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.

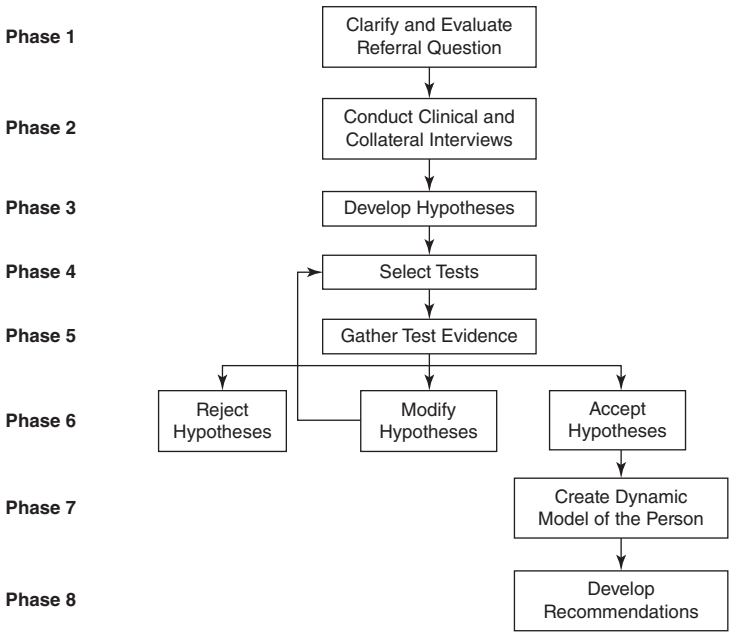

Figure 1.1 Hypothesis testing model for interpreting assessment data
Source: Adapted from Wright, 2010. Reprinted with permission from *Conducting Psychological Assessment: A Guide for Practitioners*, by A. J. Wright, Hoboken, NJ: Wiley.

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.

Chapter 2

CONTEXT OF CLINICAL ASSESSMENT

Although general knowledge regarding tests and test construction is essential, practitioners must consider a wide range of additional issues to situate testing procedures and test scores within an appropriate context. These considerations include clarifying the referral question, understanding the referral context, following ethical guidelines, identifying and working with test bias, selecting the most appropriate instrument for the variable or problem being studied, and making appropriate use of computer-assisted interpretation.

TYPES OF REFERRAL SETTINGS

Throughout the assessment process, practitioners should try to understand the unique problems and demands encountered in different referral settings. Otherwise, examiners—despite being skilled in administering and interpreting tests—might administer a needless series of tests and, at worst, provide useless information to referral sources and patients themselves. That is, a thorough investigation of the underlying motive for a referral can sometimes lead to the discovery that evaluation through testing is not even warranted.

Errors in test interpretation frequently occur because clinicians do not respond to the referral question in its broadest context. In turn, requests for psychological testing are often worded vaguely: "I would like a psychological evaluation on Mr. Smith," or "Could you evaluate Jimmy because he is having difficulties in school?" The request often does not state a specific question that must be answered or a decision that must be made, although many times this is the position that the referral source is in. For example, a school administrator may need testing to support a placement decision, a teacher may want to prove to parents that their child has a serious problem, or a psychiatric resident may not be comfortable with the management of a patient. An organization's surface motive for testing may be as vague as a statement that the procedure is a matter of policy. Greater clarification is necessary before clinicians can provide useful problem-solving information. Furthermore, many of these situations have hidden agendas that may not be adequately handled through psychological testing alone. One of the most useful questions in addressing these issues is to ask what decisions need to be made regarding the patient.

It must be stressed that the responsibility for exploring and clarifying the referral question lies with the clinician, who should actively work with the referral source to place the client's difficulty in a practicable context. Clinicians must understand the decisions that the referral source is facing, as well as the potential alternatives and

their possible implications. Clinicians also need to specify the potential utility of the psychological evaluation in determining different alternatives and their possible outcomes. They should make clear the advantages and usefulness of psychological testing but should also explain the limitations inherent in the process.

To help clarify the referral question, and to develop a relevant psychological evaluation, clinicians should become familiar with the types of environments in which they will be working. The most frequent environments are the psychiatric setting, the general medical setting, the legal context, the educational context, and the psychological clinic.

Psychiatric Setting

Levine (1981) summarized the important factors for a psychologist to be aware of in a psychiatric setting. These referrals typically come from a psychiatrist, who may be asking the referral question in the role of administrator, therapist, or physician. Each role presents unique issues for the psychiatrist, and clinicians have the primary responsibility to develop evaluations that directly address the problems at hand.

One of the main roles a psychiatrist fills is administrator on a ward. Ward administrators must frequently make decisions about problems such as suicide risk, admission/discharge, and the suitability of a wide variety of medical procedures. While retaining ultimate decision-making responsibility, psychiatrists often use information from other persons to help with decisions. This represents a change from the typical role of psychiatrists 40 years ago when they were mainly concerned with diagnosis and treatment. Currently, issues about custody, freedom of the patient, and the safety of society have taken over as the primary focus. From the perspective of psychologists performing assessments, this means that making a formal *DSM-5* (American Psychiatric Association, 2013) psychiatric diagnosis is usually not sufficient in and of itself. For example, a patient may be diagnosed bipolar, but this label does not indicate the level of dangerousness that the patient poses to him- or herself or to others. After patients have been admitted to a psychiatric setting, many practical questions have to be answered, such as the type of ward in which to place them, the activities in which they should be involved, and the method of therapy that would be most likely to benefit them.

Initially, the psychologist must determine exactly what information the ward administrator is looking for, particularly concerning any decisions that must be made about the patient. Psychologists in psychiatric settings who receive vague requests for "a psychological" sometimes develop a standard evaluation based on what they have learned about what this term implies on their specific unit. They may evaluate the patient's defense mechanisms, diagnosis, cognitive style, and psychosocial history without addressing the specific decisions that have to be made or perhaps covering only two or three relevant issues and omitting others. To maximize the usefulness of an evaluation, examiners must be especially aware of, and sensitive to, psychiatric administrators' legal and custodial responsibilities.

In contrast to the concerns of ward administrators, the standard referral questions from psychiatrists evaluating a patient for possible psychotherapy involve the appropriateness of the client for such therapy, the strategies that are most likely to be effective, and the likely outcome of therapy. These assessments are usually clear-cut and typically do not present many difficulties. Such evaluations can elaborate on likely problems

that may occur during the course of therapy, capacity for insight, diagnosis, coping style, level of resistance, degree of functional impairment, and problem complexity (see Chapter 14).

If a referral is made during therapy, however, a number of problem areas may exist that are not readily apparent from the referral question. The evaluator must investigate these complicating factors, along with potential decisions that may flow from the assessment information. An area of potential conflict arises when psychiatrists are attempting to fulfill roles of both administrator (caretaker) and psychotherapist and yet have not clearly defined these roles either for themselves or for their patients. The resulting ambiguity may cause the patient to feel defensive and resistant and the psychiatrist to feel that the patient is not living up to the therapist's expectations. Elaboration of a specific trait or need in the patient cannot resolve this conflict but must occur in the context of interactions between the therapist and the patient. A standard psychological evaluation investigating the internal structure of the patient will not address this issue.

A second possible problem area for clients referred in the midst of therapy can be the result of personal anxiety and discomfort on the therapist's part. Thus, issues such as therapist bias and possible unreasonable expectations may be equally or even more important than looking at a patient's characteristics. If role ambiguity, countertransference, bias, or unreasonable expectations are discovered, they must be elaborated and communicated in a sensitive manner.

When psychiatrists are acting in the role of physician, they and the psychologist may have different conceptual models for describing a patient's disorder. Whereas psychiatrists function primarily from a disease or medical model, psychologists may speak in terms of difficulties in living with people and society. In effectively communicating the results of psychological evaluations, examiners must bridge this conceptual difference. For example, a psychiatrist may ask whether a patient has a dissociative disorder, whereas a psychologist may not believe that the label dissociative disorder is useful or even a scientifically valid concept. The larger issue, however, is that the psychiatrist is still faced with some practical decisions. In fact, the psychiatrist may share some of the same concerns regarding dissociative disorders, but this conceptual issue may not be particularly relevant in dealing with the patient. Legal requirements or hospital policies might require that the patient be given a traditional diagnosis. The psychiatrist may also have to decide whether to give antipsychotic medication, electroconvulsive therapy, or psychotherapy. An effective examiner should be able to see beyond possible conceptual differences and instead address practical considerations. A psychiatrist may refer a defensive patient who cannot or will not verbalize his or her concerns and ask whether this person has schizophrenia. Beyond this diagnosis are factors such as the quality of the patient's thought processes and whether the person poses a danger to him- or herself or to others. Thus, the effective examiner must translate his or her findings into a conceptual model that is both understandable by a psychiatrist and useful from a task-oriented point of view.

General Medical Setting

It has been estimated that as many as two-thirds of patients seen by physicians have primarily psychosocial difficulties, and of those with clearly established medical diagnoses, between 25% and 50% have psychological disorders in addition to medical

ones (Asaad, 2000; Katon & Walker, 1998; McLeod, Budd, & McClelland, 1997; Mostofsky & Barlow, 2000). Most of these psychological difficulties are neither diagnosed nor referred for treatment (American Journal of Managed Care, 1999; Blount et al., 2007; Borus, Howes, Devins, & Rosenberg, 1988; Mostofsky & Barlow, 2000). In addition, many traditionally "medical" disorders, such as coronary heart disease, asthma, allergies, rheumatoid arthritis, ulcers, and headaches, have been found to possess a significant psychosocial component (Blount et al., 2007; Groth-Marnat & Edkins, 1996). Not only are psychological factors related to disease; of equal importance, they are related to the development and maintenance of health. In addition, the treatment and prevention of psychosocial aspects of "medical" complaints have been demonstrated to be cost-effective for areas such as preparation for surgery, smoking cessation, rehabilitation of chronic pain patients, obesity, interventions for coronary heart disease, and patients who are somatizing psychosocial difficulties (Blount et al., 2007; Chiles, Lambert, & Hatch, 1999; Groth-Marnat & Edkins, 1996; Groth-Marnat, Edkins, & Schumaker, 1995; Sobel, 2000). A complete approach to the patient, then, involves an awareness of the interaction among physical, psychological, and social variables (Kaslow et al., 2007; G. Schwartz, 1982). Thus, psychologists have the potential to make an extremely important contribution. To adequately work in general medical settings, psychologists must become familiar with medical descriptions, which often means learning a complex and extensive vocabulary (see J. D. Robinson & Baker, 2006). Another issue is that, even though they often draw information from several sources to aid in decision making, physicians must take ultimate responsibility for their decisions.

The most frequent situations in which physicians might use the services of a psychologist involve the presence of an underlying psychological disorder, possible emotional factors associated with medical complaints, assessment for neuropsychological deficit, psychological treatment for chronic pain, the treatment of chemical dependency, patient management, and case consultation (Bamgbose et al., 1980; Groth-Marnat, 1988; Pincus, Pechura, Keyser, Bachman, & Houtsinger, 2006). Regardless of whether a medical exam uncovers any physical basis for a patient's complaints, the physician still has to devise some form of treatment or at least an appropriate referral. This process is crucial in that a significant portion of patients referred to physicians do not have any detectable physical difficulties, and their central complaint is likely to be psychological (Asaad, 2000; Blount et al., 2007; Maruish & Nelson, 2014; Mostofsky & Barlow, 2000). The psychologist can then elaborate and specify how a patient can be treated for possible psychosocial difficulties (Kaslow et al., 2007; Wickramasekera, 1995a, 1995b). Doing this may require using not only the standard assessment instruments, but also more specialized ones, such as the Millon Behavioral Health Inventory or the Millon Behavioral Medicine Diagnostic (Bockian, Meagher, & Millon, 2000; Maruish, 2000; Millon, 1997).

Another area that has greatly increased in importance is the psychological assessment of a patient's neuropsychological status (see Chapter 12). Whereas physicians attempt to detect physical lesions in the nervous system, the neuropsychologist has traditionally been more concerned with the status of higher cortical functions. Another way of stating this is that physicians evaluate how the *brain* is functioning, while neuropsychologists evaluate how the *person* is functioning as a result of possible

brain abnormalities. The typical areas of assessment focus primarily on the presence of possible intellectual deterioration in areas such as memory, sequencing, abstract reasoning, spatial organization, and executive abilities (Groth-Marnat, 2000b). Such referrals, or at least screening for neuropsychological deficit, typically account for approximately one-third of all psychological referrals in psychiatric and medical settings. In the past, neuropsychologists have been asked to help determine whether a patient's complaints were "functional" or "organic." The focus now is more on whether the person has neuropsychological deficits that may contribute to or account for observed behavioral difficulties than on either/or distinctions (Loenberger, 1989). Physicians often want to know whether a test profile suggests a specific diagnosis, particularly malingering, conversion disorder, hypochondriasis, organic brain syndrome, or depression with pseudoneurological features. Further issues that neuropsychologists often address include the nature and extent of identified lesions, localization of lesions, emotional status of neurologically impaired patients, extent of disability, and suggestions for treatment planning such as recommendations for cognitive rehabilitation, vocational training, and readjustment to family and friends (Lemsky, 2000; Lezak, Howieson, Bigler, & Tranel, 2012; P. J. Snyder, Nussbaum, & Robins, 2006).

A physician might also request a psychologist to conduct a presurgical evaluation to assess the likelihood of a serious stress reaction to surgery. Finally, physicians—particularly pediatricians—are often concerned with detecting early signs of serious psychological disorder, which may have been brought to their attention by parents, other family members, or teachers. In such situations, the psychologist's evaluation should assess not only the patient's current psychological condition but also the contributing factors in his or her environment and should provide a prediction of the patient's status during the next few months or years. When the patient's current condition, current environment, and future prospects have been evaluated, the examiner can then recommend the next phase in the intervention process. A psychologist may also consult with physicians to assist them in effectively discussing the results of an examination with the patient or the patient's family.

Legal Context

During the past 40 years, the use of psychologists in legal settings has become more prevalent, important, and accepted (see Goldstein, 2007; Otto & Heilburn, 2002). Psychologists might be called in at any stage of legal decision making. During the investigation stage, they might be consulted to assess the reliability or quality of information presented by a witness. The prosecuting attorney might also need to have a psychologist evaluate the quality of another mental health professional's report, evaluate the accused person's competency, or help determine the specifics of a crime. A defense attorney might use a psychologist to help in supporting an insanity plea, to help in jury selection, or to document that brain damage has occurred. A judge might use a psychologist's report as one of a number of factors to help determine a sentence, a penal officer might wish consultation to help determine the type of confinement or level of dangerousness, or a parole officer might need assistance to help plan a rehabilitation program. Even though a psychologist might write a legal report, he or she is likely to actually appear in court in only about 1 in every 10 cases.

The increasing use and acceptance of psychologists in legal contexts have resulted in a gradual clarification of their roles (Goldstein, 2007; Otto & Heilburn, 2002) as well as a proliferation of forensic specific assessment instruments (Archer, 2006; Archer, Buffington-Vollum, Stredny, & Handel, 2006; Heilbrun, Marczyk, & Dematteo, 2002). However, acclimatizing to the courtroom environment is often difficult for multiple reasons, including the quite distinct differences between courtrooms and clinics, as well as the need to become familiar with specialized legal terms like *diminished capacity* and *insanity*. In addition, many attorneys are familiar with the same professional literature that psychologists read and may use this information to discredit a psychologist's qualifications, methods of assessment, or conclusions (Ziskin & Faust, 2008). Psychologists are also required to become increasingly sophisticated in their evaluation of possible malingering and deception (see review on kspope.com/assess/malinger.php).

Each psychologist appearing in court must have his or her qualifications approved. Important areas of consideration are the presence of clinical expertise in treating specialty disorders and relevant publication credits. Evaluation of legal work by psychologists indicates they are generally viewed favorably by the courts and may have reached parity with psychiatrists (Sales & Miller, 1994).

As outlined by the American Board of Forensic Psychology (www.abfp.com), the practice of forensic psychology includes training/consultation with legal practitioners, evaluation of populations likely to encounter the legal system, and the translation of relevant technical psychological knowledge into usable information. Psychologists are used most frequently in child custody cases, competency of a person to dispose of property, juvenile commitment, comprehension of Miranda rights, potential for having given a false confession, and personal injury suits in which the psychologist documents the nature and extent of the litigant's suffering or disability (e.g., stress, anxiety, cognitive deficit).

An essential requirement when working in the legal context is for psychologists to modify their language. Many legal terms have exact and specific meanings that, if misunderstood, could lead to extremely negative consequences. Words such as *incompetent, insane*, or *reasonable certainty* may vary in different judicial systems or from state to state. Psychologists must familiarize themselves with this terminology and the different nuances involved in its use. Psychologists may also be requested to explain in detail the meaning of their conclusions and how these conclusions were reached. Whereas attorneys rarely question the actual data that psychologists generate, the inferences and generalizability of these inferences are frequently placed under scrutiny or even attacked. Often this questioning can seem rude or downright hostile, but in most cases, attorneys are merely doing their best to defend their client. Proper legal protocol also requires that the psychologist answer questions directly rather than respond to the implications or underlying direction suggested by the questions. Furthermore, attorneys (or members of the jury) may not be trained in or appreciate the scientific method, which is the mainstay of a psychologist's background. In contrast, attorneys are trained in legal analysis and reasoning, which subjectively focus on the uniqueness of each case rather than on a comparison of the person to a statistically relevant normative group (see Hilsenroth & Stricker, 2004).

Two potentially problematic areas lie in evaluating insanity and evaluating competency. Although the insanity plea has received considerable publicity, very few people

make the appeal; and, of those who do, few have it granted. It is usually difficult for an expert witness to evaluate such cases because of the problem of possible malingering to receive a lighter sentence and the possible ambiguity of the term *insanity*. Usually a person is considered insane in accordance with the McNaughton Rule, which states that persons are not responsible if they did not know the nature and extent of their actions and if they cannot distinguish that what they did was wrong according to social norms. In some states, the ambiguity of the term is increased because defendants can be granted the insanity plea if it can be shown they were insane at the time of the incident. Other states include the clause of an "irresistible impulse" to the definition of insanity. Related to insanity is whether the defendant is competent to stand trial. *Competence* is usually defined as the person's ability to cooperate in a meaningful way with the attorney, understand the purpose of the proceedings, and understand the implications of the possible penalties. To increase the reliability and validity of competency and insanity evaluations, specialized assessment techniques have been developed; these include the MacArthur Competence Assessment Tool (Poythress et al., 1999), the Evaluation of Competency to Stand Trial–Revised (R. Rogers, Tillbrook, & Sewell, 2004), and the Rogers Criminal Responsibility Assessment Scales (R. Rogers, 1984).

The prediction of dangerousness has also been a problematic area. Because actual violent or self-destructive behavior is a relatively unusual (low base rate) behavior, any cutoff criteria will typically produce a high number of false positives (Mulvey & Cauffman, 2001). Thus, people incorrectly identified may potentially be detained and will understandably be upset. However, the negative result of failure to identify and take action against people who are potentially violent makes erring on the side of caution more acceptable. Attempts to use special scales on the Minnesota Multiphasic Personality Inventory (MMPI; Overcontrolled Hostility Scale; Megargee & Mendelsohn, 1962) or a 4-3 code type (see Chapter 7) have not been found to be sufficiently accurate for individual decision making. However, significant improvements have been made in predicting dangerousness and reoffending by assessing for the presence of antisocial features and using actuarial strategies, collateral sources, formal ratings, and summed ratings, which include relevant information on developmental influences, possible events that lower thresholds, arrest record, life situation, and situational triggers, such as interpersonal stress and substance intoxication (Monahan & Steadman, 2001; Monahan et al., 2000; Tolman & Rotzien, 2007). The legal/justice system is most likely to give weight to those individual assessment strategies that combine recidivism statistics, tests specifically designed to predict dangerousness, summed ratings, and double administrations of psychological tests to assess change over time. Frequently used tests include the Historical Clinical Risk–20 (for violence risk assessment; Webster, Douglas, Eaves, & Hart, 1997) and the Static 99 (for sexual reoffending risk; Hanson & Thornton, 1999). In contrast, informal clinical interviews are clearly considered to be insufficient (Tolman & Rotzien, 2007).

Psychologists are sometimes asked to help with child custody decisions. Guidelines for developing child custody evaluations and child protection evaluations have been developed by the American Psychological Association: (Guidelines for Child Custody Evaluations in Family Law Proceedings, 2010; www.apa.org/practice/guidelines/child-custody.aspx). The central consideration is to determine which arrangement is in the child's best interest. Areas to be considered include the mental health of the

parent, the quality of love and affection between the parent and child, the nature of the parent-child relationship, and the long-term potential effects of the different decisions on the child (M. J. Ackerman, 2006a, 2006b). Often psychological evaluations are conducted on each member of the family using traditional testing instruments. Specific tests, such as the Bricklin Perceptual Scales (Bricklin, 1984), have also been developed.

A final, frequently requested service is to aid in the classification of inmates in correctional settings. One basic distinction is between merely managing the person and attempting a program of rehabilitation. Important management considerations are levels of suicide risk, appropriateness of dormitory versus a shared room, possible harassment from other inmates, or degree of dangerousness to others. Rehabilitation recommendations may need to consider the person's educational level, interests, skills, abilities, and personality characteristics related to employment.

