sex, geographic location, age, level of education, socioeconomic status, urban versus rural environment, or even diagnostic history. Knowledge of each of these subgroup norms allows for a more appropriate and meaningful interpretation of scores.

Standardization can also refer to administration procedures. A well-constructed test should have clear instructions that permit examiners to give the test in a manner similar to that of other examiners and also similar to themselves from one testing session and the next. Research has demonstrated that varying the instructions between one administration and the next can alter the types and quality of responses the examinee gives, thereby compromising the test's reliability. Standardization of administration should refer not only to consistent administration procedures but also to ensuring adequate lighting, quiet, no interruptions, and good rapport.

Reliability

The reliability of a test refers to its degree of stability, consistency, and predictability. It addresses the extent to which scores obtained by a person are or would be the same if the person is reexamined by the same test on different occasions. Underlying the concept of reliability is the possible range of error, or error of measurement, of a single score. This is an estimate of the range of possible random fluctuation that can be expected in an individual's score. Because psychological constructs cannot be measured directly (e.g., through measuring a level in blood), test scores are at best an approximation of these constructs, and thus error is always present in the system. It may arise from such factors as a misreading of the items, poor administration procedures, or the changing mood of the client. If there is a large degree of error, the examiner cannot place a great deal of confidence in an individual's scores. The goal of a test constructor is to reduce, as much as possible, the degree of measurement error. If this error reduction is achieved, the difference between one score and another for a measured characteristic is more likely to result from some true difference than from some chance fluctuation.

Two main issues relate to the degree of error in a test. The first is the inevitable, natural variation in human performance. Typically variability is less for measurements of ability than for those of personality and state of being. Whereas ability variables (intelligence, mechanical aptitude, etc.) may show gradual changes resulting from growth and development, many personality traits and states of being are much more highly dependent on factors such as mood. This is particularly true in the case of a characteristic such as anxiety. The practical significance of this in evaluating a test is that certain factors outside the test itself can serve to reduce the reliability that the test can realistically be expected to achieve. Thus, an examiner should generally expect higher reliabilities for an intelligence test than for a test measuring a personality variable such as anxiety. It is the examiner's responsibility to know what is being measured, especially the degree of variability to be expected in the measured trait.

The second important issue relating to reliability is that psychological testing methods are necessarily imprecise. For the hard sciences, researchers can make direct measurements, such as the concentration of a chemical solution, the relative weight of one organism compared with another, or the strength of radiation. In contrast, many constructs in psychology are often measured indirectly. For example, intelligence cannot be perceived directly; it must be inferred by measuring behavior that has been defined as being intelligent. Variability relating to these inferences is likely to produce a certain degree of error resulting from the lack of precision in defining and observing inner psychological constructs. Variability in measurement also occurs simply because people have true (not because of test error) fluctuations in performance between one testing session and the next. Whereas it is impossible to control for the natural variability in human performance, adequate test construction can attempt to reduce the imprecision that is a function of the test itself. Natural human variability and test imprecision make the task of measurement extremely difficult. Although some error in testing is inevitable, the goal of test construction is to keep testing errors within reasonably accepted limits. A high measure of reliability is generally .80 or more, but the variable being measured also changes the expected strength of the statistic. Likewise, the method of determining reliability alters the relative strength of the statistic. Ideally, clinicians should hope for reliability statistics of .90 or higher in tests that are used to make decisions about individuals, whereas a reliability of .70 or more is generally adequate for research purposes.

The purpose of reliability is to estimate the degree of test variance caused by error. The four primary methods of obtaining reliability involve determining (1) the extent to which the test produces consistent results upon retesting (test-retest), (2) the relative accuracy of a test at a given time (alternate forms), (3) the internal consistency of the items (split-half and coefficient alpha), and (4) the degree of agreement between two examiners (interscorer). Another way to summarize this is that reliability can be time to time (test-retest), form to form (alternate forms), item to item (split-half/coefficient alpha), or scorer to scorer (interscorer). Although these are the main types of reliability, there is a fifth type, the Kuder-Richardson; like the split-half and coefficient alpha, it is a measurement of the internal consistency of the test items. However, because this method is considered appropriate only for tests that are relatively pure measures of a single variable, it is not covered in this book.

Test-Retest Reliability

Test-retest reliability is determined by administering the test and then repeating it on a second occasion. The reliability coefficient is calculated by correlating the scores obtained by the same person on the two different administrations. The degree of correlation between the two scores indicates the extent to which the test scores can be generalized from one situation to the next. If the correlations are high, the results are less likely to be caused by random fluctuations in the condition of the examinee or the testing environment. Thus, when the test is being used in actual practice, the examiner can be relatively confident that differences in scores are the result of an actual change in the trait being measured rather than error.

A number of factors must be considered in assessing the appropriateness of test-retest reliability. One is the potential for practice and memory of a test taken on one occasion to affect performance on a second occasion, termed practice effect. Some tasks can simply improve between one administration and the next because of practice. This is a particular problem for speeded and memory tests, such as those found on the Coding and Arithmetic subtests of the WAIS-IV. Another factor to consider is that the interval between administrations, which can affect reliability. A test manual should specify the time interval, as well as any likely significant life changes that the examinees may have experienced, such as counseling, career changes, or psychotherapy. For example, tests of preschool intelligence often give reasonably high correlations if the second administration is within several months of the first one. However, correlations with later childhood or adult IQ are generally low because of innumerable, unavoidable intervening life changes. Additional sources of variation may be the result of random, short-term fluctuations in the examinee or of variations in the testing conditions. In general, test-retest reliability is the preferred method only if the variable being measured is relatively stable. If the variable is highly changeable (e.g., anxiety), this method is usually not adequate.

Alternate Forms

The alternate forms method avoids many of the problems encountered with test-retest reliability. The logic behind alternate forms is that, if the trait is measured several times on the same individual by using parallel forms of the test, the different measurements should produce similar results. The degree of similarity between the scores represents the reliability coefficient of the test. As in the test-retest method, the interval between administrations should always be included in the manual, as well as a description of any likely significant intervening life experiences. If the second administration is given immediately after the first, the resulting reliability is more a measure of the correlation between forms and not across occasions. Correlations determined by tests given with a wide time interval, such as two months or more, provide a measure of both the relation between forms and the degree of temporal stability.

The alternate forms method eliminates many carryover effects, such as the recall of specific items. However, there is still likely to be some carryover effect in that the examinee can learn to adapt to the overall style of the test even when the specific item content between one test and another is unfamiliar. This is most likely when the test involves some sort of problem-solving strategy in which the same principle in solving one problem can be used to solve the next one. An examinee, for example, may learn to use mnemonic aids to increase his or her performance on an alternate form of the WAIS-IV Digit Span subtest.

Perhaps the primary difficulty with alternate forms lies in determining whether the two forms are actually equivalent. For example, if one test is more difficult than its alternate form, the difference in scores may represent actual differences in performance on the two tests rather than differences resulting from the unreliability of the measure. Because the test constructor is attempting to measure the reliability of the test itself and not the differences between the tests, the difference between test scores could confound and lower the reliability coefficient. Alternate forms should be independently constructed tests that use the same specifications, including the same number of items, type of content, format, and manner of administration.

A final difficulty is encountered because of personal examinee differences between one administration and the next. If the alternate forms are administered on different days, the examinee may perform differently because of short-term fluctuations such as mood, stress level, or the relative quality of the previous night's sleep. Thus, an examinee's abilities may vary somewhat from one examination to another, thereby affecting test results. Despite these problems, alternate forms reliability has the advantage of at least reducing, if not eliminating, many carryover and practice effects of the test-retest method. A further advantage is that the alternate test forms can be useful for other purposes, such as assessing the effects of a treatment program (used as pre- and posttests) or monitoring a patient's changes over time by administering the different forms on separate occasions.