Academic/Educational Context

Psychologists are frequently called on to assess children who are having difficulty, or may need special placement, in the school system. The most important areas are evaluating the nature and extent of a child's learning difficulties, measuring intellectual strengths and weaknesses, assessing behavioral difficulties, creating an educational plan, estimating a child's responsiveness to intervention, and recommending changes in a child's program or placement (Sattler, 2008, 2014). Any educational plan should be sensitive to the interactions among a child's abilities, diversity considerations, the child's personality, the characteristics of the teacher, and the needs and expectations of the parents.

A typical educational placement begins with a visit to the classroom for observation of a child's behavior under natural conditions. A valuable aspect of this visit is to observe the interaction between the teacher and child. Often, behavioral difficulty is closely linked with the child–teacher interaction. Sometimes the teacher's style of responding to a student can be as much a part of the problem as the student. Consequently, classroom observations can cause discomfort to teachers and should be handled sensitively.

Observing the child in a wider context is, in many ways, contrary to the tradition of individual testing. However, individual testing all too frequently provides a relatively limited and narrow range of information, especially because children are not reliable self-reporters and parents or caregivers may be biased. If testing is combined with a family or classroom assessment, additional crucial data may be collected, though there may also be significant resistance. This resistance may result from legal or ethical restrictions regarding the scope of the services the school can provide or the demands that a psychologist can make on the student's parents. Often there is an initial focus on, and need to perceive, the student as a "problem child" or "identified patient." This focus may obscure larger, more complex, and yet more significant, issues, such as marital conflict, a disturbed teacher, misunderstandings between teacher and parents, or a conflict between the school and the parents. All or some of the involved individuals may have an investment in perceiving the student as the person with the problem, rather than acknowledging that a disordered school system or family difficulties may be responsible. An individually oriented assessment may be conducted with excellent

interpretations, but unless wider contexts are considered, understood, and addressed, the assessment may very well be ineffective in solving both the individual difficulties and the larger organizational or interpersonal problems.

Most assessments of children in a school context include behavioral observations, a test of intellectual abilities such as the Weschsler Intelligence Scale for Children–V, Stanford Binet–V, Woodcock-Johnson Psychoeducational Battery–IV (Woodcock, Schrank, Mather, & McGrew, 2014), or Kaufman Assessment Battery for Children–II (K-ABC-II; Kaufman & Kaufman, 2004), and tests of emotional and behavioral functioning. In the past, assessment of children's emotional functioning generally relied on projective techniques. However, many projective tests have been found to have inadequate psychometric properties and are time consuming to administer, score, and interpret. As a result, a wide variety of behavioral ratings instruments have begun to replace the use of projective instruments (Kamphaus, Petoskey, & Rowe, 2000). These include the Achenbach Child Behavior Checklist (Achenbach & Rescorla, 2001), Conners–3 Parent and Teacher Rating Scales (Conners, 2008), and the Behavior Assessment System for Children–3 (BASC-3; C. R. Reynolds & Kamphaus, 2015). A number of sound objective instruments, such as the Personality Inventory for Children–2 (PIC-2; Lachar & Gruber, 2001), have also been developed. This inventory was designed along similar lines as the MMPI but is completed by a child's parent. It produces four validity scales to detect faking and 12 clinical scales, such as Depression, Family Relations, Delinquency, Anxiety, and Hyperactivity. Assessment of adolescent personality can be done effectively with the MMPI-A or the Millon Adolescent Clinical Inventory (MACI; Millon, 1993). Additional well-designed scales that are increasingly used are the Vineland Adaptive Behavior Scales–II (Sparrow, Cicchetti, & Balla, 2005), the Wechsler Individual Achievement Test—III (WIAT-III; Pearson, 2009a), and the Wide Range Achievement Test–IV (WRAT-IV; Wilkinson & Robertson, 2007).

Any report written for an educational setting should focus not only on a child's weaknesses but also on his or her strengths. Understanding of a child's strengths can potentially be used to increase a child's self-esteem as well as to create change in a wide context. Recommendations should be realistic and practical. Recommendations can be developed most effectively when the clinician has a thorough understanding of relevant resources in the community, the school system, and the classroom environment. This understanding is particularly important because the quality and resources available in one school or school system can vary tremendously from another. Recommendations typically specify which skills need to be learned, how these can be learned, a hierarchy of objectives, and possible techniques for reducing behaviors that make learning difficult.

Recommendations for special education should be made only when a regular class would clearly not be equally beneficial. However, the recommendations are not the end product. They are beginning points that should be elaborated and modified depending on the initial results. Ideally, a psychological report should be followed up with continuous monitoring.

The psychoeducational assessment of children should be carried out in two phases. The first phase should assess the nature and quality of the child's learning environment. If the child is not exposed to adequate quality instruction, he or she cannot be

expected to perform well. Thus, it must first be demonstrated that a child is struggling despite appropriate instruction. The second phase involves a comprehensive assessment battery, which includes measures of intellectual abilities, academic skills, adaptive behavior, and screening out any biomedical disorders that might disrupt learning. Intellectual abilities might involve memory, spatial organization, abstract reasoning, and sequencing. Regardless of students' academic and intellectual abilities, they will not perform well unless they have relevant adaptive abilities, such as social skills, adequate motivation and attention, and ability to control impulses. Assessing a child's values and attitudes toward education may be particularly important because they determine whether the student is willing to use whatever resources he or she may have. Likewise, the person's level of personal efficacy helps to determine whether he or she is able to perform behaviors leading toward attaining the goals the person values. Physical difficulties that might interfere with learning include poor vision, poor hearing, hunger, extreme fatigue, malnutrition, or endocrine dysfunction.

The preceding considerations clearly place the assessment of children in educational settings into a far wider context than merely the interpretation of test scores. Relationships among the teacher, family, and student need to be assessed, along with the relative quality of the learning environment. Furthermore, the child's values, motivation, and sense of personal efficacy need to be taken into consideration, along with possible biomedical difficulties. Examiners need to become knowledgeable regarding the school and community resources as well as learn population-specific instruments that have demonstrated relatively high levels of reliability and validity.

Psychological Clinic

In contrast to the medical, legal, and educational institutions where the psychologist typically serves as a consultant to the decision maker, the psychologist working in a psychological clinic often is the decision maker. A number of frequent types of referrals come into the psychological clinic. Perhaps the most common ones are individuals who are self-referred and are seeking relief from psychological turmoil. For many of these individuals, extensive psychological testing may not be relevant and, in fact, may be contraindicated, as their diagnoses and issues may be relatively straightforward and time spent in testing could best be applied toward treatment. However, brief instruments targeted toward assessing client characteristics most relevant to treatment planning can help develop treatments that will speed the rate of improvement as well as optimize outcome (see Chapters 13 and 14). Brief instruments can also be used to monitor response to therapy or inform relevant alterations, thus increasing the likelihood of successful intervention (Lambert & Hawkins, 2004). In addition, there may be certain groups of self-referred clients about whom the psychologist may question whether the treatment available in a psychological clinic is appropriate. These clients can include persons with extensive medical problems, individuals with legal complications that need additional clarification, and persons who may require higher levels of care. With these cases, it might be necessary to obtain additional information through psychological testing. The main purpose of the testing would be to aid in decision making rather than to serve as a direct source of help for the client. Still other clients in clinics who may benefit from psychological testing are those who are being seen in the clinic

already, either who have unclear diagnoses or whose treatment has stalled or plateaued. These cases may benefit from the clear guidance of a comprehensive assessment.

Two other situations in which psychological assessment may be warranted involve children who are referred by their parents for school or behavioral problems and referrals from other decision makers. When referrals are made for poor school performance or involving legal complications, special precautions must be taken before testing. Primarily, the clinician must develop a complete understanding of the client's social network and the basis for the referral. This complete understanding may include a history of previous attempts at treatment and a summary of the relationship among the parents, school, courts, and child. Usually a referral comes at the end of a long sequence of events, and it is important to obtain information regarding these events. After the basis of the referral has been clarified, the clinician may decide to have a meeting with different individuals who have become involved in the case, such as the school principal, previous therapists, probation officer, attorney, or teacher. This meeting may uncover myriad issues that require decisions, such as referral for family therapy, placement in special education, a change in custody agreements between divorced parents, individual therapy of other members of the family, and a change in school. All of these issues may affect the relevance of, and approach to, testing, but they may not be apparent if the initial referral question is taken at face value. Sometimes psychologists are also confronted with referrals from other decision makers. For example, an attorney may want to know if an individual is competent to stand trial. Other referrals may involve a physician who wants to know whether a patient with a head injury can readjust to his or her work environment or drive a car, or the physician may need to document changes in a patient's recovery.

So far, this discussion on the different settings in which psychological testing is used has focused on when to test and how to clarify the manner in which tests can be most helpful in making decisions. Several additional summary points must be stressed. As has been discussed previously, a referral source sometimes is unable to adequately formulate the referral question. In fact, the referral question is usually neither clear nor concise. It is the evaluator's responsibility to look beyond the referral question and determine the basis for the referral in its widest scope. Thus, psychologists must develop an understanding of the complexity of the client's social setting, including interpersonal factors, family dynamics, and the sequence of events leading to the referral. In addition to clarifying the referral question, a second major point is that psychologists are responsible for developing knowledge about the setting for which they are writing their reports. This knowledge includes learning the proper language, the roles of the individuals working in the setting, the choices facing decision makers, and the philosophical and theoretical beliefs they adhere to. It is also important that clinicians understand the values underlying the setting and assess whether these values coincide with their own. For example, psychologists who do not believe in aversion therapy, capital punishment, or electroconvulsive therapy may come into conflict while working in certain settings. Psychologists, thus, should clearly understand how the information they give their referral source will be used. It is essential for them to appreciate that they have a significant responsibility, because decisions made regarding their clients, which are often based on assessment results, can frequently be major changing points in a client's life. If the possibility exists for the information to be used in a manner

that conflicts with the evaluator's value system, he or she should reconsider, clarify, or possibly change his or her relationship to the referral setting.

A final point is that clinicians should not allow themselves to be placed into the role of a "testing technician" or psychometrist. This role ultimately does a disservice to the client, the practitioner, and the profession. Clinicians should not merely administer, score, and interpret tests but should also understand the total referral context in its broadest sense. This means they also take on the role of an expert who can integrate data from a variety of sources. Tests, by themselves, are limited in that they are not flexible or sophisticated enough to address themselves to complex referral questions. D. Levine (1981) wrote:

> [The formal research on test validity is] not immediately relevant to the practical use of psychological tests. The question of the value of tests becomes not "Does this test correlate with a criterion?" or "Does the test accord with a nomological net?" but rather "Does the use of the test improve the success of the decision making process?" by making it either more efficient, less costly, more accurate, more rational, or more relevant. (p. 292)

All of these concerns are consistent with the emphasis on an evaluator fulfilling the role of an expert clinician performing psychological assessment rather than a psychometrist acting as a technician.

ETHICAL PRACTICE OF ASSESSMENT

Ethical guidelines reflect values that professional psychology endorses. These values include client safety, confidentiality, the reduction of suffering, fairness, and advancing science. These guidelines have largely evolved through careful consideration of how these values are expressed in ideal practice. Notably, many of the ethical codes have been refined due to conflicts and criticisms related to assessment procedures. Criticism has been directed at the use of tests in inappropriate contexts, confidentiality, cultural bias, invasion of privacy, release of test data, and the continued use of tests that are inadequately validated. These criticisms have resulted in restrictions on the use of certain tests, greater clarification within the profession regarding ethical standards, and, unfortunately, increased skepticism from the public. To deal with these potential difficulties as well as conduct useful and accurate assessments, clinicians need to be aware of the ethical use of assessment tools. The American Educational Research Association (AERA) and other professional groups have published guidelines for examiners in their *Standards for Educational and Psychological Tests* (1999) and the *Ethical Principles of Psychologists and Code of Conduct* (American Psychological Association, 2002). A special series in the *Journal of Personality Assessment* (Russ, 2001) also elaborated on ethical dilemmas found in training, medical, school, and forensic settings. The next section outlines the most important of these guidelines along with additional related issues.

Developing a Professional Relationship

Assessment should be conducted only in the context of a clearly defined professional relationship. This means that the nature, purpose, and conditions of the relationship

must be discussed and agreed on. Usually the clinician provides relevant information, followed by the client's signed consent. Information conveyed to the client usually relates to the type and length of assessment, alternative procedures, details related to appointments, the nature and limits of confidentiality, financial requirements, and additional general information that might be relevant to the unique context of an assessment (see Pope, 2007a, 2007b, and Zuckerman, 2003, for specific guidelines, formats, and forms for informed consent).

An important area to be aware of is the impact the quality of the relationship can have on both assessment results and the overall working relationship. It is the examiner's responsibility to recognize the possible influences he or she may exert on the client and to optimize the level of rapport. For example, enhanced rapport with older children (but not younger ones) involving verbal reinforcement and friendly conversation has been shown to increase WISC-R scores by an average of 13 IQ points compared with an administration involving more neutral interactions (Feldman & Sullivan, 1971). This is a difference of nearly 1 full standard deviation. It has also been found that mildly disapproving comments, such as "I thought you could do better than that," resulted in significantly lowered performance when compared with either neutral or approving comments (Witmer, Bornstein, & Dunham, 1971). In a review of 22 studies, Fuchs and Fuchs (1986) concluded that, on average, IQ scores were 4 points higher when the examiner was familiar with the child being examined than when he or she was unfamiliar with the child. This trend was particularly pronounced for children from lower socioeconomic status. Whereas there is little evidence (Lefkowitz & Fraser, 1980; Sattler, 1973a, 1973b; Sattler & Gwynne, 1982) to support the belief that African American students have lower performance when tested by European American examiners, it has been suggested that African American students are more responsive to tangible reinforcers (money, candy) than are European American students, who generally respond better to verbal reinforcement (Schultz & Sherman, 1976). However, in a later study, Terrell, Taylor, and Terrell (1978) demonstrated that the main factor was the cultural relevance of the response. They found a remarkable 17.6-point increase in IQ scores when African American students were encouraged by African American examiners with culturally relevant comments such as "nice job, blood" or "good work, little brother." Thus, positive rapport and feedback, especially if that feedback is culturally relevant, can significantly improve test performance. As a result, the feedback, and level of rapport should, as much as possible, be held constant from one test administration to the next.

A variable extensively investigated by Rosenthal and his colleagues is that a researcher/examiner's expectations can influence another person's level of performance (R. Rosenthal, 1966). This finding has been demonstrated with humans as well as with laboratory rats. For example, when an experimenter was told to expect better performances from rats that were randomly selected from the same litter as "maze bright" (compared with "maze dull"), the descriptions of the rats' performance given by the experimenter conformed to the experimenter's expectations (R. Rosenthal & Fode, 1963). Despite criticisms that have been leveled at his studies and the finding that the magnitude of the effect was not as large as originally believed (Barber & Silver, 1968; Elashoff & Snow, 1971), Rosenthal maintains that an expectancy effect exists in some situations and suggests that the mechanisms include minute, nonverbal

behaviors (H. Cooper & Rosenthal, 1980). He maintains that the typical effects on an individual's performance are usually small and subtle and occur in some situations but not in others. The obvious implication for clinicians is that they should continually question themselves regarding their expectations of clients and check to see whether they may in some way be communicating these expectations to their clients in a manner that confounds the results.

An additional factor that may affect the nature of the relationship between the client and the examiner is the client's relative emotional state. It is particularly important to assess the degree of the client's motivation and his or her overall level of anxiety. There may be times in which it would be advisable to discontinue testing because situational emotional states may significantly influence the results of the tests. At the very least, examiners should consider the possible effects of emotional factors and incorporate these into their interpretations. For example, it might be necessary to increase the estimate of a client's optimal intellectual functioning if the client were obviously extremely anxious during administration of an intelligence test.

A final consideration, which can potentially confound both the administration and, more commonly, the scoring of responses, is the degree to which the examiner likes the client and perceives him or her as warm and friendly. Several studies (Sattler, Hillix, & Neher, 1970; Sattler & Winget, 1970) have indicated that the more the examiner likes the client, the more likely he or she will be to score an ambiguous response in a direction favorable to the client. Higher scores can occur even on items in which the responses are not ambiguous (Egeland, 1969; Simon, 1969). Thus, "hard" scoring, as opposed to more lenient scoring, can occur at least in part because of the degree of subjective liking the examiner feels toward the client. Again, examiners should continually check themselves to assess whether their relationship with the client is interfering with the objectivity and standardization of the test administration and scoring.

Issues Related to Informed Consent

Psychologists should obtain informed consent for assessment procedures. Any consent involves a clear explanation of what procedures will occur, the relevance of the testing, and how the results will be used (see Pope, 2007a; kspope.com/consent/index.php). This means that examiners should always have a clear conception of the specific reasons for giving a test. It should be stressed what information is confidential and what limitations to confidentiality exist. Exceptions to confidentiality may occur in situations involving child/elder abuse, danger to self or others, and information that has been requested based on a subpoena. The information should be provided in clear, straightforward language that can be understood by the client. Unfortunately, many formal consent forms are written at a level far above the reading comprehension level of a large proportion of clients.

Informed consent involves communicating not only the rationale for testing but also the kinds of data obtained and the possible uses of the data, when this will not adversely affect the results. This fact does not mean the client should be shown the specific test subscales beforehand, but rather that the nature and intent of the test should be described in a general way. For example, if a client is told that a scale measures "sociability," this foreknowledge might alter the test's validity in that the client

may answer questions based on popular, but quite possibly erroneous, stereotypes. Introducing the test format and intent in a simple, respectful, and forthright manner significantly reduces the chance that the client will perceive the testing situation as an invasion of privacy.

Sometimes clinicians will have provided clear information and the client will have agreed to the procedures, but unforeseen events not covered in the information may occur. This might happen when the examiner discovers aspects of the client that the client would rather keep secret. Thus, assessment may entail an invasion of privacy. The Office of Science and Technology (1967), in a report entitled "Privacy and Behavioral Research," defined privacy as "the right of the individual to decide for him/herself how much he will share with others his thoughts, feelings, and facts of his personal life." This right is considered to be "essential to insure dignity and freedom of self determination" (p. 2). The invasion of privacy issue usually becomes most controversial with personality tests, as items relating to motivational, emotional, and attitudinal traits are often disguised. Thus, persons may unknowingly reveal characteristics about themselves that they would rather keep private. Similarly, many individuals consider their IQ scores to be highly personal. Public concern over this issue culminated in an investigation by the Senate Subcommittee on Constitutional Rights and the House Subcommittee on Invasion of Privacy. Neither of these investigations found evidence of deliberate or widespread misuse of psychological tests (Brayfield, 1965).

Dahlstrom (1969) argued that public concern over the invasion of privacy is based on two basic issues. The first is that tests have been oversold to the public, with a resulting exaggeration of their scope and accuracy. The public is usually not aware of the limitations of test data and may often feel that tests are more capable of discovering hidden information than they actually are. The second misconception is that it is not necessarily wrong to obtain information about persons that they either are unaware of themselves or would rather keep private. The more important issue is how the information is used. Furthermore, the person who controls where and how this information is used is generally the client. The ethical code of the American Psychological Association (2002) specifically states that information derived by a psychologist from any source can be released only with the permission of the client. Although there may be exceptions regarding the rights of minors or when clients are a danger to themselves or others, the ability to control the information is usually clearly defined as being held by the client. Thus, the public can be uneducated regarding its rights and can underestimate the power it has in determining how the test data will be used.

Whereas concerns about invasion of privacy relate to the discovery and misuse of information that clients would rather keep secret, *inviolacy* involves the actual negative feelings created when clients are confronted with the test or test situation. Inviolacy is particularly relevant when clients are asked to discuss information they would rather not think about. For example, the MMPI contains questions about many ordinarily taboo topics relating to sexual practices, toilet behavior, bodily functions, and personal beliefs about human nature. Such questions may produce anxiety by making examinees think about deviant thoughts or repressed unpleasant memories. Many individuals obtain a certain degree of security and comfort by staying within familiar realms of thought. Even to be asked questions that may indicate the existence of unusual alternatives can serve as an anxiety-provoking challenge to personal rules

and norms. This problem is somewhat related to the issue of invasion of privacy, and it too requires one-to-one sensitivity as well as clear and accurate disclosure about the assessment procedure.

Another issue is that during personnel evaluations, participants might feel pressured to reveal personal information on tests because they aspire to a certain position. Also, applicants may unknowingly reveal information because of subtle, nonobvious test questions, and, perhaps more importantly, they have no control over the inferences that examiners make about the test data. However, if a position requires careful screening and if serious negative consequences may result from poor selection, it is necessary to evaluate an individual as closely as possible. Thus, careful testing may be required to select personnel in the police, in delicate military positions, or for important public duty overseas.

In a clinical setting, obtaining personal information regarding clients usually does not present problems. The agreement that the information be used to help clients develop new insights and change their behavior is generally clear and straightforward. However, should legal difficulties arise relating to areas such as child abuse, involuntary confinement, or situations in which clients may be a danger to themselves or others, ethical questions often arise. Usually there are general guidelines regarding the manner and extent to which information should be disclosed. These are included in the American Psychological Association's *Ethical Principles of Psychologists and Code of Conduct* (2002), and test users are encouraged to familiarize themselves with these guidelines. Professional psychologists can also consult with colleagues, their insurance companies, or the APA's ethics office (apa.org/ethics).

Labeling and Restriction of Freedom

When individuals are given a medical diagnosis for physical ailments, the social stigmata are usually relatively mild. In contrast are the potentially damaging consequences of many psychiatric diagnoses. A major danger is the possibility of creating a self-fulfilling prophecy based on the expected roles associated with a specific label. Many of these expectations are communicated nonverbally and are typically beyond a person's immediate awareness (H. Cooper & Rosenthal, 1980; R. Rosenthal, 1966). Other self-fulfilling prophecies may be less subtle; for example, a juvenile with minor but poor sexual boundaries might be labeled as a "sex offender," which would then result in quite restrictive treatment along with quite public distribution of the label.

Just as labels imposed by others can have negative consequences, self-acceptance of labels can likewise be detrimental. Clients may use their labels to excuse or deny responsibility for their behavior. This is congruent with the medical model, which usually assumes that a "sick" person is the victim of an "invading disorder." Thus, in our society, "sick" persons are not considered responsible for their disorders. However, the acceptance of this model for behavioral problems may perpetuate behavioral disorders because persons see themselves as helpless, passive victims under the power of mental health "helpers" (Szasz, 1987). This sense of helplessness may serve to lower people's ability to deal effectively with new stress. In contrast to this sense of helplessness is the belief that clients require an increased sense of responsibility for their lives and actions to effectively change their behavior.

A final difficulty associated with labeling is that it may unnecessarily impose limitations on either an individual or a system by restricting progress and creativity. For example, an organization may conduct a study to determine the type of person who has been successful at a particular type of job and may then develop future selection criteria based on this study. This can result in the future selection of relatively homogeneous employees, which in turn could prevent the organization from changing and progressing. There may be a narrowing of the "talent pool," in which people with new and different ideas are never given a chance. In other words, what has been labeled as adaptive in the past may not be adaptive in the future. One alternative to this predicament is to look at future trends and develop selection criteria based on these trends. Furthermore, diversity might be incorporated into an organization so that different but compatible types can be selected to work on similar projects. Thus, clinicians should be sensitive to the potential negative impact resulting from labeling by outside sources or by self-labeling, as well as to the possible limiting effects that labeling might have.

Competent Use of Assessment Instruments

To correctly administer and interpret psychological tests, an examiner must have proper training, which generally includes adequate graduate coursework combined with lengthy supervised experience (Turner, DeMers, Fox, & Reed, 2001). Clinicians should have knowledge of tests and test limitations and should be willing to accept responsibility for competent test use. Intensive training is particularly important for individually administered intelligence tests and for the majority of personality tests. Students who are administering tests as part of a class requirement are generally not yet adequately trained to administer and interpret tests professionally. Thus, test results obtained by students have questionable validity, and students should clearly inform their subjects that the purpose of their testing is for training purposes only.

In addition to the preceding general guidelines for training, examiners should also acquire a number of specific skills (Moreland, Eyde, Robertson, Primoff, & Most, 1995; Turner et al., 2001). These include the ability to evaluate the technical strengths and limitations of a test, the selection of appropriate tests, knowledge of issues relating to the test's reliability and validity, and interpretation with diverse populations. Examiners need to be aware of the material in the test manual as well as relevant research both on the variable the test is measuring and the status of the test since its publication. This is particularly important with regard to newly developed subgroup norms and possible changes in the meaning of scales resulting from further research. After examiners evaluate a test itself, they must also be able to evaluate whether the purpose and context for which they would like to use it are appropriate. Sometimes an otherwise valid test can be used for purposes it was not intended for, resulting in either invalid or useless inferences based on the test data. In addition, examiners must be continually aware of, and sensitive to, conditions affecting the examinee's performance. These conditions may include expectations on the part of the examiner, minor variations from the standardized instructions, degree of rapport, mood of the examinee, or timing of the test administration in relation to an examinee's life changes. To help develop accurate conclusions, examiners should have a general knowledge of the diversity of human behavior. Different considerations and interpretive strategies may be necessary

for various ethnic groups, sexes, sexual orientations, or persons from different countries (see Dana, 2005; Nguyen, Huang, Arganza, & Liao, 2007). A final consideration is that, if interns or technicians are administering the tests, an adequately trained psychologist should be available as a consultant or supervisor.

Specific data-based guidelines for test user qualifications have been developed by relevant professional organizations (American Psychological Association, 1987; Turner et al., 2001), and these guidelines have been incorporated by most organizations selling psychological tests. Qualification forms request information regarding the purpose for using tests (counseling, research, personnel selection), area of professional expertise (marriage and family, social work, school), level of training (degrees, licenses), specific courses taken (descriptive statistics, career assessment), and quality control over test use (test security, appropriate tailoring of interpretations). Persons completing the forms certify that they possess appropriate training and competencies and agree to adhere to ethical guidelines and legal regulations regarding test use.

In addition to being appropriately trained to use tests themselves, psychologists should not promote the use of psychological techniques by persons who are not qualified. This does not mean that all psychological tests should be used exclusively by psychologists, as many tests are available to other professionals. However, psychologists should be aware of which tests require a high level of training (e.g., individually administered IQ tests) and those that are more generally available.

One of the important aspects of competent test use is that tests should be used only for the purposes they were designed for. Typically, tests being extended beyond what they were designed for have been done in good faith and with good intentions. For example, an examiner might use a Thematic Apperception Test or Rorschach as the primary means of inferring an individual's IQ. Similarly, the MMPI-2 or MCMI-IV, which were designed to assess the extent of psychopathology in an individual, might be inappropriately used to assess a normal person's level of functioning. Although some conclusions can be drawn from the MMPI-2 relating to certain aspects of a normal person's functioning, and although IQ estimates based on projectives can be made, they should be considered extremely tentative. These tests were not designed for these purposes, and, as a result, such inferences do not represent the strengths of the tests. A somewhat more serious misuse can occur when a test such as the MMPI-2 is used to screen applicants for some types of personnel selection. Results from MMPI-2-type tests are likely to be irrelevant for assessing most job-related skills. Of equal importance is that the information derived from the MMPI-2 is typically of a highly personal nature and, if used in many types of personnel selection, is likely to represent an invasion of privacy.

Interpretation and Use of Test Results

Interpreting test results should never be considered a simple, mechanical procedure. Accurate interpretation means not simply using norms and cutoff scores but also taking into consideration unique characteristics of the person combined with relevant aspects of the test itself. Whereas tests themselves can be validated, the integration of information from a test battery is far more difficult to validate. It is not infrequent, for example, to have contradictions among different sources of data. It is up to the clinician

to evaluate these contradictions to develop the most appropriate, accurate, and useful interpretations. If the clinician has significant reservations regarding the test interpretation, these should be communicated, usually in the psychological report itself.

A further issue is that test norms and stimulus materials eventually become outdated. As a result, interpretations based on these tests may become inaccurate. For this reason, clinicians need to stay current on emerging research and new versions of tests. A rule of thumb is that if a clinician has not updated his or her test knowledge in at most the past 10 years, he or she is probably not practicing competently.

Part of remaining current means that psychologists should select their testing instruments, as well as any scoring and interpretation services, based on evidence related to the validity of the programs or tests. Part of this selection process requires knowledge of the context of the situation (Turner et al., 2001). A well-validated test might have been found to be quite valid in one context or population but not for another. Another issue that might have ethical considerations is conversion to or use of computerized or Internet-assisted technology (McMinn, Bearse, Heyne, Smithberger, & Erb, 2011; McMinn, Buchanan, Ellens, & Ryan, 1999; McMinn, Ellens, & Soref, 1999). Ultimately, any interpretations and recommendations regarding a client are the responsibility of the clinician. Placing a signature on a report means that the clinician is taking responsibility for the content of the report. Indeed, an important difference between an actuarial formula or automated report and a practitioner is that the practitioner ultimately will be held accountable.

Communicating Test Results

Psychologists should ordinarily give feedback to the client and referral source regarding the results of assessment (Lewak & Hogan, 2003; see Pope, 1992, 2007b, and on kspope.com/assess/feedabs1.php for forms and guidelines). This feedback should be given using clear, everyday language. If the psychologist is not the person giving the feedback, this should be agreed on in advance, and the psychologist should ensure that the person providing the feedback presents the information in a clear, competent manner. Unless the results are communicated effectively, the purpose of the assessment is not likely to be achieved. Effective feedback involves understanding the needs and vocabulary of the referral source, client, and other persons, such as parents or teachers, who may be affected by the test results. Initially, there should be a clear explanation of the rationale for testing and the nature of the tests being administered. This explanation may include the general type of conclusions that are drawn, the limitations of tests, and common misconceptions surrounding the tests or test variables. If a child is being tested in an educational setting, a meeting should be arranged with the school psychologist, parents, teacher, and other relevant persons. Such an approach is crucial for IQ tests, which are more likely to be misinterpreted than achievement tests. Assessment results feedback should be given in terms that are clear and understandable to the receiver. Descriptions are generally most meaningful when performance levels are clearly indicated along with behavioral references. For example, in giving IQ results to parents, it is only minimally relevant to say that their child has an IQ of 130 with relative strengths in spatial organization, even though this may be appropriate language for a formal psychological evaluation. A more effective description might be that their

child is "currently functioning in the top 2% when compared with his or her peers and is particularly good at organizing nonverbal material, such as piecing together puzzles, putting together a bicycle, or building a playhouse."

In providing effective feedback, the clinician should also consider the personal characteristics of the receiver, such as his or her general educational level, relative knowledge regarding psychological testing, and possible emotional response to the information (Finn, 2007). The emotional reaction is especially important when a client is learning about his or her personal strengths or shortcomings. Facilities should be available for additional counseling, if needed. If properly given, feedback is not merely informative but can actually serve to reduce symptomatic distress and enhance self-esteem (Armengol, Moes, Penney, & Sapienza, 2001; Finn & Tonsager, 1992; Lewak & Hogan, 2003). Thus, providing feedback can actually be part of the intervention process itself. Because psychological assessment is often requested as an aid in making important life decisions, the potential impact of the information should not be underestimated. Clinicians are usually in positions of power, and with that power comes responsibility, as the information clients receive and the decisions they make based on this information are often with them for many years.

Maintenance of Test Security and Release of Test Data

If test materials were widely available, it would be easy for persons to review the tests, learn the answers, and respond according to the impression they would like to make. Thus, the materials would lose their validity. Not only is maintaining test security an ethical obligation, but it is a legal requirement related to trade secrets and agreements made with test publishers when materials are purchased. Psychologists should make all reasonable efforts to ensure that test materials are secure. Specifically, all tests should be kept locked in a secure place, and no untrained persons should be allowed to review them. Any copyrighted material should not be duplicated (see Zuckerman, 2003, for forms and guidelines).

The security of assessment results should also be maintained. This security usually means that only persons designated by the client (often the referral source and client) should see the results. In reality, however, this ethical principle may sometimes be difficult to achieve. For example, many medical contexts expect most relevant treatment information (including psychological assessment results) to be kept in clients' charts. Typically, all members of the treatment team have access to the charts (Claassen & Lovitt, 2001). On one level, this access represents a conflict between psychological and medical guidelines. On another level, it represents a conflict between benefit to the patient (that may be enhanced by the treatment team having access to his or her records) and patient autonomy (patient control over to whom and where information should go). Security of assessment results can also be compromised when a large number of organizations (e.g., insurance company, interacting rehabilitation provider, referral source) all want access to patient records. This issue arises frequently in the managed care environment. The security of client records also becomes more tenuous when large interconnected databases potentially have access to patient data (McMinn, Bearse et al., 2011; McMinn, Buchanan, et al., 1999; McMinn, Ellens et al., 1999).

In some clinical and legal contexts, the court or the opposing counsel may wish to see actual client data. These data can be released if the client authorizes it or if the material has been subpoenaed. Ideally, however, the examiner should recommend that a qualified person be present to explain the results. This recommendation is consistent with the principle that the examiner protect the client from potential harm. If the examiner feels that releasing the test data may result in "substantial harm" to the client or "misuse or misrepresentation of the data" (American Psychological Association, 2002, p. 12), he or she may have the option of refusing to release the data. This situation may result in a conflict between legal and ethical requirements.

One important distinction is between "test data" and "test materials." The term *test data* refers to raw and scaled scores, such as subscale scores and test profiles. In contrast, *test materials* refers to "manuals, instruments, protocols, and test questions or stimuli" (American Psychological Association, 2002, p. 13). Interestingly, test materials turn into test data when a psychologist places the client's name on the materials. Since actual items should not be released, it is important for clinicians to make sure they do not place client-identifying information on what might be copyrighted or restricted material. This is crucial since psychologists can release test data, but they cannot release test materials (e.g., actual test items). As stated, the release of test materials would constitute a breach of trade secrets, copyright, and the conditions of purchase (Behnke, 2004). One exception to this point is that the material may be released to persons who are properly qualified (Tranel, 1994). Another exception is when a subpoena specifically squashes these terms of purchase, copyright, and trade secrets.

ASSESSING DIVERSE GROUPS

Competence in assessing diverse groups is an essential part of professional practice. This fact is highlighted by increased globalization, extensive immigration, controversies over potential test bias when used with diverse groups, cross-national adaptation of common instruments, and the American Psychological Association's requirement that professional psychologists be trained to work with diverse groups. In the United States, as of 2013, more than one-third of the population was classified as members of an ethnic minority group (United States Census Bureau, 2015). Many minority populations in the United States are underrepresented and underserved (A. Levine, 2007). Thus, it is crucial that guidelines for competent assessment be developed. The guidelines pertain to language skills, cultural competency, assessing cultural/racial identity, appropriate use of instruments, diagnostic issues, and interpretation guidelines (see Dana, 2005).

Language Skills

Evaluating a client's language proficiency is a first step in assessing diverse clients. Based on this evaluation, it may be necessary, or at least advisable, to conduct the assessment in the client's native language. A sufficiently knowledgeable clinician can conduct the assessment him- or herself. Sometimes a translator or referral to another

clinician who speaks the language may be required. If the client is reasonably proficient in English, then it may be possible to conduct the assessment in English. However, clinicians should be aware of how this might alter the interaction and must take these potential differences into account when interpreting test scores. For example, a client who is struggling with English may appear to be uncooperative or to have flat affect when in reality this impression is created primarily because of language difficulties. It may also be advisable to use assessment instruments that have been translated into the client's native language.

Cultural Competency

Cultural competency on the part of clinicians begins with self-exploration of personal histories, attitudes, and knowledge. Doing this involves clinicians understanding their exposure to various cultures, biases about various cultures, and the degree of comfort with these cultures. It is natural to feel more resonant with some cultures as opposed to others. Often attitudes can be subtle and unconscious; for example, clinicians might have a sense of white privilege yet may have difficulty acknowledging these feelings. These attitudes are typically transmitted through nonverbal and subtle means.

Based on personal exploration and knowledge of a culture, clinicians need to develop optimal strategies of service etiquette. One strategy may involve level of formality. For example, Native Americans are likely to be more comfortable with minimal formality whereas Asian Americans usually expect more formal interactions characterized by a logical, structured approach. Other factors are the extent of eye contact, physical proximity, volume of voice, and the extent to which emotions are conveyed. For example, some cultures defer to persons perceived as being of higher status by decreasing the volume of their voice and minimizing eye contact. Clinicians who are knowledgeable about these differences should be both accepting of them and not misinterpret these behaviors as indicating depression or evasiveness. At the same time, these behaviors may make it more difficult to detect depression when it is actually present. A further variable is the time involved prior to the client becoming self-disclosing. Some cultures expect extensive preliminaries, perhaps to the point of having mutual acquaintances approve of the clinician prior to more formal clinical work. In contrast, other cultures are quite comfortable with becoming more self-disclosing and "task oriented" with minimal preliminaries. Taking into account each of these factors may make the difference between developing good rapport with accurate assessment results versus poor rapport resulting in inaccurate assessment data.

Cultural/Racial Identity

Cultural identity is a crucial aspect of explaining thoughts, feelings, and behaviors. It is thus important to understand this fact when conducting individual assessment. However, cultural identity varies according to the extent that a person identifies with his or her culture. Some individuals have quite strong identifications with their cultures. As a result, careful consideration of whether standard tests are appropriate to use with them is required. It may be necessary to use test translations, different norms, translators, or instruments specific to their culture. Further, cultural identity should be taken

into account when interpreting test scores. In contrast, other clients might have early experience with a culture but have later become quite acculturated into the dominant culture. As a result, standard tests might be used with more confidence.

Level of identity can be assessed informally through interview. There are also a variety of more formal instruments that ask questions related to variables such as language proficiency/preference, religious beliefs, foods, family structure, value orientation, socioeconomic status, collectivism/individualism, and culture-specific traditions, customs, and identifications. A sample of frequently used measures follows (see review by Dana, 2005):

African Americans: African American Acculturation Scale (Landrine & Klonoff, 1994)
Asian Americans: Asian Values Scale (Kim, Atkinson, & Yang, 1999)
Hispanic/Latinos: Acculturation Rating Scale for Mexican Americans (Cuellar, Arnold, & Maldonado, 1995)
Native Americans and Alaska Natives: Northern Plains Bicultural Immersion Scale (Allen, 1998)

One caution with these instruments is that sometimes individuals have quite different origins within the general group the instrument is trying to measure. This is particularly true for Hispanics and Asian Americans. For example, there are significant differences between Hispanics from Mexico and those from Argentina. Similarly, Japanese, Chinese, Koreans, and Hmong have many differences between their cultures. Despite this fact, a scale such as the Asian Values Scale is at least a start at looking at some of the common cultural values of these groups.

Test Equivalence and Appropriate Use of Instruments

Whether an instrument is culturally appropriate is based on a number of considerations, including the client's level of acculturation, language preference, language proficiency, availability of translations of the instrument, whether the construct is the same for the client's culture, availability of norms, and availability of possibly more appropriate alternatives specific to the client's culture. At the core of whether or not the test is appropriate is evaluating the equivalence of the test. Equivalence can be organized according to linguistic, conceptual, and metric equivalence. (See Table 2.1.)

If a test is not equivalent, it may result in bias against the group or individual it is evaluating. The term *bias in testing* refers to the presence of systematic error in the measurement of certain factors (e.g., academic potential, intelligence, psychopathology) among certain individuals or groups (Suzuki & Ponterotto, 2007). The possible presence of bias toward minority groups has resulted in one of the most controversial issues in psychological testing. More specifically, critics believe that psychological tests are heavily biased in favor of, and reflect the values of, European American, middle-class society. They argue that such tests cannot adequately assess intelligence or personality when applied to minority groups. Whereas the greatest controversy has arisen from the use of intelligence tests, the presence of cultural bias is also relevant in the use of personality testing. Over the past 20 years, discussion of bias has shifted from controversy

Table 2.1 Summary of Test Equivalence

Type	Definition	Issues/Strategies
Linguistic	Wording and content	Translate into new language and then retranslate again ("back-translate"), consider idioms and pictures, *adapt* not merely literal translation
Conceptual	Construct has same meaning	Same as construct validity, makes similar predictions, correlation, and factor analysis
Metric	Same psychometric features	Distributions, ranges, stability, comparable reliability, and validity, do scores mean the same things

over the nature and extent of bias to a more productive working through of how to make the most valid and equitable assessment based on current knowledge (see Dana, 2005; Geisinger, 2003; Handel & Ben-Porath, 2000).

The original controversies over test bias centered on determining whether tests are as valid for minority groups as for nonminorities. Differences often do exist in mean test scores; however, the meaning that can be attributed to these differences has been strongly debated. The major question lies in identifying the cause of these differences. Differences in test scores could stem from genuine differences in ability, which could be the result of environmental factors (Kamin, 1974; R. Rosenthal & Jacobson, 1968) or actual hereditary determination (A. R. Jensen, 1969, 1972; Rushton, 1994), or they could be artifacts of tests that are inherently biased. Although the debate is not resolved, guidelines have been established by the Equal Employment Opportunity Commission (EEOC, 1970) for the use of psychological tests with minority groups in educational and industrial settings. The basic premise is that a screening device (psychological test) can have an adverse impact if it screens out a proportionally larger number of minorities than nonminorities. Furthermore, it is the responsibility of the employer to demonstrate that the procedure produces valid inferences for the specific purposes for which the employer would like to use it. If an industrial or educational organization does not follow the guidelines as defined by the EEOC (1970), the Office of Federal Contract Compliance has the direct power to cancel any government contract that the institution might have.

Linguistic Equivalence

As summarized in Table 2.1, the first area of concern is *linguistic equivalence*, which is whether the test has been translated accurately. On the surface, this may mean simply translating the administration instructions and test items into the language of interest. One strategy to assist with this is to use "back-translation." In back-translation, once the test is translated, it is then translated back into the original language. If the meanings of the items are still the same, then the back-translation helps to ensure that the translation is conceptually adequate. A further issue is that sometimes idioms need to be comparable. Similar to this issue is that not only verbal materials but also pictures should be made comparable. For example, a picture of a stereotypically appearing person depicted in one culture should be similarly made to look stereotypical in the culture

the test has been translated into. Doing this goes beyond merely translating the test and into an "adaptation" of the test (sometimes referred to as "functional equivalence").