Internal Consistency: Split-Half Reliability and Coefficient Alpha

The split-half method and coefficient alpha are the best techniques for determining reliability for a trait with a high degree of fluctuation. Because the test is given only once and the items are correlated with each other, there is only one administration, and it is not possible for the effects of time to intervene as they might with the test-retest method. Thus, the split-half method and coefficient alpha give measures of the internal consistency of the test items rather than the temporal stability of different administrations of the same test. To determine split-half reliability, the test is often split on the basis of odd and even items. This method is usually adequate for most tests. Dividing the test into a first half and second half can be effective in some cases but is often inappropriate because of the cumulative effects of warming up, fatigue, and boredom, all of which can result in different levels of performance on the first half of the test compared with the second. This technique also would not work on a test on which items get progressively harder as the test goes on. In contrast, coefficient alpha correlates the items with each other to determine their consistency.

As is true with the other methods of obtaining reliability, the split-half method and coefficient alpha have limitations. When a test is split in half, there are fewer items on each half, which results in wider variability because the individual responses cannot stabilize as easily around a mean. As a general principle, the longer a test is, the more reliable it is because the larger the number of items, the easier it is for the majority of items to compensate for minor alterations in responding to a few of the other items.

Interscorer Reliability

For some tests, scoring is based partially on the judgment of the examiner. Because judgment may vary between one scorer and the next, it may be important to assess the extent to which reliability might be affected. This is especially true for projectives and even for some ability tests where hard scorers may produce results somewhat different from easy scorers. This variance in interscorer reliability may apply for global judgments based on test scores, such as those with brain damage versus normal, or for small details of scoring, such as whether a person has given a shading versus a texture response on the Rorschach. The basic strategy for determining interscorer reliability is to obtain a series of responses from a single client and to have these responses scored by two different individuals. A variation is to have two different examiners test the same client using the same test and then to determine how close their scores or ratings of the person are. An interscorer reliability coefficient can be calculated using a percentage agreement, a correlation, or a kappa coefficient (which takes into account how much agreement would happen by chance). Any test that requires even partial subjectivity in scoring should provide information on interscorer reliability.

Selecting Forms of Reliability

The best form of reliability is dependent on both the nature of the variable being measured and the purposes for which the test is used. If the trait or ability being measured is highly stable, the test-retest method is preferable, whereas internal consistency is more appropriate for characteristics that are highly subject to fluctuations. When using a test to make predictions, often the test-retest method is preferable because it gives an estimate of the dependability of the test from one administration to the next. This is particularly true if, when determining reliability, an increased time interval existed between the two administrations. If, on the other hand, the examiner is concerned with measuring an individual's state (e.g., current, context-bound feelings of anxiety), split-half or coefficient alpha would likely be best.

Another consideration in evaluating the acceptable range of reliability is the format of the test. Longer tests usually have higher reliabilities than shorter ones. Also, the format of the responses affects reliability. For example, a true-false format is likely to have a lower reliability than multiple choice because each true-false item has a 50% possibility of the answer matching or being correct by chance. In contrast, each question in a multiple-choice format having five possible choices has only a 20% possibility of matching or being correct by chance. A final consideration is that tests with various subtests or subscales should report the reliability for the overall test as well as for each of the subtests. In general, the overall test score has a significantly higher reliability than its subtests. For example, the overall IQ on the WAIS-IV has a higher reliability than any of the more specific and shorter subtests used to calculate the IQ. In estimating the confidence with which test scores can be interpreted, the examiner should take into account the lower reliabilities of the subtests. For example, based on reliability alone, a Full Scale IQ on the WAIS-IV can be interpreted with more confidence than the specific subscale scores.

Most test manuals include a statistical index of the amount of error that can be expected for test scores, which is referred to as the *standard error of measurement* (SEM). The logic behind the SEM is that test scores consist of both truth and error. Thus, there is always noise or error in the system, and the SEM provides a range to indicate how extensive that error is likely to be. The range depends on the test's reliability so that the higher the reliability, the narrower the range of error. The SEM is a standard deviation score so that, for example, a SEM of 3 on an intelligence test would indicate that an individual's score has a 68% chance of being within 3 IQ points from the estimated true score. This is because the SEM of 3 represents a band extending from –1 to +1 standard deviations around the mean. Likewise, there would be a 95% chance that the individual's score would fall in a range within 6 points from the estimated true score. From a theoretical perspective, the SEM is a statistical index of how a person's repeated scores on a specific test are expected to fall around a normal distribution. Thus, it is a statement of the relationship among a person's obtained score, his or her theoretically true score, and the test reliability. Because it is an empirical statement of the probable range of scores, the SEM has more practical usefulness than knowledge of the test reliability. This band of error is also referred to as a *confidence interval*.

The acceptable range of reliability is difficult to identify and depends on several factors. First is the method of reliability that is used. Alternate forms are considered to give the lowest estimate of the actual reliability of a test, while split-half provides the highest estimate. Another consideration is the length of the test. As stated previously, longer tests are expected to have higher reliability coefficients than shorter tests. One way to estimate the adequacy of reliability is by comparing the reliability derived on other similar tests, whether of the same construct or a similar design. The examiner can then develop a sense of the expected levels of reliability, which provides a baseline for comparisons. For example, when evaluating a test measuring anxiety, a clinician may not know what is an acceptable level of reliability. A general estimate can be made by comparing the reliability of the test under consideration with other tests measuring the same or a similar variable. Alternatively, a clinician may look at tests similar in construction (types of questions asked, length, etc.) but measuring a different construct for comparison. The most important thing to keep in mind is that lower levels of reliability usually suggest that less confidence can be placed in the interpretations and predictions based on the test data. However, practitioners are less likely to be concerned with low statistical reliability if they have some basis (e.g., theoretical) for believing the test is a valid measure of the client's state at the time of testing. The main consideration is that a test score should not mean one thing at one time and something different at another.

Validity

The most crucial issue in test construction is validity. Whereas reliability addresses issues of consistency, validity assesses whether a test truly measures the trait it is supposed to measure. A test that is valid for clinical assessment should measure what it is intended to measure and should also produce information useful to clinicians.

A psychological test cannot be said to be valid in any abstract or absolute sense, but more practically, it must be valid in a particular context and for a specific group of people (Messick, 1995). Although a test can be reliable without being valid, the opposite is not true; a necessary prerequisite for validity is that the test must have achieved an adequate level of reliability. That is, a test cannot truly measure what it is supposed to measure if it cannot even measure the same thing each time it is administered. Thus, a valid test is one that accurately measures the variable it is intended to measure. For example, a test comprising questions about a person's musical preference might erroneously state that it is a test of creativity. The test might be reliable in the sense that if it is given to the same person on different occasions, it produces similar results each time. However, it would not be valid in that an investigation might indicate it does not correlate highly with other more valid measurements of creativity.

Establishing the validity of a test can be extremely difficult, primarily because psychological variables are usually abstract and intangible concepts, such as intelligence, anxiety, and personality. These concepts have no tangible reality, so their existence must be inferred through indirect means. In addition, conceptualization and research on constructs undergo change over time requiring that test validation go through continual refinement (G. Smith & McCarthy, 1995). In constructing a test, a test designer must follow two necessary, initial steps. First, the construct must be theoretically evaluated and described; second, specific operations (test questions) must be developed to measure it. Even when the designer has followed these steps closely and conscientiously, it is sometimes difficult to determine what the test really measures. For example, IQ tests are good predictors of academic success, but many researchers question whether they adequately measure the concept of intelligence as it is theoretically described. Another hypothetical test that, based on its item content, might seem to measure what is described as musical aptitude may in reality be highly correlated with verbal abilities. Thus, it may be more a measure of verbal abilities than of musical aptitude.

Any estimate of validity is concerned with relationships between the test and some external independently observed event. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 1999; G. Morgan, Gliner, & Harmon, 2001) list the three main methods of establishing validity as content-related, criterion-related, and construct-related.

Content Validity

During the initial construction phase of any test, the developers must first be concerned with its content validity. *Content validity* refers to the representativeness and relevance of the assessment instrument to the construct being measured. During the initial item development, the constructors must carefully consider the skills, knowledge, or content area of the variable they would like to measure. The items are then generated based on this conceptualization of the variable. At some point, it might be decided that the item content overrepresents, underrepresents, or excludes specific areas, and alterations in the items might be made accordingly. If experts on subject matter are used to determine the items, the number of these experts and their qualifications should be included in the test manual. The instructions they received and the extent of agreement between judges should also be provided. A good test covers not only the subject matter being measured but also additional variables. For example, factual knowledge may be one criterion, but the application of that knowledge and the ability to analyze data are also important. Thus, a test with high content validity must cover all major aspects of the content area and must do so in the correct proportion.