Conceptual Equivalence

A further concern is *conceptual equivalence*, which requires the constructs to have the same meaning in various cultures. Sometimes the equivalence of the constructs is clear, whereas other times it is more difficult to determine. For example, "dominance" as a personality trait may seem to be something that would be conceptually equivalent in all cultures. This fact is partially true, but nuances may make the concept somewhat different in various cultures. More collectivist cultures may emphasize the obligation to the group or family as being a more important aspect of dominance than individualistic cultures. It should be noted that various aspects of conceptual equivalence may emerge during translations of the test. For this reason, linguistic and conceptual equivalence are somewhat overlapping strategies.

More formal procedures for establishing conceptual equivalence might include investigating patterns of convergent and discriminant validity. A favored means of determining equivalence is factor analysis. It would be predicted that if indeed the concepts are comparable, then the same factors should emerge on the test when evaluated using samples from different cultures.

Metric Equivalence

The final means of establishing equivalence is through *metric equivalence*. This term refers to whether the instrument has similar psychometric properties across different groups/cultures. Assessing the extent to which the psychometric properties are different can include evaluating such areas as content, criterion, and construct validity. Note that a prerequisite for metric equivalence is that conceptual equivalence needs to be demonstrated first.

One of the initial things that persons reviewing tests will notice is that there are items on tests that appear irrelevant and possibly unfair for various groups. For example, a person from a different country could not reasonably be expected to know prominent political leaders in the country where the test was developed. On the surface, it would appear that such a test is culturally biased. Within the United States, early intuitive observations seemed to suggest that many African American children and other minorities usually do not have the opportunity to learn the types of material contained on many tests. Thus, their lower scores may represent not a lack of "intelligence," but rather a lack of familiarity with European American, middle-class culture. Critics of tests point out that it would clearly be unfair to assess a European American's "intelligence" based on whether he or she knows idioms or facts specific to a certain ethnic minority or national group. Low scores would simply measure an individual's relative familiarity with the knowledge contained within the group rather than his or her specific "mental strengths."

If this reasoning is used, many IQ and aptitude tests may appear on the surface to be culturally biased. However, studies in which researchers, to the best of their ability, eliminated biased test items or items that statistically discriminated between minorities and nonminorities did not alter total test scores (C. R. Reynolds, 2000).

In a representative study, 27 items were removed from the SAT that consistently differentiated minorities from nonminorities. This removal did little to change either the test takers' individual scores or the differences between the two groups (Flaugher & Schrader, 1978). Thus, the popular belief, based on a superficial appraisal of many psychological tests, that biased items are responsible for test differences does not appear to be supported by research.

Although test differences between minority and nonminority groups have frequently been found, the meaning and causes of these differences continue to be debated. For example, it has been demonstrated that African Americans consistently scored 12 to 15 IQ points lower than European Americans on the WISC-III and WAIS-III (Heaton, Taylor, & Manly, 2003; Prifitera, Weiss, & Saklofske, 1998). When African Americans and European Americans of equal socioeconomic status were compared, the differences in IQ scores were reduced to 11 to 13 IQ points (Heaton et al., 2003). Performance by Hispanics is about 7 IQ points lower than that of European Americans, and Asian Americans have been found to have IQ scores roughly equal to those of European Americans. Personality tests have also been found to have differences among various ethnic groups within the United States. For example, some studies (Dahlstrom, Lachar, & Dahlstrom, 1986; Timbrook & Graham, 1994) have found that African Americans have means five T-score points higher for MMPI scales F, 8, and 9. However, these differences were either decreased or found to be insignificant when groups were matched for age and education. This point suggests that socioeconomic factors may be an important reason for score differences. Socioeconomic status still accounts for only part of the reason for differences in test performance on cognitive tests, however (Sackett, Borneman, & Connelly, 2008). Other possible reasons are lack of belief in the impact of effort, level of acculturation, the effects of discrimination, gaps in general skills, or possible genetic differences. The reasons for these differences have been hotly debated and at this point are still unclear (see Neisser et al., 1996, and W. M. Williams, 2000).

Another consideration related to metric equivalence is the adequacy of the predictive validity of various tests when used with minority groups. Because one of the main purposes of these tests is to predict later performance, it is essential to evaluate the extent to which the scores in fact adequately predict areas such as performance in college for different populations. A representative group of studies indicates that SAT scores actually overpredict how well some minorities will perform in college (Hunter & Schmidt, 1996, 2000; A. R. Jensen, 1984; Sackett et al., 2008). Intelligence test scores have also been found to consistently predict African American work performance as accurately as European American performance (J. E. Hunter & Schmidt, 2000). Furthermore, both the WISC and the WISC-R were found to be equally as effective in predicting the academic achievement of both African Americans and European Americans in primary and secondary schools (Neisser et al., 1996; Reynolds & Hartlage, 1979).

A number of tests have been developed with the partial intent of using them in the assessment of ethnic minorities and cross-national groups, and they tend to emphasize nonverbal tasks. Included are the Leiter International Performance Scale, Peabody Picture Vocabulary Test–IV, Raven's Progressive Matrices, the Universal Nonverbal Intelligence Test–2, and the Test of Nonverbal Abilities (Bracken & McCallum, 2015;

McCallum, Bracken, & Wasserman, 2001). Some of these tests have been found to have minimal cultural bias (see Kaufman & Lichtenberger, 2006). In addition, the K-ABC-II (Kaufman et al., 2005) demonstrates minimal cultural bias. Mean IQ scores for European Americans, African Americans, and Hispanics are relatively close, and there is some evidence that reliability and concurrent validity is comparable for different ethnic populations (Kaufman et al., 2005).

As is true for ability tests and tests of scholastic aptitude, personality tests have the potential to be biased. The main research in this area has been performed on the MMPI/MMPI-2, and it has consistently indicated that minority groups do score differently than do nonminorities (see section titled "Use with Diverse Groups" in Chapter 7). However, these differences have not been found to be consistent across all populations (Greene, 1987, 1991, 2000). For example, African Americans from forensic, psychiatric, and vocational populations have been found to have varying patterns of mean scale elevations when compared to the mean scale elevations for European Americans. Even if consistent score differences were found, this does not mean these differences will be of sufficient magnitude to alter a clinician's interpretations, nor does it mean that predictions based on empirical criteria will be different. Studies using empirical criteria for prediction indicate that the MMPI does not result in greater descriptive accuracy for European Americans than for African Americans (Elion & Megargee, 1975; Greene, 1991, 2000).

Reviews of MMPI/MMPI-2 performance for Asian Americans, African Americans, Hispanics, and Native Americans have concluded that since no consistent patterns have emerged between various ethnic groups, it is premature to use different ethnically based norms (J. R. Graham, 2011; Greene, 1987, 1991; G. C. N. Hall, Bansal, & Lopez, 1999; Schinka, LaLone, & Greene, 1998). What seems to affect MMPI profiles more than ethnicity are moderator variables, such as socioeconomic status, intelligence, and education. Furthermore, the existing differences may result from true differences in behavior and personality caused by the greater stresses often encountered by minorities. J. R. Graham (1987) suggested that, when MMPI scores are deviant, the clinician should tentatively accept these scores but make special efforts to explore the person's life situation and level of adjustment and integrate this information with the test scores.

From this discussion, it should be obvious that developing test equivalence is complicated and that the meanings of various patterns of scores are far from being resolved. Several general solutions have been suggested (see Suzuki & Ponterotto, 2007). These include improving selection devices, developing different evaluation criteria, improving general skills, and changing social environments. Improving the use of selection devices involves paying continual attention to, and obtaining greater knowledge of, the meaning of different scores for different subgroups. Doing this may include tailoring specific test scores to the types of decisions individuals may make in their lives.

Another approach to solving the problem of potential test equivalence and bias is to develop different and more adequate criterion measures. For example, objective measures of work performance may be more accurate predictors than formal tests. These predictions of work performance may be higher if made by persons who share similar ethnic backgrounds. Related to this point, it may be crucial to consider the impact of various settings. For example, if a European American and a Latino attorney are placed in settings in which they work with Latinos, it is probable that the Latino

attorney would be more effective because he or she will have increased rapport and greater familiarity with the language and values of his or her clientele.

Another solution involves changing the social environment. Part of the rationale for emphasizing this approach is the belief that the differences in test scores between minorities and nonminorities are not because of test bias but rather because tests accurately reflect the effects of an unequal environment and unequal opportunities (C. R. Reynolds, 2000). Even though, in some situations, different minority norms and additional predictive studies on minority populations are necessary, the literature suggests that tests are not as biased as they have been accused of being (see Sackett et al., 2008). Removing seemingly biased or discriminating items still results in the same mean test scores, ability tests often still provide accurate predictions of grade point average for both minorities and nonminorities, and the MMPI-2 often makes behavioral predictions that are equally as accurate for various ethnic groups. These facts suggest that tests themselves are often not the problem but merely the means of establishing that, often, inequalities exist between ethnic groups. The goal should be to change unequal environments that can ideally increase a population's skills as measured by current tests of aptitude, IQ, and achievement. Whereas improving selection devices and developing different criterion measures are still important, future efforts should also stress more equal access to educational and other opportunities.

Probably the most important strategy is to maintain a flexible attitude combined with the use of alternative assessment strategies. Doing this changes the focus from merely establishing test equivalence to using a wide array of alternate assessment strategies. Thus, nonverbal techniques might be used, such as the Universal Nonverbal Intelligence Test (Bracken & McCallum, 1998), Raven's Progressive Matrices Test, or emphasis on the perceptual/nonverbal subtests of the WAIS-IV/WISC-V. In addition, "dynamic testing," in which actual observations of the benefit a person receives from learning situations, also shows promise in assessing the extent to which a client can benefit from educational interventions (learning potential; Grigorenko & Sternberg, 1998). Material beyond tests, such as teacher reports, discussions with parents, history, and behavioral observations, should also be given greater significance.

Diagnostic Issues

DSM-5 diagnosis needs to be considered within the context of cultural considerations. In addition to noting the cultural identity of the client, it is also crucial to carefully listen to cultural explanations of the client's difficulty. One category of presentation is the cultural concepts of distress that are outlined in the *DSM-5*. For example, *dhat syndrome* (mainly South Asia) is a cluster of symptoms that includes anxiety, depressive mood, and multiple somatic complaints. Also important is how the presence of oppression and discrimination among ethnic groups might contribute to misdiagnosing a person as being paranoid. A further example is how a disorder such as depression might be presented in primarily physiological terms within some cultures. In such cases, the external presentation would need to be decoded in order to identify the underlying depression. Research has clearly demonstrated varying rates of diagnoses in various cultures (Nguyen et al., 2007). What is less clear is whether these varying diagnoses represent genuine differences in rates or possible underdiagnosis, overdiagnosis,

or misdiagnosis. The practical implication is that when errors in diagnosis do occur, they have the potential to result in poor decisions and inappropriate treatment. It might also be necessary to consider combining standard psychological treatments with culture-specific interventions.

Interpretation Guidelines

The preceding discussion clearly indicates that ensuring accurate interpretations for diverse groups is challenging but also essential. Acculturation, equivalence, cultural competence, and the client's self-description within the context of his or her culture all need to be taken into account. Clinicians also need to incorporate what is known about how the instruments function within various cultures, including translations, idioms, norms, and various types of validity. However, it is nearly impossible to definitively demonstrate equivalence, due to the many steps and issues involved as well as the basic fact that error is inherent in any process aimed at equating two different cultures. Due to this fact, clinicians need to be both flexible and sensitive. For example, the pathological aspects of a high score on MMPI-2 Scale 6 (Paranoia) may need to be moderated if elevated for a client who has experienced significant racial discrimination. Similarly, indicators of low emotional expressiveness on the Rorschach may need to be modified if the person's emotional responses seemed to be "blunted" due to struggles with English as a second language. Often a phrase needs to be included in a report like " … results need to be treated with caution as the instruments have not been adequately adapted for use within the client's culture." Inserting a phrase such as this means that there are no clear specific strategies, but instead there are general guidelines to work with. Information and guidelines relevant to specific tests are included in each of the test-related chapters (see the sections titled "Use with Diverse Groups" in those chapters).

SELECTING PSYCHOLOGICAL TESTS

The most important factor in test selection is the extent to which the test is useful in answering the referral question. An assessment of neurological patients might use tests sensitive to cerebral deficit; patients with depression might be given the Beck Depression Inventory–II (A. T. Beck, Steer, & Brown, 1996); and patients with pain might be given the McGill Pain Questionnaire (Melzack, 1975), Millon Behavioral Health Inventory (Millon, Green, & Meagher, 2000), or Illness Behavior Questionnaire (Pilowski, Spence, Cobb, & Katsikitis, 1984).

Another important factor in test selection is a particular practitioner's training, experience, personal preferences, and familiarity with relevant literature. For example, a clinician who has received training in the MMPI-2 might be concerned about its ability to assess personality disorders and may rather choose to use an instrument such as the MCMI-IV (Millon, Grossman, & Millon, 2015). Clinicians might also select an instrument because it has practical efficiency in terms of time and economy (Groth-Marnat, 1999). Thus, they may wish to use simple behavioral predictions made by the client rather than use more expensive, time-consuming, and, quite possibly, less accurate tests (Shrauger & Osberg, 1981). Computer-assisted instruments may also

help to lower the cost of assessment, primarily by reducing direct practitioner time and achieving greater speed for scoring and hypothesis generation. A final crucial factor is that the assessment instrument should have good psychometric properties (see Hunsley & Mash, 2008).

The most frequently used assessment techniques are discussed in Chapters 3 to 13. Contact details for the major psychological distributors, along with a partial listing of tests they carry, are listed in Appendix A. Additional information on tests and assessments can be found by contacting various organizations that focus on assessment, listed in Appendix B. Various combinations of tests typically constitute a core battery used by clinicians. However, it is often necessary to expand such a core battery depending on the specifics of the referral question. Table 2.2 provides a listing of the domains for assessment along with relevant tests.

Although some of the tests described in Table 2.2 are thoroughly described in specific chapters dedicated to them, others may be relatively unfamiliar, and practitioners should obtain additional information on them. Various sources are available for information about these and other tests. Such sources can provide important information for deciding whether to obtain the tests and incorporate them into a battery. Probably the most useful is the *Mental Measurements Yearbook*, which contains a collection of critical test reviews that include evaluations of the tests and an overview on the tests. The *Nineteenth Mental Measurements Yearbook* was published in 2014 (Carlson, Geisinger, & Jonson, 2014), but it may be necessary to consult previous editions as not all tests are reviewed again in each new edition. The reviews are available in book form as well as online (*Mental Measurement Database*; see www.buros.org). *Tests in Print VIII* (L. L. Murphy, Geisinger, Carlson, & Spies, 2011) is associated with the *Mental Measurements Yearbook* but, rather than focusing on evaluating tests, lists information on each test, such as title, population it was designed for, available subtests, updating, author(s), and publisher. A further listing, description, and evaluation of tests can be found in Maddox (2003), *Tests: A Comprehensive Reference for Assessment in Psychology, Education, and Business* (5th ed.), which provides descriptive information on more than 3,500 tests. Practitioners interested in obtaining information on rating scales and other measures used in clinical practice might consult *Measures for Clinical Practice and Research: A Sourcebook* (Fischer & Corcoran, 2007). In *A Guide to Assessments That Work*, Hunsley and Mash (2008) present tests according to types of disorders and provide descriptions of these tests along with ratings of their psychometric properties. Neuropsychological tests are reviewed in the preceding resources as well as in Lezak and colleagues' (2012) *Neuropsychological Assessment*; Strauss, Sherman, and Spreen's (2006) *Compendium of Neuropsychological Tests*; and specialty journals in neuropsychology, particularly *Neuropsychology Review*. A careful review of the information included in these references will often answer questions clinicians might have related to a test's psychometric properties, usefulness, appropriateness for different populations, details for purchasing, and strengths and limitations. Most of the questions listed in Table 1.1 (see Chapter 1) can be answered by consulting the preceding resources.

An important and current trend in research and practice on psychological assessment is the use of tests to generate a treatment plan (Harwood, Beutler, & Groth-Marnat, 2011; Groth-Marnat & Davis, 2014; Jongsma, Peterson, & Bruce, 2006; Maruish, 2004; Wright, 2010). Indeed, a basic objective of psychological

Table 2.2 Assessment Instruments Relevant for Specific Response Domains

Cognitive Functioning
General functioning
Mental Status Examination
Mini-Mental State Examination (MMSE)
Intellectual functioning
Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV)
Wechsler Intelligence Scale for Children–Fifth Edition (WISC-V)
Stanford-Binet–Fifth Edition (SB5)
Kaufman Assessment Battery for Children–Second Edition (KABC-2)
Woodcock-Johnson Psychoeducational Battery–Fourth Edition (WJ-IV)
Memory functioning
Wechsler Memory Scale–Fourth Edition (WMS-IV)
Rey Auditory Verbal Learning Test
California Verbal Learning Test (CVLT)
Benton Visual Retention Test
Visuoconstructive abilities
Bender Visual Motor Gestalt Test–Second Edition (Bender-2)
Drawing tests
Content of thought processes
Thematic Apperception Test (TAT)
Children's Apperception Test (CAT)
Roberts' Apperception Test for Children (RATC)
Academic Achievement
Woodcock Johnson Tests of Achievement–Fourth Edition (WJ-IV)
Wide Range Achievement Test–Third Edition (WRAT-III)
Wechsler Individual Achievement Test–Third Edition (WIAT-III)
Personality Functioning, Emotional Functioning, and Level of Psychopathology
General patterns and severity
Minnesota Multiphasic Personality Inventory–Second Edition (MMPI-2)
Minnesota Multiphasic Personality Inventory–Second Edition RF (MMPI-2-RF)
Millon Clinical Multiaxial Inventory–Fourth Edition (MCMI-IV)
Millon Adolescent Clinical Inventory (MACI)
Rorschach
Symptom Checklist 90–Revised (SCL-90)
Brief Symptom Inventory (BSI)
Personality Inventory for Children–Second Edition (PIC-2)

(*continued*)

Table 2.2 *(Continued)*

General personality measures
Sixteen Personality Factors (16-PF)
NEO-PI-R
Adjective Checklist
Sentence completion tests
Diagnosis
Diagnostic Interview Schedule
Schedule for Affective Disorders and Schizophrenia
Structured Clinical Interview for DSM (SCID)
Structured Interview for DSM Personality Disorders (SCID-2)
Diagnostic Interview for Children and Adolescents
Depression
Beck Depression Inventory–Second Edition (BDI-2)
Hamilton Rating Scale for Depression
Children's Depression Inventory
Anxiety
State-Trait Anxiety Inventory
Fear Survey Schedule
Anxiety Disorders Interview Schedule
Sexual disturbance
Derogatis Sexual Functioning Inventory
Alcohol abuse
Michigan Alcoholism Screening Test
Alcohol Use Inventory
Interpersonal patterns
California Psychological Inventory (CPI)
Rathus Assertiveness Schedule
Therapeutic Reactance Scale
Taylor Johnson Temperament Analysis
Marital/family disturbance
Dyadic Adjustment Scale
Family Environment Scale
Marital Satisfaction Inventory
Academic/school adjustment
Achenbach Child Behavior Checklist (CBCL)
Vineland Social Maturity Scale
Connors Behavior Rating Scales
Kinetic School Drawing
Behavior Assessment System for Children–Second Edition (BASC-2)

Table 2.2 *(Continued)*

Adaptive level
AAMD Adaptive Behavior Scale
Vineland Adaptive Behavior Scale
Prognosis and risk
Suicide potential
Scale of Suicide Ideation
Beck Hopelessness Scale
Schizophrenia prognosis
Camberwell Family Interview
Vocational Interests
Career Assessment Inventory
Kuder Occupational Interest Survey
Self-Directed Search
Strong Interest Inventory (SII)

assessment is to provide useful information regarding the planning, implementation, and evaluation of treatment. With the increased specificity of both treatment and assessment, this goal is becoming possible. For example, oppositional, resistant clients have been found to have optimal treatment outcomes when either self-directed or paradoxical interventions have been used (Beutler, Clarkin, & Bongar, 2000; Beutler, Sandowicz, Fisher, & Albanese, 1996). In addition, a problem's severity has clear implications for the restrictiveness of treatment (inpatient, day treatment, outpatient) as well as treatment duration and intensity. Thus, clinicians should not select tests based simply on their diagnostic accuracy or psychometric properties; they should also be concerned with the functional utility of the tests in treatment planning. Accordingly, Chapter 14 presents a systematic, integrated approach to transforming assessment results into a series of clear treatment recommendations.

One special concern in selecting tests is faking. In many situations, clinicians might be concerned that persons will either consciously or unconsciously provide inaccurate responses (see kspope.com/assess/). Malingering ("inconsistent effort") is becoming an increasingly important issue, especially in forensic settings, where personal gain may result in presenting "fake bad" results. Thus, clinicians may want to pay particular attention to validity scales built in to tests (e.g., MMPI-2, MCMI-IV) and use specialty instruments designed to detect faking (e.g., Test of Memory Malingering, Structured Interview of Reported Symptoms). Although controversial, many projective techniques may be resistant to attempts at faking.

Another special concern in selecting tests relates to the time required for assessment, which may cause examiners to consider selecting short forms of instruments such as the WAIS-IV or WISC-V. Although many short forms for cognitive tests seem sufficiently valid for screening purposes, their use as substitutes for the longer forms is not acceptable (Kaufman, Kaufman, Balgopal, & McLean, 1996; Kaufman & Lichtenberger, 2002). Most past attempts to develop short forms for the longer

objective personality tests, such as the MMPI-2, have not been successful (Butcher, 2011). However, future computerized applications that tailor items based on a client's previous responses (adaptive testing) may result in the development of shortened administrations with acceptable psychometric properties (Forbey & Ben-Porath, 2007). In addition, the recent 338-item MMPI-2 Restructured Form is a shorter and psychometrically improved version (Ben-Porath & Tellegen, 2008/2011).

During the evaluation of single cases, such as in clinical diagnosis and counseling, clinicians do not usually use formal combinations of test scores. Rather, they rely on their past judgment, clinical experience, and theoretical background to interpret and integrate test scores. However, for personnel decisions, academic predictions, and some clinical decisions (recidivism rate, suicide risk), clinicians may be advised to use statistical formulas (Aegisdottir et al., 2006). The two basic approaches for combining test results are multiple regression equations and multiple cutoff scores. Multiple regression equations are developed by correlating each test or subtest with a criterion. The higher the correlation, the greater is the weight in the equation. The correlation of the entire battery with the criterion measure gives an indication of the battery's highest predictive validity. For example, high school achievement can be predicted with this regression equation, which combines IQ and California Psychological Inventory (CPI) subtests:

$$\begin{aligned}\text{Achievement} = {} & .786 + .195\ \text{Responsibility} + .44\ \text{Socialization} \\ & - .130\ \text{Good Impression} + .19\ \text{Achievement via Conformance} \\ & + .179\ \text{Achievement via Independence} + .279\ \text{IQ}\end{aligned}$$

This equation raises the correlation with grade point average to .68 as compared with .60 when using IQ alone (Megargee, 1972). This correlation indicates that academic achievement is dependent not only on intellectual factors but also on psychosocial ones, such as responsibility, socialization, achievement via independence, and achievement via conformance, all of which are measured by the CPI. The second strategy, multiple cutoff scores, involves developing an optimum cutoff for each test or subtest. If the person is above a certain specified score (e.g., above the range for brain damage or schizophrenia), the score can be used to indicate the presence of a certain characteristic. Although equations or cutoffs have not been developed for all tests, the decision to include a test in a battery may depend in part on the presence of such formal extensions of the test. In addition, many of the computer-assisted interpretive packages use various actuarial formulas (usually in combination with expert interpretations) to develop their interpretations.

COMPUTER-ASSISTED ASSESSMENT

During the past 40 years, computer-assisted assessment has grown exponentially. By 1990, 17% of practicing psychologists frequently used computer-generated narratives, with an additional 36% using them on an occasional basis (Spielberger & Piotrowski, 1990). By 1999, the number of psychologists stating that they used some form of computer-assisted testing had increased to 40% (McMinn, Buchanan et al., 1999). More than 400 software packages are available, many of which are listed in various

catalogs published and distributed by test suppliers. At present, computers are used mainly for their clerical efficiency in scoring and data storage and to generate interpretive reports. However, more and more, testing is becoming available in computer/tablet-assisted formats. Future uses of computers are likely to continue to experiment with features such as innovative presentation of items (e.g., adaptive testing), networked norms, novel presentation of stimuli (e.g., virtual reality), psychophysiological monitoring, and artificial intelligence (Garb, 2000; Groth-Marnat, 2000a, 2009). Computing in mental health has included not only computer-assisted assessment but also computer interviews, computerized diagnosis, computer-aided instruction, direct treatment intervention, clinical consultation, and simulated psychiatric interviews (Lichtenberger, 2006; McMinn, Buchanan et al., 1999).

There have been a number of particular advances in computer-assisted administration and interpretation in neuropsychology (see special series review by Kane, 2007). Batteries have been developed mainly in large organizational contexts (military, Federal Aviation Authority) and focused on specialized types of problems. For example, the Neurobehavioral Evaluation System (NES) is particularly sensitive to the impact of environmental toxins (Groth-Marnat, 1993), CogScreen has been used in the selection of airline pilots, and the military's Unified Tri-service Cognitive Performance Assessment Battery (UTC-PAB) was originally developed to assess the impact of drugs in the workplace. The Cambridge Neuropsychological Test Automated Batteries (CANTAB) have been found to detect and locate brain damage including early signs of Alzheimer's, Parkinson's, and Huntington's diseases (Fray, Robbins, & Sahakian, 1996; Luciana, 2003). Although computer-assisted programs show considerable promise, they are currently used less than the more familiar individually administered neuropsychological tests or test batteries (Camara, Nathan, & Puente, 2000; Luciana, 2003).

Computer-assisted assessment has a number of advantages. Use of computers can save valuable professional time, potentially improve reliability and fidelity to standardized administration, reduce possible tester bias, and reduce the cost to the consumer by improving efficiency (Butcher, Perry, & Hahn, 2004; Groth-Marnat, 1999: Kane, 2007; Luciana, 2003). Even greater benefits may someday be realized by incorporating more complicated decision rules in interpretation, collecting data on response latency and key pressure, incorporating computer-based models of personality, tailoring future questions to a client based on past responses, and estimating the degree of certainty of various interpretations (Groth-Marnat, 2000a, 2000b; Lichtenberger, 2006).

In the past, computer-assisted assessment has resulted in considerable controversy within mental health publications (Faust & Ziskin, 1989; Groth-Marnat & Schumaker, 1989), the popular media (C. Hall, 1983), and professional publications outside the mental health area (Groth-Marnat, 1985). A primary issue has been untested reliability and validity. Research on reliability, however, has typically indicated that computerized administrations have generally excellent reliability that is at least equivalent to the paper-and-pencil versions (Campbell et al., 1999; Kane, 2007; Luciana, 2003). In addition, computer-administered versus paper-and-pencil outcomes for traditional tests have generally been found to result in negligible differences in scores (Butcher et al., 2004; Finger & Ones, 1999). This finding supports the view that if a paper-and-pencil version of the test is valid, a computerized version will have equal validity resulting from the comparability in scores.

A further issue is the validity of computer-based test interpretation. Butcher et al. (2004) concluded that in the vast majority of computer-based interpretations, 60% of the interpretations were appropriate. Shorter to mid-length narratives were generally considered to have a higher proportion of valid interpretations when compared with longer ones. In addition, the narrative statements contained in the computer-based reports were comparable to the types of statements made by clinicians. Although this finding generally supports the use of computer-based interpretations, the fact that 40% or more of interpretations were not considered accurate means that the computer-based interpretations should be carefully evaluated. Thus, cutting and pasting computerized narratives into reports results in unacceptably high error rates. Indeed, 42% of psychologists surveyed felt this procedure raised ethical concerns (McMinn, Ellens et al., 1999). The previous summary clearly emphasizes that computer-based reports should not be used to replace clinical judgment but should instead be used as an adjunct to provide possible interpretations for the clinician to consider.

The Association of Test Publishers (2000) attempted to clarify standards in its *Guidelines for Computer-Based Testing* (as did the 2010 American Psychological Association's ethics code). The association stressed that only persons who meet the requirements for using psychological tests in general should use computer-based assessments (Turner, DeMers, Fox, & Reed, 2001). Specifically, users should have an understanding of psychological measurement, validation procedures, and test research. They should also limit their use of computerized techniques to those areas they are competent to use. They should be knowledgeable regarding how computer-based scores were generated and how interpretations have been made. Finally, they should be able to evaluate whether the computer-based procedures are applicable to how they will be used.

The preceding difficulties associated with computer-assisted assessment suggest a number of guidelines for users (Butcher et al., 2004; Groth-Marnat & Schumaker, 1989). First, practitioners should not blindly accept computer-based narrative statements but rather should ensure, to the best of their ability, that the statements are both linked to empirically based research and placed in the context of the unique history and unique situation of the client. Computers have, among other benefits, the strong advantage of offering a wide variety of possible interpretations to the clinician, but these interpretations still need to be critically evaluated. Far greater research needs to be performed on both the meaning of computer-administered test scores and on the narrative interpretations based on these scores. The developers of software should also be encouraged to provide enough information in the manual to allow proper evaluation of the programs and should develop mechanisms to ensure that obsolete programs are updated.

RECOMMENDED READING

Butcher, J. N., Perry, J. N., & Hahn, J. (2004). Computers in clinical assessment: Historical developments, present status, and future challenges. *Journal of Clinical Psychology*, *60*, 331–345.

Dana, R. H. (2005). *Multicultural assessment: Principles, applications, and examples*. Mahwah, NJ: Erlbaum.

Heilbrun, K., Marczyk, G. G., & Dematteo, D. (2002). *Forensic mental health assessment: A case-book*. New York, NY: Oxford University Press.

Hunsley, J., & Mash, E. J. (Eds.). (2008). *A guide to assessments that work*. New York, NY: Oxford University Press.

Pope, K. (2007a). Informed consent in psychotherapy and counseling: Forms, standards & guidelines, & references. Retrieved from http://kspope.com/consent/index.php.

Pope, K. (2007b). Responsibilities in providing psychological test feedback to clients. Retrieved from http://kspope.com/assess/feedabs1.php.

Yalof, J., & Brabender, V. (2001). Ethical dilemmas in personality assessment courses: Using the classroom for in vivo training. *Journal of Personality Assessment*, *77*, 203–213.

Zuckerman, E. L. (2003). *The paper office: Forms, guidelines, and resources to make your practice work ethically, legally, and profitably*. New York, NY: Guilford Press.

Chapter 3

THE ASSESSMENT INTERVIEW

The single most important means of data collection to provide context for psychological evaluation is the assessment interview. Without interview data, most psychological test results are meaningless. The interview also provides potentially valuable information that may be otherwise unobtainable, such as behavioral observations, idiosyncratic features of the client, and the person's reaction to his or her current life situation. In addition, interviews are the primary means for developing rapport.

Sometimes an interview is mistakenly thought to be simply a conversation. In fact, an interview and a conversation differ in many ways. An interview typically has a clear sequence and is organized around specific, relevant themes, because it is meant to achieve defined goals. Its general objectives are to gather information that cannot easily be obtained through other means, establish a relationship that is conducive to obtaining the information, develop greater understanding in both the interviewer and interviewee regarding problems, and provide direction and support in helping the interviewee deal with problems. The interviewer must have knowledge about the areas to be covered during the interview and direct and control the interaction to achieve specific goals.

A basic dimension of an interview is its degree of structure. Some interviews allow the participants to freely drift from one area to the next, whereas others are highly directive and goal oriented, often using structured ratings and checklists. The more unstructured formats offer flexibility, possibly higher rapport, the ability to assess how clients organize their responses, and the potential to explore unique details of a client's history. Unstructured interviews, however, have received frequent criticism, resulting in widespread distrust of their reliability and validity. As a result, highly structured and semistructured interviews have been developed that provide sound psychometric qualities, the potential for use in research, and the ability to be administered by less trained personnel.

Regardless of the degree of structure, any interview needs to accomplish specific goals, such as assessing the client's strengths, level of adjustment, the nature and history of the problem, diagnosis, and relevant personal and family history. Techniques for accomplishing these goals vary from one interviewer to the next. Most practitioners use at least some structured aids, such as intake forms that provide identifying data and basic elements of history. Obtaining information through direct questions on intake forms frees the clinician to investigate other aspects of the client in a more flexible, open-ended manner. Clinicians might also use a checklist to help ensure that they have covered all relevant areas. Other clinicians use one of the formally developed structured interviews, such as the Schedule for Affective Disorders and Schizophrenia (SADS) or Structured Clinical Interview for the DSM-IV (SCID).

HISTORY AND DEVELOPMENT

Early Developments

The earliest method of obtaining information from clients was through clinical interviewing. At first, these interviews were modeled after question-and-answer medical formats, but later, the influence of psychoanalytic theories resulted in a more open-ended, free-flowing style. Parallel to the appearance of the psychoanalytically oriented interview was the development of the more structured and goal-oriented mental status examination, originally formulated by Adolf Meyer in 1902. The mental status examination assessed relevant areas of a client's current functioning, such as general appearance, behavior, thought processes, thought content, memory, attention, speech, insight, and judgment. Professionals also expressed early interest in the relationship between biographical data and the prediction of occupational success or prognosis for specific disorders.

Regardless of the style used, the interviews all had these common objectives: to obtain a psychological portrait of the person, to conceptualize what is causing the person's current difficulties, to make a diagnosis, and to formulate a treatment plan. The difficulty with unstructured interviews is that they were (and still are) considered to have questionable reliability, validity, and cost-effectiveness. The first standardized psychological tests were developed to overcome these limitations. Tests could be subjected to rigorous psychometric evaluation and were more economical because they required less face-to-face contact with the person(s) being evaluated.

Developments during the 1940s and 1950s

During the 1940s and 1950s, researchers and clinicians began conceptualizing and investigating five critical dimensions of interviews:

1. Content versus process
2. Goal orientation (problem solving) versus expressive elements
3. Degree of directiveness
4. Amount of structure
5. The relative amount of activity expressed by the participants

These issues have been the focus of numerous research studies. A representative and frequently cited study on interviewer style was reported by W. Snyder (1945), who found that a nondirective approach was most likely to create favorable changes and self-exploration in clients. In contrast, a directive style using persuasion, interpretation, and interviewer judgment typically resulted in clients being defensive and resistant to expressing difficulties. Strupp (1958) investigated the experience-inexperience dimension and found, among other things, that experienced interviewers expressed more warmth, a greater level of activity, and a greater number of interpretations. Level of empathy did not differ based the interviewer's degree of experience. Further representative studies include Porter's (1950) in-depth evaluation of the effects of different types of responses (evaluative, probing, reassuring) and R. Wagner's (1949) early review, which questioned the reliability and validity of employment interviews.

Developments During the 1960s

A considerable amount of research in the 1960s was stimulated by C. Rogers (1961), who emphasized understanding the proper interpersonal ingredients necessary for an optimal therapeutic relationship (warmth, positive regard, genuineness). Elaborating on Rogers's ideas, Truax and Carkhuff (1967) developed a 5-point scale to measure interviewer understanding of the client. This scale was used for research on interviewing and therapist training and as support for a client-centered theoretical orientation. Additional research efforts were directed toward listing and elaborating on different categories of interactions, such as clarification, summarizing, and confrontation.

Other investigators conceptualized interviewing as an interactive system in which the participants simultaneously influenced each other (Matarazzo, 1965; Watzlawick, Beavin, & Jackson, 1966). This emphasis on an interactive, self-maintaining system became the core for most early and later formulations of family therapy. The 1960s also saw the development and formalization of behavioral assessment, primarily in the form of goal-directed interviews that focused on understanding current and past reinforcers, as well as on establishing workable target behaviors. Proponents of behavioral assessment also developed formal rating instruments and self-reports for areas such as depression, assertiveness, and fear.

Some attempts were made at integrating different schools of thought into a coherent picture, such as Beier's (1966) conceptualization of unconscious processes being expressed through nonverbal behaviors that could then be subject to covert social reinforcement. However, the 1960s (and part of the 1970s) were mostly characterized by a splintering into different schools of conflicting and competing ideologies. For example, client-centered approaches emphasized the importance of staying with the client's self-exploration; behavioral interviews emphasized antecedents and consequences of behavior; and family therapy focused on interactive system processes. Parallel progress was made within each of these different schools and within different disciplines, but little effort was devoted to cross-fertilization and/or integration.

Throughout the 1950s and 1960s, child assessment was conducted primarily through interviews with parents. Direct interviews with the child were considered to be for therapeutic purposes rather than for assessment. Differential diagnosis was unusual; almost all children referred to psychiatric clinics were either undiagnosed or diagnosed as "adjustment reactions" (Rosen, Bahn, & Kramer, 1964). Early research by Lapouse and Monk (1958, 1964) using structured interviews indicated that mothers were more likely to report overt behaviors that are bothersome to adults (thumb-sucking, temper tantrums), but children were more likely to reveal covert difficulties (fears, nightmares). Somewhat later, P. Graham and Rutter (1968), using structured interviews of children (rather than a parent), found interrater agreement was high for global psychiatric impairment (.84); moderate for attentional deficit, motor behavior, and social relations (.61–.64); and low for more covert difficulties, such as depression, fears, and anxiety (.30).

Developments During the 1970s

Assessment with adults and children during the 1970s saw a further elaboration and development of the trends of the 1960s, as well as increased emphasis on structured

interviews. The interest in structured interviews was fueled largely by criticisms about the poor reliability of psychiatric diagnosis. Typical structured interview data would be transformed into such scales as organicity, disorganization, or depression-anxiety.

Initial success with adult structured interviews (e.g., Present State Examination, Renard Diagnostic Interview) encouraged thinking regarding the further development of child-structured interviews both for global ratings and for specific content areas. Child assessment became concerned not only with information derived from parents but also with the child's own experience. There was a trend toward direct questioning of the child, greater emphasis on differential diagnosis, and the development of parallel versions of structured interviews for both the parent(s) and child.

Behavioral strategies of interviewing for both children and adults not only emphasized the interviewee's unique situation but also provided a general listing of relevant areas for consideration. Kanfer and Grimm (1977) outlined the areas an interviewer should assess as:

1. Behavioral deficiencies,
2. Behavioral excesses,
3. Inappropriate environmental stimulus control,
4. Inappropriate self-generated stimulus, and
5. Problem reinforcement contingencies.

In a similar categorization, Lazarus (1973, 2005) developed his BASIC-ID model, which described a complete assessment as involving *b*ehaviors, *a*ffect, *s*ensation, *i*magery, *c*ognition, *i*nterpersonal relations, and need for pharmacological intervention/*d*rugs.

Additional themes in the 1970s included interest in biographical data, online computer technology, and the training of interviewer skills. Specifically, efforts were made to integrate biographical data for predicting future behavior (suicide, dangerousness, prognosis for schizophrenia) and for inferring current traits. J. W. Johnson and Williams (1980) were instrumental in developing some of the earliest online computer technology to collect biographical data and to integrate it with test results. Although training programs included interviewing skills, a central debate was whether these skills could actually be significantly learned or improved (Wiens, 1976).

Whereas most reviews of the literature in the 1970s emphasized the advantages of a comprehensive structured format, family therapists were dealing with group processes in which formal interview structure was typically deemphasized. Because most family therapists were observing fluid interactional processes, they needed to develop a vocabulary different from that used in traditional psychiatric diagnosis. In fact, *DSM* categories were usually considered irrelevant because they described static characteristics of individuals rather than ongoing group processes. Few, if any, structured formats were available to assess family relationships.

Developments During the 1980s

Many of the trends, concepts, and instruments developed in the 1960s and 1970s were further refined and adapted for the 1980s. One important effort was the adaptation

of many instruments to the *DSM-III* (1980) and *DSM-III-R* (1987). In addition, the increased delineation of childhood disorders required greater knowledge related to differential diagnosis and greater demand for structured interviews as adjuncts to assessment. Many of the efforts were consistent with the use of specific diagnostic criteria, along with a demand for efficiency, cost-effectiveness, and accountability. Despite concerns regarding computer-based interpretations (Groth-Marnat & Schumaker, 1989), some of these functions were beginning to be performed by specific computer programs. Because interviews were becoming increasingly structured, with the inclusion of scales and specific diagnostic strategies, the distinction between tests and interviews was becoming less clear. In some contexts, aspects of interviewing were even replaced with computer-requested and computer-integrated information and combined with simple programs to aid in diagnosis, such as DIANO III (Spitzer, Endicott, & Cohen, 1974) and CATEGO (Wing, Cooper, & Sartorius, 1974). During the mid- and late 1980s, most clinicians, particularly those working in large institutions, used a combination of structured interviews and open-ended unstructured approaches. Some research focused on the importance of the initial interview regarding clinical decision making and later therapeutic outcome (Hoge, Andrews, Robinson, & Hollett, 1988; Turk & Salovey, 1985). There was also a greater appreciation and integration of the work from different disciplines and from differing theoretical persuasions (Bellack & Hersen, 1988). Finally, greater emphasis was placed on the impact and implications of culture and gender on the assessment process (L. Brown, 1990).

The 1990s and Into the Millennium

Two of the defining features of psychology in the 1990s were managed health care and the controversy over the validity of repressed memories. Both of these issues had significant implications for interviewing. Managed health care emphasized the cost-effectiveness of providing health services; and for interviewing, this means developing the required information in the least amount of time. Doing this may mean streamlining interviews by maximizing computer-derived information or self-administered forms. The use of computer-assisted interviewing brings up the larger issue of the extent to which practitioners need to spend face-to-face time with the client rather than deriving information through other means. The development of single-session therapy (Hoyt, 1994) illustrates the potential brevity of information gathering that might be required before making therapeutic interventions. There was also recognition that precise patient–treatment matching can optimize the treatment and potentially the cost-effectiveness of psychosocial interventions (Antony & Barlow, 2011; Beutler & Clarkin, 1990; Beutler, Clarkin, & Bongar, 2000).

The controversy over repressed memories has forced interviewers to clarify the extent to which the information they derive from clients represents literal as opposed to narrative truth. Research has consistently indicated that client self-reports are reconstructions of events (Henry, Moffitt, Caspi, Langley, & Silva, 1994; Loftus, 1993) and are likely to be particularly questionable for retrospective reports of psychosocial variables (Garb, 2007; Henry et al., 1994; Piasecki, Hufford, Solhan, & Trull, 2007). The even greater challenge to interviewers is to ensure that their interviewing style and method of questioning are not distorting the information derived from clients.