A concept somewhat related to content validity is face validity. These terms are not synonymous, however, because content validity pertains to judgments made by experts, whereas face validity concerns judgments made by the test users. *Face validity* refers to the degree to which a test seems like it is measuring what it purports to measure. For example, a test of arithmetic with a significant collection of arithmetic math problems to solve has high face validity. One issue in face validity is client rapport. A group of potential mechanics who are being tested for basic skills in mathematics may be better served by word problems that relate to machines rather than to business transactions. However, some tests may deliberately have low face validity, in order to decrease opportunities for examinees to skew results purposely. For example, a test like the Rorschach has low face validity for measuring a construct like psychotic thinking—examinees may not realize the test is measuring this construct—specifically to make it more difficult to fake the results in a specific direction. Despite the potential importance of face validity in regard to test-taking attitudes, disappointingly few formal studies on face validity are performed and/or reported in test manuals.

In the past, content validity has been conceptualized and operationalized as being based on the subjective judgment of the test developers. As a result, it has been regarded as the least preferred form of test validation, albeit necessary in the initial stages of test development. In addition, its usefulness has been focused primarily on achievement tests (how well has this student learned the content of the course?) and personnel selection (does this applicant know the information relevant to the potential job?). More recently, content validity has been used more extensively in personality and clinical assessment (Ben-Porath & Tellegen, 2008/2011; Butcher, Graham, Williams, & Ben-Porath, 1990; Harkness, McNulty, Ben-Porath, & Graham, 2002; Millon, Grossman, & Millon, 2015). More recent use of content validity has paralleled more rigorous and empirically based approaches to establishing validity from multiple perspectives.

Criterion Validity

A second major approach to determining validity is criterion validity, which has also been called *concurrent, empirical,* or *predictive validity*. Criterion validity is determined by comparing test scores with some sort of performance on an outside measure. The outside measure should have a theoretical relation to the variable that the test is supposed to measure. For example, an intelligence test might be correlated with grade point average; an aptitude test, with independent job ratings; or a test of anxiety, with other tests measuring similar constructs. The relation between the two measurements is usually expressed as a correlation coefficient.

Criterion-related validity is most frequently divided into either concurrent or predictive validity. *Concurrent validity* refers to measurements taken at the same, or approximately the same, time as the test. For example, an intelligence test might be administered at the same time as assessments of a group's level of academic achievement. *Predictive validity* refers to outside measurements that were taken some time after the test scores were derived. Thus, predictive validity might be evaluated by correlating the intelligence test scores with measures of academic achievement a year after the initial testing. Concurrent validation is often used as a substitute for predictive validation because it is simpler, less expensive, and less time consuming and because participant attrition is not an issue. However, the main consideration in deciding whether concurrent or predictive validation is preferable depends on the test's purpose. Predictive validity is most appropriate for tests used for selection and classification of personnel. This may include hiring job applicants, placing military personnel in specific occupational training programs, screening out individuals who are likely to develop emotional disorders, or identifying which category of psychiatric populations would be most likely to benefit from specific treatment approaches. These situations all require that the measurement device provide a prediction of some future outcome. In contrast, concurrent validation is preferable if an assessment of the client's current state is required rather than a prediction of what might occur to the client at some future time. The distinction can be summarized by asking "Is Mr. Jones maladjusted?" (concurrent validity) rather than "Is Mr. Jones likely to become maladjusted at some future time?" (predictive validity).

An important consideration is the degree to which a specific test can be applied to a unique work-related environment (see Hogan, Hogan, & Roberts, 1996). This consideration relates more to the social value and consequences of the assessment than the formal validity as reported in the test manual (Messick, 1995). In other words, can the test under consideration provide accurate assessments and predictions for the environment in which the examinee is working? To answer this question adequately, the examiner must refer to the manual and assess the similarity between the criteria used to establish the test's validity and the situation to which he or she would like to apply the test. For example, can an aptitude test that has adequate criterion-related validity in the prediction of high school grade point average also be used to predict academic achievement for a population of college students? If the examiner has questions regarding the relative applicability of the test, he or she may need to undertake a series of specific tasks. The first is to identify the required skills for adequate performance in the situation involved. For example, the criteria for a successful teacher may include such attributes as verbal fluency, flexibility, and good public speaking skills. The examiner then must determine the degree to which each skill contributes to the quality of a teacher's performance. Next, the examiner has to assess the extent to which the test under consideration measures each of these skills. The final step is for the examiner to evaluate the extent to which the attribute that the test measures is relevant to the skills he or she needs to predict. Based on these evaluations, the examiner can estimate the confidence that he or she places in the predictions developed from the test. This approach is sometimes referred to as *synthetic validity* because examiners must integrate or synthesize the criteria reported in the test manual with the variables they encounter in their clinical or organizational settings.

The strength of criterion validity depends in part on the type of variable being measured. Usually, intellectual or aptitude tests give relatively higher validity coefficients than personality tests because there are generally a greater number of variables influencing personality than intelligence. As the number of variables that influences the trait being measured increases, it becomes progressively more difficult to account for them. When a large number of variables are not accounted for, the trait can be affected in unpredictable ways. This situation can create a much wider degree of fluctuation in the test scores, thereby lowering the validity coefficient. Thus, when evaluating a personality test, the examiner should not expect as high a validity coefficient as for intellectual or aptitude tests. A helpful guide is to look at the validities found in similar tests and compare them with the test being considered. For example, if an examiner wants to estimate the range of validity to be expected for the extraversion scale on the Myers Briggs Type Indicator (MBTI), he or she might compare it with the validities for similar scales found in the NEO-PI-3 and Eysenck Personality Questionnaire. The relative level of validity, then, depends both on the quality of the construction of the test and on the variable being studied.

An important consideration is the extent to which it is realistically expected that the trait being measured should predict the trait to which it is being compared. For example, the typical correlation between intelligence tests and academic performance is about .50 (Neisser et al., 1996). Because no one would say that grade point average is entirely the result of intelligence, the relative extent to which intelligence determines grade point average has to be estimated. It can be calculated by squaring the correlation coefficient and changing it into a percentage. Thus, if the correlation of .50 is squared, it comes out to 25%, indicating that 25% of academic achievement can be accounted for by IQ as measured by the intelligence test. The remaining 75% may include factors such as motivation, quality of instruction, and past educational experience. The problem facing the examiner is to determine whether 25% of the variance is sufficiently useful for the intended purposes of the test. This determination ultimately depends on the personal judgment of the examiner.

The main problem confronting criterion validity is finding an agreed-upon, definable, acceptable, and feasible outside criterion. Whereas for an intelligence test, grade point average might be an acceptable criterion, it is far more difficult to identify adequate criteria for most personality tests. Even with so-called intelligence tests, many researchers argue that it is more appropriate to consider them tests of scholastic aptitude rather than of intelligence. Yet another difficulty with criterion validity is the possibility that the criterion measure will be inadvertently biased. Referred to as *criterion contamination*, this occurs when knowledge of the test results influences an individual's later performance. For example, a supervisor in an organization who receives such information about subordinates may act differently toward a worker placed in a certain category after being tested. This situation may set up negative or positive expectations for the worker, which could influence his or her level of performance. The result is likely to artificially increase the level of the validity coefficients. To work around these difficulties, especially in regard to personality tests, a third major method must be used to determine validity.

Construct Validity

The method of construct validity was developed in part to correct the inadequacies and difficulties encountered with content and criterion approaches. Early forms of content validity relied too heavily on subjective judgment, while criterion validity was too restrictive in working with the domains or structure of the constructs being measured. Criterion validity had the further difficulty in that there was often a lack of agreement in deciding on adequate outside criteria. The basic approach of construct validity is to build a strong case that the test measures a theoretical construct or trait. This assessment involves three general steps. Initially, the test constructor must make a careful analysis of the trait. Then the test designer must consider the ways in which the trait should relate to other variables. Finally, the test designer needs to test whether these hypothesized relationships actually exist (Foster & Cone, 1995). For example, a test measuring dominance should have a high positive correlation with the individual accepting leadership roles, a high negative correlation with measures of submissiveness, and a very low correlation to measure of some unrelated trait, like openness. Likewise, a test measuring anxiety should have a high positive correlation with individuals who are measured during an anxiety-provoking situation, such as an experiment involving some sort of physical pain. As these hypothesized relationships are verified by research studies, the case for the measure's construct validity gets stronger and the degree of confidence that can be placed in the test increases.

There is no single, best approach for determining construct validity; rather, a variety of different possibilities exists. For example, if some abilities are expected to increase with age, correlations can be made between a population's test scores and age. This method may be appropriate for variables such as general fund of knowledge or motor coordination, but it would not be applicable for most emotional measurements. Even in the measurement of fund of knowledge or motor coordination, this approach may not be appropriate beyond the age of maturity. Another method for determining construct validity is to measure the effects of experimental or treatment interventions. Thus, a posttest measurement may be taken following a period of instruction to see if the intervention affected the test scores in relation to a previous pretest measure. For example, after an examinee completes a course in arithmetic, it would be predicted that scores on a test of arithmetical ability would increase. Often correlations can be made with other tests that supposedly measure a similar variable. However, a new test that correlates too highly with existing tests may represent needless duplication, unless it incorporates some additional advantage, such as a shortened format, ease of administration, or superior predictive validity. Related to this line of validation is presenting an argument that the test method is not majorly responsible for test scores. That is, a true/false test developed to measure anxiety should have a low correlation with a true/false test used to measure food preferences. If these scores are highly related (despite being theoretically unrelated), it may be that the scores on these tests are heavily influenced by the fact that they are true/false tests rather than by the content they are supposed to be measuring.

Factor analysis is of particular relevance to construct validation because it can be used to identify and assess the relative strength of different psychological traits. Factor analysis can also be used in the design of a test to identify the primary factor or factors measured by a series of different tests. Thus, it can be used to simplify one or more tests by reducing the number of categories to a few common factors or traits. The factorial validity of a test is the relative weight or loading that a factor has on the test. For example, if a factor analysis of a measure of anxiety determined that the test was composed of three clear factors that seemed to be measuring cognitive aspects of anxiety, affective aspects of anxiety, and physiological aspects of anxiety, the test could be considered to have factorial validity. This would be especially true if the three factors seemed to be accounting for a clear and large portion of what the test was measuring.

Another method used as a component to build construct validity is to estimate the degree of internal consistency by correlating specific subtests with the test's total score. For example, if a subtest on an intelligence test does not correlate adequately with the overall or Full Scale IQ, it should be either eliminated or altered in a way that increases the correlation. A final method for obtaining construct validity is for a test to converge or correlate highly with variables that are theoretically similar to it. The test should not only show this convergent validity but also have discriminant validity, in which it would demonstrate low correlations with variables that are dissimilar to it. Thus, scores on reading comprehension should show high positive correlations with performance in a literature class and low correlations with performance in a class involving mathematical computation.

Related to discriminant and convergent validity is the degree of sensitivity and specificity an assessment device demonstrates in identifying different categories. *Sensitivity* refers to the percentage of true positives that the instrument has identified, whereas *specificity* is the relative percentage of true negatives. A structured clinical interview might be quite sensitive in that it would accurately identify 90% of people with schizophrenia in an admitting ward of a hospital. However, it may not be sufficiently specific in that 30% of individuals without schizophrenia would be incorrectly classified as having schizophrenia (a true negative rate of 70%). The difficulty in determining sensitivity and specificity lies in developing agreed-upon, objectively accurate outside criteria for categories such as psychiatric diagnosis, intelligence, or personality traits.

As indicated by the variety of approaches discussed, no single, quick, efficient method exists for determining construct validity. Establishing construct validity is the building of a strong case, an amassing of evidence. The process is similar to testing a series of hypotheses for which the results of the studies determine the meanings that can be attached to later test scores (Foster & Cone, 1995; Messick, 1995). Almost any data can be used, including material from the content and criterion approaches. The greater the amount of supporting data, the greater is the level of confidence with which the test can be used. As a result, construct validity represents the strongest and most sophisticated approach to test validation. In many ways, all types of validity can be considered subcategories of construct validity. Construct validation involves theoretical knowledge of the trait or ability being measured, knowledge of other related variables, hypothesis testing, and statements regarding the relationship of the test variable to a network of other variables that have been investigated (G. T. Smith, 2005). Thus, construct validation is a never-ending process in which new relationships always can be verified and investigated.

VALIDITY IN CLINICAL PRACTICE

Although a test may have been found to have a high level of validity during its construction, it does not necessarily follow that the test is also valid in a specific situation with a particular client. A test can never be valid in any absolute sense because, in practice, numerous variables might affect the test results. A serious issue, then, is the degree of validity generalization that is made. In part, this generalization depends on the similarity between the population used during various stages of test construction and the population and situation that it is being used for in practice. Validity in clinical practice also depends on the extent to which tests can work together to improve each other's accuracy. Some tests thus show incremental validity in that they improve overall accuracy in increments as increasing numbers of data sources are used. *Incremental validity*, then, refers to the ability of tests to produce information above what is already known. Another important consideration is the ability of the clinician to generate hypotheses, test these hypotheses, and blend the data derived from hypothesis testing into a coherent, integrated picture of the person (for a full discussion of this process, see Wright, 2010). Maloney and Ward (1976) refer to this latter approach to validity as *conceptual validity* because it involves creating a conceptually coherent description of the person.

Incremental Validity

For a test to be considered useful and efficient, it must be able to produce accurate results above and beyond the results that could be obtained with greater ease and less expense (Hunsley & Meyer, 2003). If equally accurate clinical descriptions could be obtained through such basic information as biographical data and knowing the referral question, there would be no need for psychological tests. Incremental validity also needs to be evaluated in relation to cost-effectiveness. A psychological test might indeed demonstrate incremental validity by increasing the relative proportions of accurate diagnoses, or hit rates, by 2%. However, practitioners need to question whether this small increase in accuracy is worth the extra time and cost involved in administering and interpreting the test. Clinicians might focus their time more productively directly toward treatment.

In the 1950s, one of the theoretical defenses for tests having low reliabilities and validities was that, when used in combination, their accuracy could be improved. In other words, results from a series of different tests could provide checks and balances to correct for inaccurate interpretations. A typical strategy used to empirically test for this was to first obtain biographical data, make interpretations and decisions based on these data, and then test their accuracy based on some outside criterion. Next, a test such as the MMPI could be given; then the interpretations and decisions based on it could likewise be assessed for accuracy. Finally, clinicians could be given both sets of data to assess any improvements in the accuracies of interpretation/decisions between either of the first two conditions and the combined information.

It would seem logical that the greater the number of tests used, the greater would be the overall validity of the assessment battery. However, research on psychological tests used in clinical practice has often demonstrated that they have poor incremental validity. An older but representative study by Kostlan (1954) on male psychiatric outpatients compared the utility of a case history, Rorschach, MMPI, and a sentence completion test. Twenty experienced clinicians interpreted different combinations of these sources of test data. Their conclusions were combined against criterion judges who used a lengthy checklist of personality descriptions. The conclusions were that, for most of the data, the clinicians were no more accurate than if they had used only age, occupation, education, marital status, and a basic description of the referral question. The exception was that the most accurate descriptions were based on a combination of social history and the MMPI. In contrast, psychological tests have sometimes clearly demonstrated their incremental validity. S. Schwartz and Wiedel (1981) demonstrated that neurological residents gave more accurate diagnoses when an MMPI was used in combination with history, electroencephalogram (EEG), and physical exam. This was probably not so much because of a specific MMPI neurological profile but rather because the MMPI increased diagnostic accuracy by enabling the residents to rule out other possible diagnoses.