This issue becomes intensely highlighted during interviews to investigate the possibility of childhood sexual abuse (see guidelines in S. White & Edelstein, 1991).

Further themes in the 1990s and into the millennium were the importance of interview strategies for special populations and the development of new technologies. It is clear that many diverse populations are more likely to be misdiagnosed. At least in part, this misdiagnosis results in worse outcomes compared with majority groups (Neighbors et al., 2007; Nguyen, Huang, Arganza, & Liao, 2007). The potential for misdiagnosis for minority groups demands that clinicians be aware of their own biases, become knowledgeable regarding these subgroups, and make appropriate modifications to their interviews (Ponterotto & Grieger, 2007). Several new technologies are both available and becoming progressively more utilized. These include computer-administered interviews (Garb, 2007) as well as data derived from electronic diaries (Piasecki et al., 2007) and ambulatory sensors (Haynes & Yoshioka, 2007) that become a part of clinical interviews. The themes and issues related to cost-effectiveness, patient–treatment matching, recovered memories, use of new interview technologies, and strategies for interviewing special populations will continue to be important themes throughout the first few decades of the millennium.

ISSUES RELATED TO RELIABILITY AND VALIDITY

Although the interview is not a standardized test, it is a means of collecting data and, as such, can and should be subjected to some of the same types of psychometric considerations as formal tests. Evaluating the psychometric properties of interviews is important because interviews can introduce numerous sources of bias, particularly if the interviews are relatively unstructured. Reliability of interviewers is usually discussed in relation to interrater (interviewer) agreement. R. Wagner's (1949) early review of the literature found tremendous variation, ranging from .23 to .97 (*Mdn*.57) for ratings of personal traits and .20 to .85 (*Mdn*.53) for ratings of overall ability. Later reviews have generally found similar variations in interrater agreement (Arvey & Campion, 1982; L. Ulrich & Trumbo, 1965). The problem then becomes how to determine which ratings to trust and which to view with skepticism. Of particular relevance is why some interviewers focus on different areas and have different biases. A consistent finding is that, when interviewers were given narrow areas to assess and were trained in interviewer strategies, interrater agreement increased (Dougherty, Ebert, & Callender, 1986; Zedeck, Tziner, & Middlestadt, 1983). The consensus is that highly structured interviews were more reliable (Garb, 2007; Huffcutt & Arthur, 1994). However, increased structure undermines one of the greatest strengths of interviews—their flexibility. In many situations, a free-form, open-ended approach may be the best way to obtain some types of information.

Research on interview validity has typically focused on various sources of interviewer bias. Halo effects result from the tendency of an interviewer to develop a general impression of a person and then infer other seemingly related characteristics. For example, clients who are considered to express warmth may be seen as more competent or mentally healthy than they actually are. This clustering of characteristics may be incorrect, thereby producing distortions and exaggerations. Similarly, first

impressions have been found to bias later judgments (W. Cooper, 1981). Confirmatory bias might occur when an interviewer makes an inference about a client and then directs the interview to elicit information that confirms the original inference. This bias typically occurs when clinicians develop initial diagnostic impressions and then ignore later relevant information since they are somehow invested in confirming their initial impressions. Similarly, a psychoanalytically oriented interviewer might focus on questions related to early childhood traumas, possibly incorrectly confirming traditional psychoanalytic explanations of current adult behaviors. Similar to halo effects is the finding that one specific outstanding characteristic (e.g., educational level, physical appearance) can lead an interviewer to judge other characteristics that he or she incorrectly believes are related to the outstanding one. For example, physical attractiveness has been found to create interviewer bias in job applicants (Gilmore, Beehr, & Love, 1986). In a clinical context, physical attractiveness may result in practitioners either deemphasizing pathology or, on occasion, exaggerating pathology because of discomfort the interviewers may feel over their feelings of attraction (L. Brown, 1990). Interviewers also may focus incorrectly on explanations of behavior that emphasize traits rather than situational determinants (Ross, 1977).

In addition to the interviewer's perceptual and interactional biases, interviewees themselves may distort their responses. Some specific areas of distortions include victims of automobile accidents typically exaggerating the amount of time they lost from work; 40% of respondents providing overestimates of their contributions to charity; and 17% of respondents reporting their ages incorrectly (R. Kahn & Cannell, 1961). Some interviewees may present an overly favorable view of themselves, even if they are relatively naive regarding their motivations. Distortions, however subtle, are often found in sensitive areas, such as sexual behavior. More extreme cases of falsification occur with outright (conscious) lies, delusions, confabulations, and lies by pathological (compulsive) liars that they partially believe themselves (Kerns, 1986). Inaccuracies based on retrospective accounts have been found to most likely occur related to psychosocial information (e.g., family conflict, onset of psychiatric symptoms) compared with variables such as change of residence, reading skill, height, and weight (B. Henry et al., 1994).

Reviews of interview validity, in which interviewer ratings were compared with outside criterion measures, have, like reliability measures, shown tremendous variability, ranging from –.05 to +.75 (Arvey & Campion, 1982; Henry et al., 1994; Huffcutt & Arthur, 1994; J. Hunter & Hunter, 1984; L. Ulrich & Trumbo, 1965). One clear finding is that validity increases as the structure of the interview format increases (Huffcutt & Arthur, 1994; Marchese & Muchinsky, 1993). For example, a meta-analysis by Wiesner and Cronshaw (1988) found that unstructured interviews had validity coefficients of .20, structuring the interview increased the validity to .63, and structured interviews by a panel using consensus ratings increased validity coefficients to a quite respectable .64. However, the validity seems to vary according to the type of variable that is being assessed. Situational employment interviews (i.e., asking the interviewee what he or she would do in a particular situation) had higher validities (.50) than interviews used to assess past job-related behavior (.39) or rate psychological qualities such as dependability (.29; McDaniel, Whetzel, Schmidt, & Maurer, 1994). It has also been found that interview accuracy increases more when interviewees are held accountable for the

process they went through when coming to their decisions, compared to being held accountable for the accuracy of their predictions (procedural versus outcome accountability; Brtek & Motowidlo, 2002).

The previous brief review indicates that adding structure to interviews and paying close attention to the procedure by which decisions are made typically result in higher levels of validity. It also means that information derived from unstructured interviews should be treated cautiously and treated simply as hypotheses that need to be supported by other means. Interviewers should also continually question the extent to which their particular style, attitudes, and expectations might be compromising interview validity. Given the difficulties related to unstructured formats, a variety of formal structured clinical interviews have been developed. Additional information on the reliability and validity of the most frequently used structured clinical interviews is provided in the "Structured Interviews" section of this chapter.

ASSETS AND LIMITATIONS

Both structured and unstructured interviews allow clinicians to place test results in a wider, more meaningful context. In addition, biographical information from interviews can be used to help predict future behaviors; what a person has done in the past is an excellent guide to what he or she may continue doing in the future. Improving prediction of suicide risk, success in certain occupations, and prognosis for certain disorders can often be effectively accomplished by attending to biographical data rather than test scores.

Because tests are almost always structured or "closed" situations, the unstructured or semistructured interview is typically the only time during the assessment process when the clinician can observe the client in an open, ambiguous situation. Observations can be made regarding how individuals organize their responses, and inferences can be derived from subtle, nonverbal cues. These inferences can be followed up with further, more detailed questioning. This flexibility inherent in unstructured and semistructured interviews is frequently their strongest advantage over standardized tests. The focus during unstructured interviews is almost exclusively on the individual rather than on how that individual does or does not compare with a larger normative comparison group. Some types of information can be obtained only through this flexible, person-centered approach, which allows the interviewer to pay attention to idiosyncratic factors. In crisis situations when relatively rapid decisions need to be made, it can be impractical to take the time required to administer and interpret tests, leaving interviews and rapid screening devices as the only means of assessment. Finally, interviews allow clinicians to establish rapport and encourage client self-exploration. Rarely do clients reveal themselves or perform optimally on tests unless they first sense trust, openness, and a feeling of being understood.

The greatest difficulty with unstructured interviews is interviewer bias from perceptual and interactional processes such as the halo effect, confirmatory bias, and the primacy effect. This bias typically results in considerable variability for both reliability

and validity, as well as in difficulty comparing one subject with the next. One of the main reasons for diagnostic disagreement is variations in the information obtained (information variance) and variations in the criteria (criterion variance) used to conclude the presence or absence of a condition. Variation in interviewing means that different practitioners develop and ask a wide variety of questions and apply standards for the presence of a condition, such as depression, in an inconsistent fashion.

Structured interviews have many distinct advantages over unstructured approaches. Because structured interviews have more psychometric precision, the results enable comparability between one case and the next (or the population). The standardized presentation allows for the development of reliable ratings, reduces information variance, and uses consistent diagnostic criteria (Garb, 2007; Summerfeldt & Antony, 2002). In addition, the comprehensiveness of many structured interviews reduces the likelihood of missing a diagnosis or set of relevant symptoms. Partly because of these advantages, structured clinical interviews have progressed from being used primarily for research to use in a number of clinical settings. One issue, however, is the time required for structured interviews. The more recently developed, but not widely used, computer-assisted programs offer a potential method of countering this difficulty (Epstein & Klinkenberg, 2001; Garb, 2007). In addition, computer-administered interviews are comprehensive, and clients are more likely to disclose highly sensitive information when compared with clinician-administered interviews (Garb, 2007). Instruments such as the Diagnostic Interview Schedule (DIS) and Diagnostic Interview for Children and Adolescents (DICA) have been designed for administration by lay interviewers, thereby reducing the time required by professionals.

Although structured interviews generally have stronger psychometric properties than unstructured formats, they tend to overlook the idiosyncrasies and richness of the person. In many cases, these unique aspects may go undetected and yet may make a significant difference in interpreting test scores or making treatment recommendations. Although still somewhat controversial (Helzer & Robins, 1988), another criticism by many clinicians and researchers is that highly structured approaches may not create enough rapport for the client to feel sufficiently comfortable about revealing highly personal information. This is truer for the highly structured interviews, such as the DIS, than for a semistructured instrument, such as the SADS, which includes an initial, relatively unstructured component. However, M. Rosenthal (1989) noted that rapport with structured instruments can be enhanced through carefully educating the client as to the importance and procedures of these more structured approaches.

Although many of the structured interviews have demonstrated adequate reliability, studies relating to validity have primarily focused on the general level of impairment or simple discriminations between psychiatric and nonpsychiatric populations. There has been considerable controversy over what exactly is an acceptable outside criterion measure regarding the "true" diagnosis. In-depth studies of construct validity or incremental validity have yet to be performed. Furthermore, far more work needs to be done on the treatment utility of structured interviews in areas such as selection of treatment, likely response to specific forms of pharmacological or psychotherapeutic interventions, and prognosis.

THE ASSESSMENT INTERVIEW AND CASE HISTORY

General Considerations

The previously mentioned historical and psychometric considerations indicate that no single correct way exists to conduct an unstructured or semistructured interview. Interviewer style is strongly influenced by theoretical orientation and by practical considerations. Persons strongly influenced by client-centered theories tend to be nondirective and to avoid highly structured questions. This is consistent with the underlying belief that persons have the inner ability to change and organize their own behaviors. The goal of a client-centered interview, then, is to create the type of interpersonal relationship most likely to enhance this self-change. In contrast, a behavioral interview is more likely to be based on the assumption that change occurs because of specific external influences and consequences. As a result, behavioral interviews are relatively structured because they are directed toward obtaining specific information that would help to design strategies to alter external conditions. In addition, different interviewing styles and strategies work well with some clients but may be relatively ineffective with others.

A useful distinction is between a diagnostic interview and one that is more informal and exploratory. The goal of a diagnostic interview is to develop a specific diagnosis, which usually was formerly based on the multiaxial *DSM-IV* model (see Othmer & Othmer, 1994; R. Rogers, 2001; Sommers-Flanagan & Sommers-Flanagan, 2013) but now is evolving to be based on the *DSM-5* (American Psychiatric Association [APA], 2013) taxonomy. Developing a diagnosis might follow a five-step process in which the clinician develops diagnostic clues, considers these in relation to diagnostic criteria, takes a psychiatric history, and, based on this information, develops a diagnosis with corresponding estimates of prognosis (Othmer & Othmer, 1994). Such an interview is likely to be directive with a careful consideration for inclusion and exclusion criteria of different disorders. It is most likely to occur in a psychiatric or general medical setting. In contrast, many practitioners do not believe in the value of formal diagnosis and, accordingly, do not pursue a formal *DSM-5* diagnosis. Even those who do value formal diagnosis may believe that the purpose of the clinical interview is to understand context, history, and interviewee's perspective, and the full assessment can work toward determining a formal diagnosis. These interviewers might be more concerned with areas such as a client's coping style, social supports, family dynamics, or the nature of the disability. As such, their interviews might be less directive and more flexible. Again, neither style is right or wrong, but one style may be appropriate and effective in one context (or client), whereas it is ineffective or inappropriate in another context.

Often interviewers wish to construct a semistructured interview format by listing in sequence the types of questions they would like to ask the person. To construct such a list, interviewers might consult Table 3.1 to note possibly relevant areas (note that this list is not exhaustive). Each of these areas might then be converted into specific questions, often starting with the most broad and general question and becoming

Table 3.1 Checklist for an Assessment Interview and Case History

Presenting Problem and Its History

- Description of the problem
- Initial onset
- Changes in frequency
- Antecedents/consequences
- Intensity and duration
- Previous treatment
- Attempts to solve
- Formal treatment

Family Background

- Socioeconomic level
- Parents' occupations(s)
- Emotional/medical history
- Married/separated/divorced
- Family constellation
- Cultural background
- Parents' current health
- Family relationships
- Urban/rural upbringing

Personal History

Infancy

- Developmental milestones
- Family atmosphere
- Amount of contact with parents
- Early medical history
- Toilet training

Early and middle childhood

- Adjustment to school
- Academic achievement
- Hobbies/activities/interests
- Peer relationships
- Relationship with parents
- Important life changes

Adolescence

- All areas listed for early and middle childhood
- Presence of acting out (legal, drugs, sexual)
- Early dating
- Reaction to puberty
- Childhood abuse

Early and middle adulthood

- Career/occupation
- Interpersonal relationships
- Satisfaction with life goals
- Hobbies/interests/activities
- Romantic relationship/marriage
- Domestic violence
- Medical/emotional history
- Relationship with parents
- Economic stability
- Substance abuse

Late adulthood

- Medical history
- Ego integrity
- Reaction to declining abilities
- Economic stability

Miscellaneous

- Self-concept (like/dislike)
- Happiest/saddest memories
- Earliest memory
- Fears
- Somatic concerns (headaches, stomachaches, etc.)
- Events that create happiness/sadness
- Recurring/noteworthy dreams

progressively more specific as needed. For example, the first few areas might be converted into this series of questions:

- "Tell me about the most important concerns that you have right now."
- "How do these things affect you in your life?"
- "When did the difficulty first begin?"
- "How often does it occur?"
- "Have there been times when it has been better or worse?"
- "What happens after the behavior(s) occurs?"

Because clients vary regarding their personal characteristics (e.g., age, educational level, degree of cooperation) and type of presenting problem (e.g., childhood difficulties, legal problems, psychosis), interview questions necessarily need to vary from person to person. Furthermore, any series of questions should not be followed rigidly but with a certain degree of flexibility, to allow exploring unique but relevant areas that arise during the interview.

Good interviewing is difficult to define, partly because different theoretical perspectives exist regarding clinician–client interaction. Furthermore, clinicians achieve successful interviews not so much by what they do or say but by making sure they express the proper attitude. Whereas clinicians from alternative theoretical orientations might differ regarding areas such as their degree of directiveness or the type of information they should obtain, most agree that certain aspects of the relationship are essential (Patterson, 1989). These aspects include the interviewer's expression of sincerity, acceptance, understanding, genuine interest, warmth, and a positive regard for the worth of the person. If clinicians do not demonstrate these qualities, they are unlikely to achieve the goals of the interview, no matter how these goals are defined.

Patient ratings of the quality of interviews have been found to be dependent on the extent to which interviewers can understand the patient's emotions and detect emotional messages that are only partially expressed, particularly as these emotions are likely to be indirect and conveyed through nonverbal behaviors. Understanding a client's emotional responses is especially relevant in clinical interviews that focus on a client's personal difficulties. Typically, words are inadequate to accurately describe problem emotions, so interviewers must infer them from paraverbal or nonverbal expression. Reliance on nonverbal cues is highlighted by the assumption that nonverbal aspects of communication are a powerful method of conveying information. For example, eye contact can convey involvement; rigidity of posture might suggest client defensiveness; and hand movements often occur beyond the person's conscious intent, suggesting nervousness, intensity, or relaxation. Mehrabian (1972) supported this perspective with his estimates that the message received is 55% dependent on facial expression, 38% dependent on tone, and only 7% dependent on the content of what is said.

Interviewers vary in the extent to which they take notes during the interview. Some argue that note taking during an interview might increase a client's anxiety, raise questions regarding anonymity, increase the likelihood that a client will feel like an object under investigation, and create an unnatural atmosphere. In contrast, many interviewers counter these arguments by pointing out that a loss of rapport rarely

results solely from note taking during the interview, assuming, of course, that the interviewer still spends a sufficient amount of time attending to the client. Ongoing note taking is also likely to capture more details and result in less memory distortion than recording material after an interview has been completed. Thus, an intermediate amount of note taking during the interview is recommended. If the interview is audiotaped or videotaped, the reasons for this procedure need to be fully explained, along with the assurance of confidentiality and the procuring of signed consent. Although audiotape or videotape recording is often awkward at first, usually the interviewer and client quickly forget that it is occurring.

Interview Tactics

Numerous interview tactics and types of statements have been proposed and studied. These include the clarification statement, verbatim playback, probing, confrontation, understanding, active listening, reflection, feedback, summary statement, random probing, self-disclosure, perception checking, use of concrete examples, and therapeutic double binds. Additional relevant topics are the importance of eye contact, self-disclosure, active listening, and touch. These areas are beyond the scope of this chapter, but the interested reader is referred to excellent discussions by Cormier and Cormier (1998), Sommers-Flanagan and Sommers-Flanagan (2013), Sattler (2014), and Zuckerman (2005). The most relevant skills for interviewing do not come so much from memorizing interviewing tactics but from experiential practice and reviewing actual live or taped interview sessions. However, several important tactics of interviewing are described because they provide a general interviewing strategy.

Preliminaries

During the initial phase of the interview, practitioners need to ensure that they deal adequately with the next seven issues:

1. Organize the physical characteristics of the interview situation so that the room looks lived in but not untidy; utilize optimal lighting; and arrange seating so that the interviewer and client are neither too close nor too far and so that eye level is approximately equal.
2. Introduce themselves and indicate how they prefer to be addressed (Doctor, first name, etc.) and clarify how the client prefers to be addressed.
3. State the purpose of the interview, check the client's understanding of the process, and clarify any discrepancies between these two understandings.
4. Explain how the information derived from the interview will be used.
5. Describe the confidential nature of the information, the limits of confidentiality, and special issues related to confidentiality (e.g., how the information might be obtained and used by the legal system). Further, explain that the client has the right not to discuss any information he or she does not wish to disclose. If the information will be sent to other persons, obtain a signed release of information.
6. Explain the role and activities they would like the client to engage in, the instruments that are likely to be used in the assessment, and the total likely length

of time required. In some circumstances, this may be formalized into a written contract (Handelsman & Galvin, 1988).

7. Make sure that any fee arrangements have been clarified, including the hourly rate, total estimated cost, the amount the client versus a third party is likely to need to pay, and the interval between billing and the expected payment.

With the possible exception of fee arrangement (item 7), the preceding issues should be handled by a mental health practitioner rather than a secretary or receptionist. Covering these areas during the preliminary stages of the interview should reduce the likelihood of miscommunications and later difficulties.

Directive Versus Nondirective Interviews

The degree to which clinicians choose to be structured and directive during an interview depends on both theoretical and practical considerations. If time is limited, the interviewer will likely need to be direct and to the point. The interviewer will use a different approach for assessing a person who has been referred and will be returning to the referring person than for assessing a person before conducting therapy with him or her. An ambiguous, unstructured approach may make an extremely anxious person even more anxious, while a direct approach may prove more effective. A passive, withdrawn client also is likely to initially require a more direct question-and-answer style. As stated previously, a less structured style often encourages deeper client self-exploration, enables clinicians to observe the client's organizational abilities, and may result in greater rapport, flexibility, and sensitivity to the client's uniqueness.

Frequently, behavioral interviews are characterized as being structured and directed toward obtaining a comprehensive description of actual behaviors and relevant cognitions, attitudes, and beliefs (see Chapter 4). Behavioral interviewing is often contrasted with the more unstructured psychodynamic approach, which investigates underlying motivations and hidden dynamics and assesses information that may not be within the person's ordinary awareness. Typically, these approaches are perceived as competing and mutually exclusive. Haas, Hendin, and Singer (1987) pointed out that this either/or position is not only unnecessary but unproductive, because each style of interviewing provides different types of information that could potentially compensate for the other's weaknesses. Using both approaches might increase interview breadth and validity. Exploring multiple facets of the person may include direct behavioral data (public communication), self-description, and private symbolization (Leary, 1957). Each of these levels may be useful for different purposes, and the findings from each level might be quite different from one another.