Often clinical psychologists attempt to make a series of behavioral predictions based on complex psychological tests. Although these predictions may show varying levels of accuracy, a simpler and more effective means of achieving this information might be simply to ask the clients to predict their own behaviors. In some circumstances, self-prediction has been found to be more accurate than psychological tests, whereas in others, tests have been found to be more accurate (Shrauger & Osberg, 1981). Advantages of self-assessment are that it can be time-efficient and cost-effective and can facilitate a collegial relationship between assessor and client. In contrast, difficulties are that, compared with formal testing, self-assessment may be significantly more susceptible to social desirability, attributional errors, distortions caused by poor adjustment, and the relative self-awareness of the client. These factors need to be carefully considered before the clinician decides to use self-assessment versus formal psychological tests. Although the incremental validity of using self-assessment in combination with formal testing has not been adequately researched, it would seem that this is conceptually a potentially useful strategy for future research.

Reviews of studies on incremental validity (Garb, 1998, 2003, 2005b) have provided a number of general conclusions. The addition of an MMPI to background data has consistently led to increases in validity, although the increases were quite small when the MMPI was added to extensive data. The addition of projective tests to a test battery did not generally increase incremental validity. Lanyon and Goodstein (1982) have argued that case histories are generally preferable to psychological test data. Furthermore, a single test in combination with case history data is generally as effective as a large number of tests with case history data. Some studies have found that the MMPI alone was generally preferable to a battery containing the MMPI, Rorschach, and sentence completion (Garb, 1984, 1994a, 1998, 2005b). In contrast, other studies have found that the Rorschach can add incremental validity to a test battery (G. Meyer, 1997; Weiner, 1999).

The poor demonstrated incremental validity of many of the traditional clinical tests may relate to weaknesses and unanswered questions in the research. First, few studies have looked at statistically derived predictions and interpretations based on optimal multiple cutoff scores or multiple regression equations. However, more recent research, particularly on tests like the MMPI-2 and California Personality Inventory (CPI), has emphasized this approach. For example, combined weightings on such variables as specific CPI scores, Scholastic Aptitude Test (SAT) scores, grade point average (GPA), and IQ can be combined to predict success in specific programs (e.g., Aegisdottir, White, Spengler, Maugherman, Anderson, Cook et al., 2006; Grove, Zald, Lebow, Snitz, & Nelson, 2000). Further research using this approach may yield greater incremental validity for a wide number of assessment techniques. Second, few studies on incremental validity have investigated the ways in which different tests might show greater incremental validity in specific situations for specific populations. Instead, most research has focused on the validity of global personality descriptions, without tying these descriptions to the unique circumstances or contexts persons might be involved in. Finally, as most previous studies have focused on global personality descriptions, certain tests demonstrate greater incremental validity when predicting highly specific traits and behaviors.

Conceptual Validity

A further method for determining validity that is highly relevant to clinical practice is conceptual validity (Maloney & Ward, 1976). In contrast to the traditional methods (content validity, etc.), which are primarily concerned with evaluating the theoretical constructs in the test itself, conceptual validity focuses on individuals with their unique histories and behaviors. It is a means of evaluating and integrating test data so that the clinician's conclusions make accurate statements about the examinee. There are similarities with construct validity in that construct validity also tries to test specific hypothesized relationships between constructs. Conceptual validity is likewise concerned with testing constructs, but in this case the constructs relate to the individual rather than to the test itself.

In determining conceptual validity, the examiner generally begins with individuals for whom no constructs have been developed. The next phase is to observe, collect data, and form a large number of hypotheses. If these hypotheses are confirmed through consistent trends in the test data, behavioral observations, history, and additional data sources, the hypotheses can be considered to represent valid constructs regarding the person. The focus is on an individual in his or her specific situation, and the data are derived from a variety of sources. The conceptual validity of the constructs is based on the logic and internal consistency of the data. Unlike construct validity, which begins with previously developed constructs, conceptual validity produces constructs as its end product. Its aim is for these constructs to provide valid sources of information that can be used to help solve the unique problems that an individual may be facing.

CLINICAL JUDGMENT

Any human interaction involves mutual and continually changing perceptions. *Clinical judgment* is a special instance of perception in which the clinician attempts to use whatever sources are available to create accurate descriptions of the client. These sources may include test data, case history, medical records, personal journals, and verbal and nonverbal observations of behavior. Relevant issues and processes involved in clinical judgment include data gathering, data synthesis, the relative accuracy of clinical versus statistical/actuarial descriptions, and judgment in determining what to include in a psychological report. This sequence also parallels the process clinicians go through when assessing a client.

Data Gathering and Synthesis

Most of the research related to the strengths and weaknesses of data gathering and synthesis has focused on the assessment interview (see Chapter 3). However, many of the issues and problems related to clinical judgment during interviewing also have implications for the gathering and synthesis of test data. One of the most essential elements in gathering data from any source is the development of an optimum level of rapport. Rapport increases the likelihood that clients will give their optimum level of performance. If rapport is not sufficiently developed, it is increasingly likely that the data obtained from the person will be inaccurate.

Another important issue is that the interview itself is typically guided by the client's responses and the clinician's reaction to these responses. A client's responses might be nonrepresentative because of factors such as a transient condition (stressful day, poor night's sleep, etc.) or conscious/unconscious faking. The client's responses also need to be interpreted by the clinician. These interpretations can be influenced by a combination of personality theory, research data, and the clinician's professional and personal experience. The clinician typically develops hypotheses based on a client's responses and combines his or her observations with his or her theoretical understanding of the issue. These hypotheses can be further investigated and tested by interview questions and test data, which can result in confirmation, alteration, or elimination of the hypotheses. Thus, bias can potentially enter into this process from a number of different directions, including the types of questions asked, initial impressions, level of rapport, or theoretical perspective.

The clinician typically collects much of the initial data regarding a client through unstructured or semistructured interviews. Unstructured approaches in gathering and interpreting data provide flexibility, focus on the uniqueness of the person, and are ideographically rich. In contrast, an important disadvantage of unstructured approaches is that a clinician, like most other persons, can be influenced by a number of personal and cultural biases. For example, clinicians might develop incorrect hypotheses based on first impressions (primacy effect). They might end up seeking erroneous confirmation of incorrect hypotheses by soliciting expected responses rather than objectively probing for possible disconfirmation. Thus, clinicians might be unduly influenced by their preferred theory of personality, halo effects, expectancy bias, and cultural stereotypes. These areas of potential sources of error have led to numerous questions regarding the dependability of clinical judgment.

Accuracy of Clinical Judgments

After collecting and organizing their data, clinicians then need to make final judgments regarding the client. Determining the relative accuracy of these judgments is crucial. In some cases, clinical judgment is clearly in error, whereas in others it can be quite accurate. Cultural bias can come into play, and clinicians should take into consideration cultural context and personal beliefs when making clinical judgments. To increase accuracy, clinicians need to know how errors might occur, how to correct these errors, and the relative advantages of specialized training.

A possible source of inaccuracy is that clinicians frequently do not take into account the base rate, or the rate at which a particular behavior, trait, or diagnosis occurs in the general population (Faust, 1991; S. Hawkins & Hastie, 1990; Wedding & Faust, 1989). For example, an intake section of a psychiatric hospital might use a test that has been shown to be 90% accurate at telling whether a person has schizophrenia. Perhaps 5% of the time the test shows a false positive and 5% of the time it shows a false negative. If a person comes in and the test reveals a positive result for schizophrenia, it is not necessarily a 90% or 95% chance that he or she actually has schizophrenia. Because schizophrenia has a low base rate (e.g., if roughly 1% of the population has it), there is actually a much greater than 10% chance that this individual does not have schizophrenia.

It is also rare for clinicians to receive feedback regarding either the accuracy of their diagnoses or other frequently used judgments, such as behavioral predictions, personality traits, or the relative success of their recommendations (Garb, 1989, 1994a, 1998, 2005b). Thus, it is possible that inaccurate strategies for arriving at conclusions will continue with little likelihood of correction.

A further source of error is that information obtained earlier in the data collection process is frequently given more importance than information received later (primacy effect). This means that different starting points in the decision-making process may result in different conclusions. This error can be further reinforced if clinicians make early judgments and then work to confirm these judgments through seeking supporting information. The resulting *confirmatory bias* is especially likely to occur in a hypothesis-testing situation in which clinicians do not adequately seek information that could disconfirm as well as confirm their hypothesis (Haverkamp, 1993). The most problematic examples occur when clinicians interpret a client's behavior and then work to persuade the client that their interpretation is correct (Loftus, 1993).

Research on person perception accuracy indicates that, even though nobody is uniformly accurate, some persons are much better at accurately perceiving others. Taft (1955) and P. E. Vernon (1964) summarized the early research on person perception accuracy by pointing out that accuracy is not associated with age (in adults); there is little difference in accuracy between males and females (although females are slightly better); and accurate perceptions of others are positively associated with intelligence, artistic/dramatic interests, social detachment, and good emotional adjustment. Authoritarian personalities tend to be poor judges. In most instances, accuracy is related to similarity in race and cultural backgrounds (P. Shapiro & Penrod, 1986). In some cases, accuracy by psychologists may be only slightly related to their amount of clinical experience (Garb, 1989, 1992, 1994a, 1998, 2005b); and, for some judgments, psychologists may be no better than certain groups of nonprofessionals, such as physical scientists and personnel workers (Garb, 1992, 1994a, 1998, 2005b). Relatively higher rates of accuracy were achieved when clinical judgments based on interviews were combined with formal assessments and when statistical interpretive rules were used. When subjective test interpretation was combined with clinical judgment, it was questionable whether any increase in accuracy was obtained (Garb, 1984, 1989).

It would be logical to assume that the more confidence clinicians feel regarding the accuracy of their judgments, the more likely it is that their judgments are accurate. In several studies, however, confidence was often not related to accuracy (E. Kelly &

Fiske, 1951; Kleinmuntz, 1990). Kelly and Fiske even found that degree of confidence was inversely related to predicting the success of trainees in a Veterans Administration training program. Several studies (Kareken & Williams, 1994; Lichtenstein & Fischoff, 1977) concluded that persons were generally overconfident regarding judgments; and when outcome knowledge was made available, clinicians typically overestimated what they thought they knew before receiving outcome knowledge (Hawkins & Hastie, 1990). This overconfidence is usually referred to as *hindsight bias* ("I would have known it all along") and is usually accompanied by a denial that the outcome knowledge has influenced judgment. Paradoxically, as knowledge and experience in an area increase, there is generally a decrease in confidence regarding judgments. This observation was found to be true unless the clinicians were very knowledgeable, in which case they were likely to have a moderate level of confidence (Garb, 1989). Confidence was also higher if participants were made socially accountable for their judgments (Ruscio, 2000). Thus, the more experienced clinicians and persons who were more socially accountable rated their level of confidence as higher.

Crucial to clinical judgment is whether clinicians can make judgments better than laypersons and whether amount of clinical training can increase accuracy. This is a particularly important issue if psychologists are offering their services as expert witnesses to the legal justice system. Research reviews generally support the value of clinical training, but this is dependent on the domain being assessed. For example, Garb (1992) concluded, "Clinicians are able to make reliable and valid judgments for many tasks, and their judgments are frequently more valid than judgments by laypersons" (p. 451). In particular, clinicians have been found to make more accurate judgments relating to relatively complex technical areas, such as clinical diagnosis, ratings of mental status, many domains related to interview information, short-term (and possibly long-term) predictions of violence, psychological test interpretation (WAIS, MMPI), forensic knowledge, competency evaluations, neuropsychological test results, psychotherapy data, and biographical data (see primarily Garb, 1998, but also 1984, 1989, 1992, 1994a). In contrast, trained clinicians were no better than less experienced persons (laypersons, novice trainees) in making judgments based on projective test results and in making personality descriptions based on face-to-face interaction (Garb, 2005b; Witteman & van den Bercken, 2007).

The preceding material indicates that errors in clinical judgment can and do occur. It is thus crucial, especially when appearing as an expert in court, that clinicians be familiar with the relevant literature on clinical judgment and, based on this information, take steps to improve their accuracy. Accordingly, Garb (1994a, 1998, 2005b) and Wedding and Faust (1989) made the following recommendations:

1. To avoid missing crucial information, clinicians should use comprehensive, structured, or at least semistructured approaches to interviewing. This is especially important in cases where urgent clinical decisions (danger to self or others) may need to occur.
2. Clinicians should not only consider the data that support their hypotheses, but they should also carefully consider or even list evidence that does not support their hypotheses. This method will likely reduce the possibility of hindsight and confirmatory bias.

3. Diagnoses should be based on careful attention to the specific criteria contained in the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*-5; American Psychiatric Association, 2013) or *International Classification of Disorders* (*ICD–10*; World Health Organization, 1992). In particular, this means not making errors caused by inferences biased by gender and ethnicity.
4. Because memory can be a reconstructive process subject to possible errors, clinicians should avoid relying on memory and rather refer to careful notes as much as possible.
5. In making predictions, clinicians should attend to base rates as much as possible. Such a consideration potentially provides a rough estimate of how frequently the behavior will occur in a given population or context. Any clinical predictions, then, are guided by this base rate occurrence and are likely to be improvements on the base rate.
6. Clinicians should seek feedback when possible regarding the accuracy and usefulness of their judgments. For example, psychological reports should ideally be followed up with rating forms (that can be completed by the referral sources) relating to the clarity, precision, accuracy, and usefulness of the information and recommendations contained in the reports.
7. Clinicians should learn as much as possible regarding the theoretical and empirical material relevant to the person or group they are assessing. Doing this would potentially help clinicians to develop strategies for obtaining comprehensive information, allow them to make correct estimates regarding the accuracy of their judgments, and provide them with appropriate base rate information.
8. Practitioners should be familiar with the literature on clinical judgment in order to continually update their knowledge on past and emerging trends.

Sometimes in court proceedings, psychologists are challenged regarding the difficulties associated with clinical judgment. If the preceding steps are taken, psychologists can justifiably reply that they are familiar with the literature and have taken appropriate steps to guard against inaccuracies in clinical judgment. More important, by taking these steps, the clinicians' quality of service related to clients and referral sources is also likely to be enhanced.

Clinical Versus Actuarial Prediction

Over 60 years ago, Meehl (1954) published a review of research comparing the relative accuracy of clinical judgment to statistical formulas when used on identical sets of data (life history, demographic data, test profiles). The clinical approach used clinicians' judgment, whereas the actuarial approach used empirically derived formulas, such as single/multiple cutoffs and regression equations, to come to decisions regarding a client. His review covered a large number of settings including military placement, college success, criminal recidivism, and benefit from psychotherapy. He concluded that statistical decisions consistently outperformed clinical judgments (Meehl, 1954, 1965). Some lively debate in the journals ensued, with Meehl's conclusions generally being supported (Aegisdottir et al., 2006; Garb, 1994b; Grove et al., 2000; Kleinmuntz, 1990).

The magnitude of this difference has been estimated to be a 13% greater accuracy using actuarial methods when compared with clinical judgment.

Despite the empirical support for an actuarial approach, several practical and theoretical issues need to be considered. A clinical approach to integrating data and arriving at conclusions allows a clinician to explore, probe, and deepen his or her understanding in many areas. These explorations frequently involve areas that tests or statistical formulas cannot measure. Often an interview is the only means of obtaining observations of behavior and unique aspects of history. Idiosyncratic events with a low frequency of occurrence may significantly alter a clinician's conclusions although no formulas take these events into account. It is quite common for unique, rare events to have occurred at some time in a client's life; and, during the process of assessment, they are frequently relevant and can often alter the conclusions of many, if not most, clinical assessments. Not only do unique aspects of a person change interpretations, but typically an assessment for a person needs to be focused for a specific context and specific situation that he or she is involved in. When the focus changes from institutional to individual decision making, the relevance of statistical rules becomes less practical (McGrath, 2001; Vane & Guarnaccia, 1989). Not only are individuals too multifaceted for simple actuarial formulas, but their unique situations, contexts, and the decisions facing them are even more multifaceted.