Sequence of Interview Tactics

Most authors recommend that interviewers begin with open-ended questions and then, after observing the client's responses, use more direct questions to fill in gaps in their understanding (Harwood, Beutler, & Groth-Marnat, 2011; Othmer & Othmer, 2002; Sommers-Flanagan & Sommers-Flanagan, 2013). Although this sequence might begin with open-ended questions, it should typically lead to interviewer responses that are

intermediate in their level of directiveness, such as facilitating comments, requesting clarification, and possibly confronting the client with inconsistencies.

An important advantage of open-ended questions is that they require clients to comprehend, organize, and express themselves with little outside structure. This occasion is perhaps the only one in the assessment process that makes this requirement of clients, because most tests or structured interviews provide guidance in the form of specific, clear stimuli. When clients are asked open-ended questions, they will be most likely to express significant but unusual features about themselves. Verbal fluency, level of assertiveness, tone of voice, energy level, hesitations, and areas of anxiety can be noted. Hypotheses can be generated from these observations, and further questioning and testing can be used to test these hypotheses. In contrast to these advantages, open-ended questions can potentially provide an overabundance of detailed, vague, or tangential information.

Interviewer responses that show an intermediate level of directiveness are facilitation, clarification, empathy, and confrontation. Facilitation of comments maintains or encourages the flow of conversation. This might be accomplished verbally ("Tell me more … "; "Please continue … ") or nonverbally (eye contact, nodding). These requests for clarification might be used when clients indicate, perhaps through subtle cues, that they have not fully expressed something regarding the topic of discussion. Requests for clarification can bring into the open material that was only implied. In particular, greater clarification might be achieved by requesting the client to be highly specific, such as asking him or her to provide concrete examples (e.g., a typical day or a day that best illustrates the problem behavior). Empathic statements ("It must have been difficult for you") can also facilitate client self-disclosure.

Sometimes interviewers may wish to confront, or at least comment on, inconsistencies in a client's information or behavior. Carkhuff (1969) categorized the potential types of inconsistencies as being between what a person is versus what he or she wants to be, between what he or she is saying versus what he or she is doing, and between the person's self-perception versus the interviewer's experience of the person. A confrontation might also challenge the improbable content of what he or she is reporting (tall stories).

The purpose of confrontations during assessment is to obtain more in-depth information about the client. In contrast, therapeutic confrontations are used to encourage client self-exploration and behavior change. If a practitioner is using the initial interview and assessment as a prelude to therapy, this distinction is less important. However, a confrontational style can produce considerable anxiety, which should be created only if sufficient opportunity exists to work through the anxiety. Usually a client is most receptive to confrontations when they are posed either hypothetically as possibilities to consider or as curiosities on the part of the interviewer rather than as direct challenges. Confrontations also require a sufficient degree of rapport to be sustained; unless this rapport is present, confrontations may result in client defensiveness and a deterioration of the relationship.

Finally, direct, close-ended questions can be used to fill in gaps in what the client has reported. Thus, a continual flow can be formed between client-directed or client-organized responses and clinician-directed responses. This sequence, beginning with

open-ended questions, then moving to intermediately structured responses (facilitation, clarification, confrontation), and finally ending in directive questions, should not be rigid but should vary throughout the interview.

Comprehensiveness

The basic focus of an assessment interview should be to define the problem behavior (nature of the problem, severity, related affected areas) and its context (conditions that worsen or alleviate it, origins, antecedents, consequences). Interviewers might wish to use a checklist, such as the one in Table 3.1, to ensure they are covering the most relevant areas. In using such a checklist, the interviewer might begin with a general question, such as "How were you referred here?" or "What are some areas that concern you?" Observations and notes can then be made about the way the client organizes his or her responses, what he or she says, and the way he or she says it. The interviewer can use facilitating, clarifying, and confronting responses to obtain more information. Finally, the interviewer can review the checklist—for example, on family background—to see if all relevant areas were covered sufficiently. If some areas or aspects of areas were not covered, the interviewer might ask direct questions, such as "What was your father's occupation?" or "When did your mother and father divorce?" The interviewer can then begin the same sequence for personal history related to infancy, middle childhood, and so on. Table 3.1 is not comprehensive but is intended as a general guide for most interview situations. If practitioners generally evaluate specific client types (e.g., child abuse, suicide, those with brain impairments), this checklist may need amending and/or be used as an adjunct to commercially available structured interviews, such as the Personality Disorder Examination (Loranger, 1988), Neuropsychological Status Examination (Schinka, 1983), or Lawrence Psychological-Forensic Examination (Lawrence, 1984).

Avoidance of "Why" Questions

It is best to avoid "why" questions because they can increase client defensiveness. A "why" question may sound accusatory or critical and thus forces the client to account for his or her behavior. In addition, clients may become intellectual in this situation, thereby separating themselves from their emotions. An alternative approach is to preface the question with either "What is your understanding of … " or "How did it occur that … " rather than "why?" These options are more likely to result in a description rather than a justification and to keep clients more centered on their emotions.

Nonverbal Behaviors

Interviewers should also be aware of their own as well as their clients' nonverbal behaviors. In particular, interviewers might express their interest by maintaining eye contact, being facially responsive, and attending verbally and nonverbally, such as through occasionally leaning forward.

Concluding the Interview

Any interview is bound by time constraints. An interviewer might help to ensure observance of these constraints by alerting the client when only 5 or 10 minutes remain until

the arranged completion of the interview. Doing this allows the client or interviewer to focus on final relevant information. There should also be an opportunity for the client to ask any questions or provide comments. At the end of an interview or assessment session, the interviewer may summarize the main themes of the interview and, if appropriate, make any recommendations.

MENTAL STATUS EVALUATION

The mental status exam was originally modeled after the physical medical exam; just as the physical medical exam is designed to review the major organ systems, the mental status exam reviews the major systems of psychiatric functioning (appearance, cognitive function, insight, etc.). Since its introduction into American psychiatry by Adolf Meyer in 1902, it has become the mainstay of patient evaluation in most psychiatric settings. Most psychiatrists consider it as essential to their practice as the physical examination is in general medicine (Rodenhauser & Fornal, 1991).

A mental status examination can be used as part of a formal psychological assessment for a variety of reasons. A brief mental status examination might be appropriate before assessment to determine the appropriateness of more formal psychological testing. If, for example, a patient were unable to determine where he or she was and had significant memory impairments, testing with most instruments might be too difficult and could result in needless distress. Brief screenings might also be used to determine basic case management issues, such as hospitalization or placing a patient under close observation. A mental status examination can be used as part of an assessment using formal psychological tests. The "raw" data from the exam can be selectively integrated with general background information to present a coherent portrait of the person and assist in diagnosis.

Despite its popularity among psychiatrists, this form of interviewing is not typically used by psychologists, partly because many areas reviewed by the mental status exam are already covered during the assessment interview and through the interpretation of psychological test results. Many psychological tests cover these areas in a more precise, in-depth, objective, and validated manner with scores that can be compared to appropriate norms. A client's appearance, affect, and mood are usually noted by attending to behavioral observations. A review of the history and nature of the problem is likely to pick up areas such as delusions, misinterpretations, and perceptual disorders (hallucinations). Likewise, interview data and psychological test results typically assess a client's fund of knowledge, attention, insight, memory, abstract reasoning, and level of social judgment. However, the mental status examination reviews all of the preceding areas in a relatively brief, systematic manner. Furthermore, there are situations, such as intakes in an acute medical or psychiatric hospital, where insufficient time is available to evaluate the client with psychological tests.

Numerous sources in the psychiatric literature provide thorough guidelines for conducting a mental status exam (Crary & Johnson, 1981; Othmer & Othmer, 2002; Robinson, 2001; Sadock & Sadock, 2010; Sommers-Flanagan & Sommers-Flanagan, 2013), and R. Rogers (2001) has provided a review of the more structured mental status exams. This literature indicates that practitioners vary widely in how they

conduct mental status examinations. The most unstructured versions involve merely the clinician's use of the mental status examination as a set of general guidelines. The more structured versions range from comprehensive instruments that assess both general psychopathology and cognitive impairment to those that focus primarily on cognitive impairment. For example, the comprehensive North Carolina Mental Status Examination (Ruegg, Ekstrom, Evans, & Golden, 1990) includes 36 items that are rated on a 3-point scale (not present, slight or occasional, marked or repeated) to cover the important clinical dimensions of physical appearance, behavior, speech, thought processes, thought content, mood, affect, cognitive functioning, orientation, recent memory, immediate recall, and remote memory. Another similar comprehensive instrument is the Missouri Automated Mental Status Examination Checklist (Hedlund, Sletten, Evenson, Altman, & Cho, 1977), which requires the examiner to make ratings on nine areas of functioning: general appearance, motor behavior, speech and thought, mood and affect, other emotional reactions, thought content, sensorium, intellect, and insight and judgment. The checklist includes 119 possible ratings, but the examiner makes ratings in only those areas he or she judges to be relevant.

Despite extensive development, the more comprehensive mental status examinations have not gained wide acceptance. In contrast, the narrower structured mental status examinations that focus more exclusively on cognitive impairment are used quite extensively. One of the most popular has been the Mini Mental State Examination (Folstein, Folstein, & McHugh, 1975). It includes 11 items designed to assess orientation, registration, attention, calculation, and language. It has excellent interrater and test-retest reliabilities (usually well above .80), correlates with WAIS IQs (.78 for verbal IQ), and is sensitive to global and left-hemisphere deficits (but not right-hemisphere impairment; R. Rogers, 2001; Tombaugh, McDowell, Kristjansson, & Hubley, 1996). Clinicians who wish to develop knowledge and skills in conducting mental status examinations are encouraged to consult the preceding sources.

The following descriptions of the typical areas covered serve as a brief introduction to this form of MSE interviewing. The outline is organized around the categories recommended by Crary and Johnson (1981), and a checklist of relevant areas is included in Figure 3.1. Interviewers can answer the different areas on the checklist either during or after a mental status examination. The tabled information can then be used to answer relevant questions relating to the referral question, to help in diagnosis, or to add to other test data. Such a checklist is important because clinicians not using similar checklists have been found to often omit crucial information (Ruegg et al., 1990).

General Appearance, Behavior, and Relatedness

This area assesses material similar to that requested in the "behavioral observations" section of a psychological report (see Chapter 15). A client's clothing, posture, gestures, speech, personal care/hygiene, and any unusual physical features, such as physical disabilities, tics, or grimaces, are noted. Attention is given to the degree to which the client's behavior conforms to social expectations, but this is placed in the context of his or her culture and social position. Relatedness toward the evaluator is also an important factor to note. Additional important areas are facial expressions, eye contact, activity level, degree of cooperation, notable physical characteristics, and attentiveness.

Mental Status Evaluation

Appearance		**Within Norm**	**Notable**	**Details**
	Grooming			
	Motor Activity			
	Coordination/Gait			
	Notes on Appearance			

Relatedness	**Within Norm**	**Notable**	**Details**
	Cooperative Friendly Relaxed Good eye contact	Hostile Guarded Seductive Poor eye contact	
	Notes on Relatedness		

Speech/ Language		**Within Norm**	**Notable**		**Details**
	Receptive				
	Expressive		Quiet Loud Slow	Clutter/Stutter Rapid Pressured	
	Notes on Speech/Language				

Affect/Mood		**Within Norm**	**Notable**		**Details**
	Affect	Expressive Good Range	Flat Constricted Angry Mood-Incongruent	Anxious Sad Labile Inappropriate to Situation	
	Mood	Euthymic	Elevated Depressed	Angry	
	Notes on Affect/Mood				

Thought Process	**Within Norm**	**Notable**		**Details**
	Goal directed Logical Abstract Reasoning	Tangential Circumstantial Magical Concrete	Flight of Ideas Slow Rapid Loose	
	Notes on Thought Process			

Figure 3.1 Format for mental status and history

Mental Status Evaluation

Thought Content		**Present**	**Not Present**	**Details**
	Hallucinations			
	Delusions			
	Depressive Ideation			
	Suicidality			
	Aggressiveness			
	Homicidality			
	Notes on Thought Content			

Memory		**Intact**	**Impaired**	**Details**
	Short-Term			
	Long-Term			
	Notes on Memory			

Attention/ Concentration	**Within Norm**	**Notable**	**Details**
	Notes on Attention/Concentration		

Alertness/ Orientation	**Within Norm**	**Notable**	**Details**
	Alert Oriented	Lethargic Disoriented Hypervigilant	
	Notes on Alertness		

Judgment/ Planning		**Within Norm**	**Notable**	**Details**
	Judgment			
	Impulse Control			
	Notes on Judgment/Planning			

Insight	**Within Norm**	**Notable**	**Details**
	Notes on Insight		

Figure 3.1 ***(Continued)***

Is the client friendly, hostile, seductive, or indifferent? Do any bizarre behaviors or significant events occur during the interview? In particular, speech might be fast or slow, loud or soft, or include a number of additional unusual features. Figure 3.1 includes a checklist of relevant areas of behavior, appearance, and relatedness.

Speech and Language

Clients' speech and language are often proxies for their thought processes, as they relate to the primary mode of communicating thoughts to the outside world. They help clinicians determine the possibility of poor or exceptional cognitive functioning, focus, and confusion and possible thought disorder. Additionally, speech and language often highlight interpersonal characteristics, such as shyness, anxiety interacting with others, and aggressiveness. Clinicians should evaluate in general how well individuals understand language, as evidenced by responding appropriately to directions and conversations (known as *receptive* language). *Expressive* language, in contrast, relates to the client's actual speech and use of language. *Speech* relates to the quality of speaking, such as quiet, loud, rapid, slow, and so on. *Language* relates to the words used, including having difficulty with word finding, using complex and appropriate vocabulary, or misusing words often.

Feeling (Affect and Mood)

A client's *mood* refers to the dominant emotion reported during the interview, whereas *affect* refers to the client's outwardly projected range of emotions. Information related to affect is inferred from the content of the client's speech, facial expressions, and body movements. The type of affect can be judged according to variables such as its depth, intensity, duration, and appropriateness. The client might be cold or warm, distant or close, labile, or, as is characteristic of schizophrenia, his or her affect might be blunted or flattened. The client's mood might also be euphoric, hostile, anxious, or depressed, and an examiner should note the level of congruence between mood and affect.

Perception and Thinking

Perception

Different clients perceive themselves and their world in a wide variety of ways. It can be diagnostically important to note whether there are any illusions or hallucinations. For example, the presence of auditory hallucinations is most characteristic of those with schizophrenia, whereas vivid visual hallucinations are more characteristic of persons with organic brain syndromes.

Intellectual Functioning

Any assessment of higher intellectual functioning needs to be made in the context of a client's educational level, socioeconomic status, and familiarity and identification with a particular culture. If a low level of intellectual functioning is consistent with a general pattern of poor academic and occupational achievement, a diagnosis of intellectual

disability might be supported. However, if a person performs poorly on tests of intellectual functioning and yet has a good history of achievement, organicity might be suspected.

Intellectual functioning typically involves reading and writing comprehension, general fund of knowledge, ability to do arithmetic, and the degree to which the client can interpret abstract language, such as proverbs. Throughout the assessment, clinicians typically note the degree to which the client's thoughts and expressions are articulate versus incoherent. Sometimes clinicians might combine assessments of intellectual functioning with some short, formal tests such as the Bender, an aphasia screening test, or even portions of the WAIS or WISC.

Orientation

The ability of clients to be oriented can vary in the degree to which they know who they are (person), where they are (place), and when current and past events have occurred or are occurring (time). Clinical observation indicates the most frequent type of disorientation is for time; disorientation for place and person occurs less frequently. When disorientation does occur for place, and especially for person, the condition is likely relatively severe. Disorientation is most consistent with organic conditions. If a person is oriented in all three spheres, this is frequently abbreviated as "oriented X3."

Related to the orientation of clients is their *sensorium*, which refers to how intact their physical sensory processes are to receiving and integrating information. *Sensorium* might refer to hearing, smell, vision, and touch and might range from being clouded to clear. Can the client attend to and concentrate on the outside world, or are these processes interrupted? The client might experience unusual smells, hear voices, or have the sense that his or her skin is tingling. *Sensorium* can also refer to the client's level of consciousness, which may vary from hyperarousal and excitement to drowsiness and confusion. Disorders of a client's sensorium often reflect organic conditions but may also be consistent with psychosis.

Memory, Attention, and Concentration

Because memory acquisition and retrieval require attention and concentration, these three functions are frequently considered together. Long-term memory is often assessed by requesting information regarding the client's general fund of information (e.g., important dates, major cities in a country, three major heads of state since 1900). Some clinicians include the Information or Digit Span subtests from the WAIS/WISC or other formal tests of a similar nature. Recall of a sentence or paragraph might be used to assess short-term memory for longer, more verbally meaningful information. In addition, clients' long-term memory might be evaluated by measuring recall of major life events, and the accuracy of their recall can be compared with objective records of these events (e.g., year graduated from high school, date of marriage). It is often useful to record any significant distortions of selective recall in relation to life events, as well as to note the client's attitudes toward his or her memory.

Short-term memory might be assessed either by requesting that clients recall recent events (e.g., most recent meal, how they got to the appointment) or by having them repeat digits forward and backward. Again, the WAIS/WISC Digit Span subtest, or

a similar version of it, might be used. Serial sevens (counting forward by adding 7 each time) can be used to assess how distractible or focused they are. Persons who are anxious and preoccupied have a difficult time with serial sevens as well as with repeating digits forward and, especially, backward.

Insight and Judgment

Clients vary in their ability to interpret the meaning and impact of their behavior on others. They also vary widely in their ability to provide for themselves, evaluate risks, and make plans. Adequate insight and judgment involve developing and testing hypotheses regarding their own behavior and the behavior of others. Clients also need to be assessed to determine why they believe they were referred for evaluation and, in a wider context, their attitudes toward their difficulties. How do they relate their past history to current difficulties, and how do they explain these difficulties? Where do they place the blame for their difficulties? Based on their insights, how effectively can they solve problems and make decisions?

Thought Content

A client's speech can often be considered a reflection of his or her thoughts. The client's speech may be coherent, spontaneous, and comprehensible or may contain unusual features. It may be slow or fast, be characterized by sudden silences, or be loud or unusually soft. Is the client frank or evasive, open or defensive, assertive or passive, irritable, abusive, or sarcastic? Consideration of a person's thoughts is often divided into thought content and thought processes. Thought content such as delusions might suggest a psychotic condition, but delusions may also be consistent with certain organic disorders, such as dementia or chronic amphetamine use. The presence of compulsions or obsessions should be followed up with an assessment of the client's degree of insight into the appropriateness of these thoughts and behaviors. Thought processes such as the presence of rapid changes in topics might reflect flighty ideas. The client might also have difficulty producing a sufficient number of ideas, include an excessive number of irrelevant associations, or ramble aimlessly.

INTERPRETING INTERVIEW DATA

Interpreting and integrating interview data into the psychological report inevitably involves clinical judgment. Even with the use of structured interviews, the clinician still must determine which information to include or exclude. Thus, all the potential cautions associated with clinical judgment need to be considered (see Chapter 1). Caution is particularly important because life decisions and the success of later treatment may depend on conclusions and recommendations described in the report.

Several general principles can be used to interpret interview data. The interview is the primary instrument that clinicians use to develop tentative hypotheses regarding their clients. Thus, interview data can be evaluated by determining whether these hypotheses are supported by information outside the interview. Interview data that are supported by test scores can be given greater emphasis in the final report if they are

relevant to the referral question. Even material that is highly supported throughout different phases of the interview process should not be included unless it relates directly to the purpose of the referral.

There is a continuum in handling interview information that varies according to the extent the information will be interpreted. On one hand, the information might be merely reorganized into a chronological history of the person's life. This method would emphasize repeating the information in as objective and accurate a manner as possible. Typically this is done in the history section of a psychological report. On the other hand, interview data can be considered raw data to be interpreted. It is thus similar to the data from formal psychological tests. It might, therefore, be used to make inferences related to a client's personality, coping style, or mood and affect.

One method of organizing interview information is to use the information to develop a coherent narrative of the person's life. For example, describing how early family patterns resulted in emotionally sensitive areas ("scar" tissue) can be used to help explain current symptom patterns and difficulties in interpersonal relationships. A different sort of history might trace how interest in a vocation was first begun (early childhood daydreams regarding occupations) and how this progressed and developed as the person matured. Another person might present difficulties related to authority figures. Specific details relating to these difficulties might emerge, such as the client feeling like a martyr and eventually inappropriately expressing extreme anger toward authority figure(s). A careful review of the client's history might reveal how he or she becomes involved in these recurring relationships and how he or she typically attempts to resolve them. Persons who are frequently depressed might distance themselves from others by their behavior and then be confused about why relationships seem to be difficult. Often these themes emerge during a carefully conducted interview, yet aspects of the themes (or the entire themes themselves) are not apparent to the interviewee.

Interview data might also be organized around various domains (see further discussion in Chapter 15). A grid can be used to organize these domains. The various domains might be listed on the left side of the grid with the top of the grid listing the sources of data (of which the interview might be one of a variety of sources of information; see Table 15.2 in Chapter 15). Domains might include mood and affect, cognitions, level of resistance, interpersonal patterns, or coping style. This approach treats interview data in much the same manner as data from psychological tests.