A further difficulty with a purely actuarial approach is that development of both test reliability and validity, as well as actuarial formulas, requires conceiving the world as stable and static. For such approaches to be useful, the implicit assumption is that neither people nor criteria change. In contrast, the practitioner must deal with a natural world that is imperfect, constantly changing, does not necessarily follow rules, is filled with constantly changing perceptions, and is subject to chance or at least impossible-to-predict events. Thus, even when statistical formulas are available, they may not apply. This distinction between the statistical orientation of the psychometrician and the natural environment of the practitioner underlies the discrepancy between their two worlds (Beutler, 2000). Practitioners must somehow try to combine these two modes of analysis, but they often find the task difficult. It may be true that controlled studies generally favor a statistical approach over a clinical one, but, at the same time, that truth is seldom useful to the practitioner involved in the changing and unique world of practice (Bonarius, 1984).

Bonarius (1984) presented a conceptual alternative to this dilemma. The first step is to alter mechanistic views of prediction. Instead, clinicians might avoid the term *prediction* altogether and use *anticipation*. Anticipating future possibilities implies a cognitive constructional process rather than a mechanical process. It admits that the world can never be perfect in any mechanistic sense and that there is no such thing as an average person in an average situation engaged in an average interaction. Furthermore, the creation of future events is shared by coparticipants. Clients take an active part in formulating and evaluating their goals. The success of future goals depends on the degree of effort they are willing to put into them. The coparticipants share responsibility for the future. Thus, the likelihood that future events will occur is related to both cognitive constructions of an idiosyncratic world and interaction between participants.

Ideally, clinicians need to be aware of and to use, whenever available, actuarial approaches, such as multiple cutoffs and regression equations. Doing so would be particularly important for situations where there are clearly defined outcomes, errors are costly, and clinicians need to have maximum accountability. Such situations might include suicide, violence, sexual offending, recidivism, relapse, postparole adjustment, malingering, response to psychotherapy, academic performance, vocational success, psychiatric prognosis, or success in training programs. Despite over 50 years of research and debates, actuarial strategies are still not widely available except within forensic contexts. In addition, many of the formulas are "not ready for prime time" (Aegisdottir et al., 2006). It is hoped that at some time in the future, a set of optimal, well-validated actuarial formulas will be widely available along with user-friendly programs on how to use them (Groth-Marnat, 2000b, 2009). The results from such formulas will still need to be integrated with data and inferences obtainable only through clinical means. Although it is unlikely that actuarial prediction rules will replace clinical judgment, formal prediction rules can and should be used more extensively as a resource to improve the accuracy of clinical decision making.

Psychological Report

An accurate and effective psychological report requires that clinicians clarify their thinking and crystallize their interpretations. The report ties together all sources of information, often combining complex interprofessional and interpersonal issues. All the advantages and limitations involved with clinical judgment either directly or indirectly affect the report. The focus should be a clear communication of the clinician's interpretations, conclusions, and recommendations. Chapter 15 provides in-depth information on the psychological report as it relates to relevant research, guidelines, format, and sample reports.

PHASES IN CLINICAL ASSESSMENT

An outline of the phases of clinical assessment can provide both a conceptual framework for approaching an evaluation and a summary of some of the points already discussed. Although the steps in assessment are isolated for conceptual convenience, in actuality, they often occur simultaneously and interact with one another. Throughout these phases, the clinician should integrate data and serve as an expert on human behavior rather than merely an interpreter of test scores. Doing so is consistent with the belief that a psychological assessment can be most useful when it addresses specific individual problems and provides guidelines for decision making regarding these problems.

Evaluating the Referral Question

Many of the practical limitations of psychological evaluations result from an inadequate clarification of the problem. Because clinicians are aware of the assets and limitations of psychological tests, and because clinicians are responsible for providing useful information, it is their duty to clarify the requests they receive. Furthermore, they cannot assume that initial requests for an evaluation are adequately stated.

Clinicians may need to uncover hidden agendas, unspoken expectations, and complex interpersonal relationships. One of the most important general requirements is that clinicians understand the vocabulary, conceptual model, dynamics, and expectations of the referral setting in which they will be working (Turner et al., 2001). Further, clinicians must evaluate whether the referral questions are appropriate for psychological assessment and whether they have a level of competence necessary to conduct an assessment to answer the specific questions.

Clinicians are rarely asked to give a general or global assessment but instead are asked to answer specific questions. To address these questions, it is sometimes helpful to contact the referral source at different stages in the assessment process. For example, it is often important in an educational evaluation to observe the student in the classroom environment. The information derived from such an observation might be relayed back to the referral source for further clarification or modification of the referral question. Likewise, an attorney may wish to somewhat alter his or her referral question based on preliminary information derived from the clinician's initial interview with the client.

Data Collection

After clarifying the referral question and obtaining knowledge related to the problem, clinicians can proceed with the actual collection of information. The information may come from a wide variety of sources, the most frequent of which are interview data, collateral information, behavioral observations, and test scores. Collateral information may include school records, previous psychological reports, medical records, police reports, or interviews with parents or teachers. It is important to realize that the tests themselves are merely a single tool, or source, for obtaining data. The case history is of equal importance because it provides a context for understanding the client's current problems and, through this understanding, renders the test scores meaningful. In many cases, a client's history is of even more significance in making predictions and in assessing the seriousness of his or her condition than his or her test scores. For example, a high score on depression on the MMPI-2 is not as helpful in assessing suicide risk as are historical factors, such as the number of previous attempts, details regarding any previous attempts, and length of time the client has been depressed. Moreover, test scores themselves are usually not sufficient to answer the referral question. For specific problem solving and decision making, clinicians must rely on multiple sources and, using these sources, check to assess the consistency of the observations they make.

Before beginning the actual testing procedure, examiners should carefully consider the problem, the adequacy of the tests they will use, and the specific applicability of that test to an individual's unique situation. This preparation may require referring both to the test manual and to additional outside sources. Clinicians should be familiar with operational definitions for problems such as anxiety disorders, psychoses, personality disorders, and organic impairment so that they can be alert to their possible expression during the assessment procedure. Clinicians should also be familiar with problems that can arise from medical conditions and substance use. Competence in merely administering and scoring tests is insufficient to conduct effective assessment. For example, the development of an IQ score does not necessarily indicate that an examiner is aware of differing cultural expressions of intelligence or of the limitations of the assessment device. It is essential that clinicians have in-depth knowledge about the variables they are measuring; if not, their evaluations are likely to be extremely limited.

When evaluating whether a test will be useful in a specific case, a clinician should consider several factors. The relative adequacy of the test will include inquiry about certain practical considerations, the standardization sample, and reliability and validity (see Table 1.1). Specifically, a test should truly measure a construct of interest in the specific case. It is important that the examiner also consider whether a specific test or tests are appropriate to use on an individual or group. Doing this demands knowledge in such areas as the client's age, sex, ethnicity, race, culture, educational background, motivation for testing, anticipated level of resistance, social environment, and interpersonal relationships. Finally, clinicians need to assess the effectiveness or utility of the test in aiding the treatment process.

Interpreting the Data

The end product of assessment should be a set of recommendations that are clear, specific, and reasonable. In order to support these recommendations, clinicians should be able to describe the client's current level of functioning, considerations relating to etiology, and prognosis. Etiologic descriptions should avoid simplistic formulas and should instead focus on the influence exerted by several interacting factors, which may include primary, predisposing, precipitating, and reinforcing causes. Further elaborations may also attempt to assess the person from a systems perspective, in which the clinician evaluates patterns of interaction, mutual two-way influences, and the specifics of reciprocal information feedback. An additional crucial area is to use the data to develop an effective plan for intervention (see Beutler, Clarkin, & Bongar, 2000; Harwood, Beutler, & Groth-Marnat, 2011; Hersen, 2005a; Jongsma, Peterson, & Bruce, 2014; Maruish, 2004). Clinicians should also pay careful attention to research on, and the implications of, incremental validity and continually be aware of the limitations and possible inaccuracies involved in clinical judgment. If actuarial formulas are available, they should be used when possible. These considerations indicate that the description of a client should not be a mere labeling or classification but should rather provide a deeper and more accurate understanding of the person. This understanding should allow the examiner to perceive new facets of the person in terms of both his or her internal experience and his or her relationships with others.

To develop these descriptions, clinicians must make inferences from their test data. Although such data are objective and empirical, the process of developing hypotheses, obtaining support for these hypotheses, and integrating the conclusions is dependent on the theoretical knowledge and understanding, experience, and training of the clinician. This process generally follows a sequence of developing hypotheses, identifying relevant facts, making inferences, and supporting these inferences with relevant and consistent data. Wright (2010) conceptualized an eight-phase approach (Figure 1.1) for using data in a psychological assessment. It should be noted that, in actual practice, these phases are not as clearly defined as indicated in the figure, but often occur simultaneously. For example, when a clinician reads a referral question or initially observes a client, he or she is already developing hypotheses about that person and checking to assess the validity of these observations.

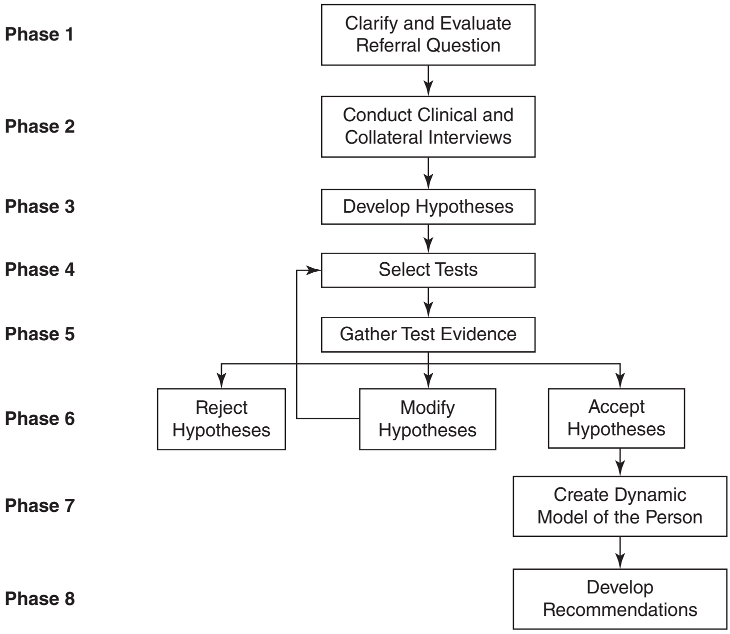

Figure 1.1 Hypothesis testing model for interpreting assessment data
Source: Adapted from Wright, 2010. Reprinted with permission from *Conducting Psychological Assessment: A Guide for Practitioners*, by A. J. Wright, Hoboken, NJ: Wiley.

Phase 1

The first phase, discussed above, is the clarification and evaluation of the referral question. As referral questions are one source of data, the clinician is already starting to develop hypotheses about what is going on for a client, what impact it has on his or her life, under what conditions the current problems developed, and even possible recommendations for how to improve the client's functioning and life in general.

Phase 2

Phase 2 focuses on collecting another source of data through clinical interviews and other background information (e.g., through collateral interviews, such as with parents or teachers, or through reviewing records or previous reports). Clinicians must understand the strengths and limitations of data collected from clinical interviews (see Chapter 3). It is from these data, though, that clearer initial hypotheses can be formed about the client's cognitive, emotional, personality, academic, neuropsychological, adaptive, and other areas of functioning.

Phase 3

Based on the information collected in Phases 1 and 2, the third phase focuses on developing hypotheses about what factors (situations, internal dynamics, etc.) may be causing and/or reinforcing whatever problems the client is having. These hypotheses require the clinician to have a firm grasp on many content areas of psychology, including personality theory, developmental psychology, abnormal psychology, developmental neurobiology, and even areas outside of psychology like biology, sociology, and cultural anthropology.

These hypotheses must be grounded in clear and logical clinical science and theory, regardless of theoretical orientation. For example, a hypothesis about the etiology of a client's low self-esteem may revolve around negative self-talk (from a cognitive-behavioral perspective) or the internalization of a mother's criticism (from a psychodynamic perspective). Regardless of theoretical orientation, the hypothesis must make sense within a specific psychological framework.

Phase 4

The importance of deliberateness when selecting tests to use in a specific assessment battery cannot be overstated. In addition to the considerations discussed earlier (see Table 1.1), the clinician must be confident that the tests selected can rule in or out the specific hypotheses generated in Phase 3 (as well as any modified hypotheses later on). Special attention should always be paid to cultural and sociodemographic characteristics of the client in order to ensure that the tests selected are appropriate, given the development, standardization, and norming procedures of the tests being considered.

Phases 5 and 6

Phase 5 centers on administering and scoring tests in order to collect data to evaluate the hypotheses generated in Phase 3. Phase 6, one of the most difficult phases, relates to the actual evaluation of test data within the context of the hypotheses generated previously. Phases 4 through 6 are iterative and recursive. As test data are collected, hypotheses can be rejected, modified, or accepted. Rejected hypotheses are abandoned, and the clinician can confidently move on to evaluating other hypotheses. Modified hypotheses may require the selection of new tests; while some tests may help develop modified hypotheses, additional tests are often necessary to actually evaluate these new hypotheses.

While rejecting and modifying hypotheses is often relatively straightforward, accepting hypotheses can be much more difficult, especially when it comes to personality or emotional functioning. It is often the case that a test or test score can rule *out* a hypothesis but cannot rule it *in*. For example, a high score on the Working Memory Index (WMI) of the WISC-V may rule out the presence of the inattentive subtype of attention-deficit/hyperactivity disorder (ADHD). This is because a child with ADHD would find it very difficult, if not impossible, to perform extremely well on WMI tasks that require both selective and sustained attention. However, a low score on the same WMI cannot rule ADHD *in*. Because multiple factors can affect performance on the WMI, more testing would be necessary to investigate the case of whether or not ADHD was present.

Phase 7

Phase 7 is a complicated phase requiring the clinician to make sense of all of the data collected in a way that can be clearly communicated to the client and/or referral source. Rather than presenting an acontextual list of a client's strengths and weaknesses or, even worse, presenting data test by test (which requires the audience to then determine which findings are important and connect the dots to make sense of the feedback), clinicians should create a dynamic understanding of how factors interact to explain what is happening for the client. To do this process well takes good training, supervision, and experience.

Phase 8

The final phase of the data interpretation process is linking the results to clear, specific, and reasonable recommendations that are likely to improve the client's life and functioning. Chapter 14 focuses on this process. In short, clinicians must understand treatment options from two different perspectives. First, clinically, clinicians must understand what is likely to link to and address the specific problems that emerged from the assessment, including the dynamics identified in Phase 7. Second, clinicians must understand the research behind interventions, how effective they have been shown to be, and what about them has been suggested or found to be the reasons that they are effective. Clinicians must consider both the empirical support of interventions and the likelihood of the interventions benefitting the specific client in his or her specific context and situation. Recommendations cannot be vague or broad, such as recommending "therapy" to a client. They should be both clear and specific. Additionally, they should be reasonable, given the circumstances. Although a specific treatment may be the best choice for a specific client, for a number of reasons, if that treatment is not available to the client (because of, for example, geographic location or financial limitations), then making a recommendation for that kind of treatment will not ultimately benefit the client.

RECOMMENDED READING

Aiken, L. R., & Groth-Marnat, G. (2006). *Psychological testing and assessment* (12th ed.). Boston, MA: Pearson Education.

Garb, H. N. (2005). Clinical judgment and decision making. *Annual Review of Psychology*, *1*, 67–89.

Groth-Marnat, G. (2000). Visions of clinical assessment: Then, now, and a brief history of the future. *Journal of Clinical Psychology*, *56*, 349–365.

Kubiszyn, T. W., Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., ... Eisman, E. J. (2000). *Empirical support for psychological assessment in clinical care settings. Professional Psychology*, *31*, 119–130.

Matarazzo, J. D. (1990). Psychological assessment versus psychological testing: Validation from Binet to the school, clinic, and courtroom. *American Psychologist*, *45*, 999–1017.

Meyer, G. J., Finn, S. E., Eyde, L., Kay, G. G., Moreland, K. L., Dies, R. R., ... Reed, G. M. (2001). Psychological testing and psychological assessment. *American Psychologist*, *56*, 128–165.

Wright, A. J. (2010). *Conducting psychological assessment: A guide for practitioners*. Hoboken, NJ: Wiley.