There is no one strategy for sensitizing interviewers to the types and patterns of recurring themes they may encounter during interviews. Inevitably, clinical judgment is a significant factor. The accuracy and types of judgment depend on the theoretical perspective of the interviewer, knowledge regarding the particular difficulty the interviewer is investigating, past experience, types of questions asked, and purpose of the interview.

STRUCTURED INTERVIEWS

Standardized psychological tests and structured interviews were developed to reduce the problems associated with open-ended interviews. They serve both to structure the stimuli presented to the person and to reduce the (potentially biased) role of clinical

judgment. Because structured interviews generate objective ratings on consistent areas, they have the advantage of making possible comparisons between one case or population and the next. Typically, these interviews vary in their degree of structure, the relative expertise required to administer them, and the extent to which they serve as screening procedures designed for global measurement or as tools used to obtain specific diagnoses.

Before structured interviews could be developed, clear, specific criteria had to be created relating to symptom patterns and diagnoses. Developing these clear, specific criteria ideally helped to reduce the amount of error caused by vague guidelines for exclusion or inclusion in different categories (*criterion variance*). These criteria then needed to be incorporated into the interview format and interview questions. *Information variance* refers to the variability in amount and type of information derived from interviews with clients. In most unstructured interviews, information variance is caused by the wide differences in content and phrasing because of factors such as the theoretical orientation or style of the interviewer. Structured interviews correct for this by utilizing the same or similar questions for each client.

The first popular system of specific criterion-based diagnosis was developed by Feighner et al. (1972) and provided clear, behaviorally oriented descriptions of 16 psychiatric disorders based on the *DSM-II* (APA, 1968). Clinicians using the Feighner criteria were found to have an immediate and marked increase in interrater diagnostic reliability. The descriptions of and relevant research on the Feighner criteria were published in Woodruff, Goodwin, and Guze's (1974) book, *Psychiatric Diagnosis*. Several interviews, such as the Renard Diagnostic Interview (Helzer et al., 1981), incorporated the Feighner criteria. Spitzer, Endicott, and Robins (1978) further altered and elaborated the Feighner criteria to develop the Research Diagnostic Criteria (RDC). Simultaneous with the development of the RDC, Endicott and Spitzer (1978) developed the SADS, which was based on the new RDC. When new versions of the *DSM* were published (APA, 1980, 1987, 1994, 2000, 2013), revisions of previous interviews typically incorporated the most recent *DSM* criteria, along with elements of the Feighner criteria and/or the RDC.

As noted earlier, the reliability of structured interviews has been found to vary depending on the specificity or precision of the rating or diagnosis. Whereas the highest reliabilities have been found for global assessment (presence/absence of psychopathology), much lower reliabilities have generally been found for the assessment of specific types of behaviors or syndromes. Likewise, high reliabilities have been found for overt behaviors, but reliability has been less satisfactory for more covert aspects of the person, such as obsessions, fears, and worries. Reliability also tends to be lower when clinicians are asked to attempt exact estimates of behavioral frequencies and for inferences of multifaceted aspects of the person derived from complex clinical judgments.

Most early studies on validity were based on item content (content validity) or degree of accuracy in distinguishing between broad areas of psychopathology (psychiatric/nonpsychiatric). More recent trends have attempted to assess the accuracy of far more specific areas. However, most validity studies have suffered from an absence of clear, commonly agreed-upon criteria. Although structured interviews were attempts to improve on previous, imperfect instruments (unstructured interviews, standardized tests), the structured interviews themselves could not be compared with anything

better. For example, the "procedural validity" strategy is based on comparing lay interviewers' diagnoses with diagnoses derived from trained psychiatrists. Although the psychiatrist's diagnosis may be better than the layperson's, diagnoses by trained psychiatrists still cannot be said to be an ultimate, objective, and completely accurate standard. Furthermore, there is confusion about whether actual validity is being measured (which would assume psychiatrists' diagnoses are the true, accurate ones) or merely a version of interrater agreement. At the core of this issue is the very nature of how diagnosis is defined and the degree to which it is actually helpful in treatment (see Beutler & Malik, 2002; Widiger & Clark, 2000).

Future studies need to involve aspects of what has previously been discussed as construct validity. The focus on construct validity means looking more carefully at structured interviews in relationship to etiology, course, prognosis, and treatment utility relating to areas such as the appropriate selection of treatments and the likelihood of favorable responses to these treatments. Validity studies also need to look at the interaction between and implications of multiple criterion measures, including behavioral assessment, checklists, rating scales, self-report inventories, biochemical indices, and neuropathological alterations.

Since the mid-1970s, there has been a proliferation of structured interviews for a wide range of areas. Clinicians working in specific areas often select structured interviews directed toward diagnosing the disorders they are most likely to encounter. For example, some situations might benefit from using the Anxiety Disorders Interview Schedule–IV (DiNardo, Brown, & Barlow, 1994) to make clear distinctions between anxiety disorders and substance abuse and between psychosis and major affective disorders. Other contexts might be best served by the Eating Disorder Examination (EDE; Z. Cooper & Fairburn, 1987) or the Structured Interview for *DSM-IV* Dissociative Disorders (SCID-D; Steinberg, 1993). Three categories of structured interviews with representative frequently used instruments are included in Table 3.2 and have been extensively reviewed in R. Rogers's (2001) *Handbook of Diagnostic and Structured Interviewing*. One consideration in selecting these instruments is that, because most structured interviews undergo continuous revisions, the most up-to-date research should be consulted to ensure that practitioners obtain the most recently revised versions. The next section provides an overview of structured interviews that are used most frequently and are the most extensively researched.

Structured Clinical Interview for the *DSM*

The SCID (First, Spitzer, Gibbon, & Williams, 1996, 1997; Spitzer et al., 1987) is the most frequently used structured interview (see description and updates at www.scid4.org and www.appi.org/pages/scid-5.aspx). It is a clinician-administered, comprehensive broad-spectrum instrument that adheres closely to the *DSM* decision trees for psychiatric diagnosis. A certain degree of flexibility is built in so that administration can be tailored to different populations and contexts. Thus, slightly different forms are used for psychiatric patients (SCID–In/Patient), outpatients (SCID–Out/Patients), and nonpatients (SCID–Non/Patients). Criticisms that the early version of the SCID had sacrificed clinical information so that it would be more user-friendly for clinicians resulted in a revision that emphasized a clear, easy-to-use version for clinical contexts

Table 3.2 Frequently Used Structured Interviews by Categories

I.	Assessment of clinical disorders
	Schedule of Affective Disorders and Schizophrenia (SADS) and Schedule of Affective Disorders and Schizophrenia for School-Age Children (K-SADS)
	Diagnostic Interview Schedule (DIS) and Diagnostic Interview for Children (DISC)
	Structured Clinical Interview for DSM-IV (SCID)
	Diagnostic Interview for Children and Adolescents (DICA)
II.	Assessment of personality disorders
	Structured Interview for DSM-IV Personality Disorders (SIDP)
	Personality Disorder Examination (PDE)
	Structured Clinical Interview for DSM-III-R Personality Disorders (SCID–II)
III.	Focused structured interviews
	Anxiety Disorders Interview Schedule (ADIS)
	Diagnostic Interview for Borderlines (DIB)
	Psychopathy Checklist (PCL)
	Structured Interview for DSM-IV-Dissociative Disorders (SCID-D)
	Structured Interview of Reported Symptoms (SIRS)
	Psychosocial Pain Inventory (PSPI)
	Comprehensive Drinker Profile (CDP)
	Eating Disorder Examination (EDE)
	Structured Interview of Sleep Disorders (SIS-D)
	Substance Use Disorders Diagnostic Schedule (SUDDS)

(the SCID-Clinical Version or SCID-CV; First et al., 1997) and a longer, more in-depth version for research (SCID-I or SCID-Research Version; First, Spitzer et al., 1996). A new version aligned with the *DSM-5* is now available as well (SCID-5; First, Williams, Karg, & Spitzer, 2015). Whereas these versions of the SCID are directed toward what used to be known as Axis I diagnoses, a separate version has been developed for the diagnosis of personality disorders (SCID-II; Spitzer, Williams, Gibbon, & First, 1990). A further variation, the SCID-D-Revised (Steinberg, 1993), was developed (though not by the team who developed the SCID) using *DSM-IV* criteria for the assessment of dissociative disorders. The SCID and its variations include several open-ended questions as well as a skip structure, which enables the interviewer to branch into new areas depending on the client's previous responses. Because clinical judgment is essential throughout the interview, the SCID should be administered only by trained professionals. To increase incremental validity, the authors encourage the inclusion of relevant additional data in making final diagnostic decisions.

The SCID, along with its variations, is the most comprehensive structured interview available. As a result, administration time can be considerable, even with the built-in screening questions and skip structure. Many individual clinicians and treatment sites deal with this problem by primarily administering the modules they are

most concerned with. For example, a treatment center specializing in substance abuse might administer the module for Psychoactive Substance Use Disorders along with the SCID-II when the comorbidity of personality disorders is suspected. Administration time might also be reduced by administering the computerized mini-SCID (First, Gibbon, Williams, & Spitzer, 1996), which has been designed to screen for possible (formerly) Axis I disorders. In addition, a computerized SCID-II (AutoSCID-II; First, Gibbon et al., 1996) that can potentially reduce clinician time is available. Although it can be administered by telephone, this procedure is discouraged, given the poor agreement between telephone and face-to-face diagnoses (Cacciola, Alterman, Rutherford, McKay, & May, 1999).

The reliability studies have resulted in overall moderate, but quite variable, test-retest and interrater reliabilities (First & Gibbon, 2004). For example, interrater agreement using the SCID-II for common diagnostic categories ranges between .40 and .86 with a mean of .59 (First, Spitzer, Gibbon, & Williams, 1995). Riskind, Beck, Berchick, Brown, and Steer (1987) found that several difficult-to-distinguish diagnostic categories had relatively good levels of interrater agreement. These included generalized anxiety disorders (.79, 86% agreement), depressive disorders (.72, 82% agreement; Riskind et al., 1987), panic disorders (k = .86), and major depression (k = .81; J. Reich & Noyes, 1987). Test-retest reliabilities over a 2-week interval for psychiatric patients was fair to good (overall weighted kappas = .61) but poor for nonpatients (overall weighted kappas = .37; J. B. Williams et al., 1992).

For the most part, validity studies of the SCID have assumed that *DSM-IV* diagnoses are the benchmark for making comparisons of diagnostic accuracy. Thus, "procedural validity" has often been assumed since the SCID has closely paralleled the diagnostic criteria derived from the *DSM* (R. Rogers, 2001). A representative validity study found good agreement (k = .83) between interviewer ratings and cross ratings of interviewer videotapes by two senior psychiatrists (Maziade et al., 1992). Other studies have found considerable diagnostic overlap within (formerly) Axis I disorders and between (formerly) Axis I and personality disorders (Alnacs & Torgerson, 1989; Brawman-Mintzer et al., 1993). However, evaluating the meaning of this overlap is difficult because the extent to which it is caused by instrument error versus true comorbidity (i.e., the frequent co-occurrence of anxiety and depression) is difficult to determine. In contrast to these mostly favorable studies, a number of studies have found generally poor agreement between SCID and clinician-based diagnosis (Shear et al., 2000; Steiner, Tebes, Sledge, & Walker, 1995). In summary, the strength of the SCID is its impressive breadth of coverage, use of modules targeted toward specific areas, and close parallel with the *DSM*. Its weaknesses are its wide variation in reliability and its need for further validity studies, particularly studies that relate it to other diagnostic measures.

Schedule for Affective Disorders and Schizophrenia

The SADS (Endicott & Spitzer, 1978) is a clinician-administered, extensive, semistructured interview that has been one of the most widely used structured interviews for clinical research purposes. Although it was originally designed for differential diagnosis between affective disorders and schizophrenia, it has evolved to include a much

wider range of symptoms and allows the interviewer to consider many different diagnostic categories. A wide range of disorders is considered within the SADS, but its primary strength lies in obtaining fine detail regarding different subtypes of affective disorders and schizophrenia (R. Rogers, Jackson, & Cashiel, 2004). The interview rates clients on six gradations of impairment from which diagnoses are reached using the clear, objective categories derived from Spitzer et al.'s (1978) RDC. The SADS is divided into adult versions for current symptoms, occurrence of lifetime symptoms, and degree of change. There is a further version for the assessment of children's difficulties (K-SADS or Kiddie-SADS). Two modifications for the SADS have been the inclusion of anxiety disorders (SADS-LA; Fyer, Endicott, Manuzza, & Klein, 1985, 1995) and eating disorders (EAT-SADS-L; Herzog, Keller, Sacks, Yeh, & Lavori, 1992).

Adult Version

The adult version of the SADS (Endicott & Spitzer, 1978) is designed to be administered in two different parts, the first focusing on the client's current illness and the second on past episodes. This division roughly corresponds with the three different versions of the SADS. The first is the regular version (SADS), the second is the lifetime version (SADS-L, which is actually the second half of the SADS), and the third is the SADS-C, which measures changes in the client. The SADS-L is directed toward diagnosing the possible presence of psychiatric disturbance throughout the person's life. The SADS and SADS-L are used most extensively. Because the questions in the SADS are directed toward current symptoms and those symptoms experienced 1 week before administration, it is most appropriate for administration when the client is having current difficulties. In contrast, the SADS-L is most appropriate when there is no current, acute illness. To make accurate ratings, interviewers are allowed to use a wide range of sources (client's family, medical records) and ask a number of different questions. Final ratings are made on a 6-point Likert-type scale. Administration involves more than 200 items and takes from 1.5 to 2 hours and should be conducted only by a psychiatrist, psychologist, or psychiatric social worker. The end product is eight summary scales:

1. Mood and ideation
2. Endogenous features
3. Depressive-associated features
4. Suicidal ideation and behavior
5. Anxiety
6. Manic syndrome
7. Delusions-hallucinations
8. Formal thought disorder

Interrater reliabilities for the specific diagnostic categories have been found to be quite high, with the exception of the Formal Thought Disorder Scale (Endicott & Spitzer, 1978). The low reliability of this scale may have been because few of the patients in the Endicott and Spitzer sample showed clear patterns of disordered thoughts, which

resulted in high variability for the ratings. Test-retest reliabilities were likewise good, ranging from .88 for manic disorders to .52 for chronic and intermittent depressive disorder (Spiker & Ehler, 1984). The exception was a low reliability for schizoaffective, depressed (.24), but this was likely due to the small number of patients included in this category, which resulted in limited variance. Using a different and possibly more appropriate statistical method, reliability increased to .84. Overall, the SADS has demonstrated excellent reliability, particularly for interrater and test-retest reliabilities related to current episodes of psychiatric disturbance.

Validity studies have been encouraging because expected relationships have been found between SADS scores and external measures of depression, anxiety, and psychosis. For example, M. H. Johnson, Margo, and Stern (1986) found that relevant SADS measures could effectively discriminate between patients with depression and paranoid and nonparanoid schizophrenia. In addition, the SADS depression measures effectively rated the relative severity of a patient's depression. For example, Coryell et al. (1994) found clear consistency between different levels of depression. The authors suggest that incremental validity might be increased by having clients referred for a medical examination to screen out physical difficulties that might be resulting in central nervous system dysfunction. The authors also recommend that interviewers try to increase validity by always including the best available information (family history, structured tests, other rating schedules) before making final ratings. The SADS has been used to predict the clinical features, course, and outcome of various disorders, including major depressive disorder (Coryell et al., 1994), schizophrenia (Stompe, Ortwein-Swoboda, Strobl, & Friedman, 2000), and bipolar disorder (Vieta et al., 2000). A number of studies have also effectively used the SADS to detect family patterns of schizophrenia (Stompe et al., 2000) and obsessive-compulsive disorders (Bienvenu et al., 2000).

Child Version

The SADS for School-Age Children (Kiddie-SADS-P, K-SADS-P; Ambrosini, 2000; Puig-Antich & Chambers, 1978) is a semistructured interview developed for children between ages 6 and 18. The K-SADS has come out in versions to be used in epidemiological research (K-SADS-E), to assess present and lifetime psychopathology (K-SADS-P/L), and to assess current levels of symptomology (K-SADS-P). Although much of the K-SADS is based on research with major depressive disorders of prepubertal children, it also covers a wide range of other disorders, such as phobias, conduct disorders, obsessive-compulsive disorders, and separation anxiety.

The interview should be administered by a professional clinician who has been trained in the use of the K-SADS and is familiar with *DSM* criteria. All versions are administered to both the parent and the child. Any discrepancies between the two sources of information are clarified before final ratings are made. Total administration time is approximately 1.5 hours per informant (3 hours total). The first phase is a 15- to 20-minute unstructured interview in which rapport is developed as well as an overview of relevant aspects of history, including the frequency and duration of presenting symptoms, their onset, and whether the parents have sought previous treatment. This interview is followed by structured questions regarding symptoms, which are rated on a Likert scale, with 1 representing "not at all" and 7 indicating

that they are "extreme." A skip structure is built into the format so that interviewers can omit irrelevant questions. Interviewers are allowed to use their judgment regarding the wording and the type and number of questions. Finally, ratings are made regarding behavioral observations (appearance, attention, affect). Interviewers are also asked to rate the completeness and reliability of the interview and to make a global assessment of pathology (degree of symptomatology and level of impairment).

Test-retest and interrater reliability for the K-SADS has been good with a general trend for each version to have improved reliabilities. Ambrosini (2000), for example, reported that the K-SADS-P/L had test-retest reliabilities ranging from 1.00 (lifetime occurrence of major depression) to .55 (for lifetime occurrence for attention-deficit/hyperactivity disorder). However, overall reliabilities have been lower for the K-SADS (and K-SADS-III-R) than for the adult SADS, although this is to be expected given the relative changeability and less well-developed language skills found with children (Ambrosini, Metz, Prabucki, & Lee, 1989; Chambers et al., 1985). Validity studies indicate that relevant K-SADS measures correlated highly with diagnoses for conduct disorders, schizophrenia, and depression (Apter, Bleich, Plutchik, Mendelsohn, & Tyrano, 1988). Additional expected correlations have been found between SADS measures and ratings of adolescent mood (Costello, Benjamin, Angold, & Silver, 1991) and the Child Behavior Checklist (CBCL; Achenbach & Edelbrock, 1983; Ambrosini, 2000). Finally, follow-up studies on adolescents diagnosed with disorders (e.g., depression) have found a continued risk for later affective difficulties (Lewinsohn, Rohde, Klein, & Seeley, 1999).

Collectively, the different versions of the SADS provide a thorough, well-organized interview with unparalleled coverage of the subtypes and gradations of the severity of mood disorders. The SADS has also been well accepted in research and clinical settings. It has strong interrater reliability, provides good ratings of symptom severity, measures associated symptoms, includes guidelines for possible malingering, and has strong evidence of convergent validity (see R. Rogers, 2001; R. Rogers et al., 2004). In contrast, its weaknesses include a relatively narrow band of diagnosis compared with some of the other available instruments, such as the SCID or DIS. In addition, the diagnoses are based on RDC rather than the more recent *DSM-IV-TR* or *DSM-5* criteria. This criticism is somewhat moderated, however, by many of the RDC and *DSM* criteria being nearly the same, especially for childhood disorders. Finally, administration and interpretation of the SADS require extensive training (usually 1 week) as well as a good working knowledge of differences between the SADS/RDC and *DSM* criteria.

Diagnostic Interview Schedule

In contrast to the SADS, which is semistructured and requires administration by trained professionals, the DIS (Robins, Helzer, Croughan, & Ratcliff, 1981) is highly structured and was designed specifically by the National Institute of Mental Health (Division of Biometry and Epidemiology) to be administered by nonprofessional interviewers for epidemiological studies (see Helzer & Robins, 1988). It has been updated for the *DSM-III-R* (Robins et al., 1989) and the *DSM-IV* (Robins, Cottler, Bucholz, & Compton, 1996) but has yet to be updated for the *DSM-5*. The latest version (DIS-IV) includes 19 modules with more than 30 diagnoses, including one

personality disorder diagnosis (antisocial personality). This modular format allows for tailoring various portions of the DIS-IV to the interests of the researcher or clinician. However, clinical judgment is reduced to a minimum by using verbatim wording, specific guidelines, a clear flow from one question to the next, and simple yes-no answers. Thus, the DIS is far more economical to administer than the SADS. Total administration time is 60 to 90 minutes. Studies have generally indicated that results are comparable between trained clinicians and nonprofessional interviewers (Helzer, Spitznagel, & McEvoy, 1987).

Adult Version

The original version of the DIS was derived from the format of the earlier Renard Diagnostic Interview. However, diagnosis for the DIS-IV is based exclusively on *DSM-IV* criteria. Initially, questions are directed toward obtaining information regarding the client's life, and information is also requested regarding more current symptoms based on the past 2 weeks, past month, past 6 months, and past year. Specific probe questions distinguish whether a symptom is clinically significant. A total of 470 potential clinical ratings are made and organized around 24 major categories. Administration time is approximately 60 to 90 minutes.

Computerized administration and scoring programs are available that can generate *DSM-IV*-based diagnoses. However, computer-based diagnoses on early versions of the DIS were found to generate an average of 5.5 possible diagnoses compared with an average of 2.6 for nonstructured interviews (Wyndowe, 1987). Patient acceptance for the computer administration has been found to be high, although the average administration time is somewhat longer than the clinician-interviewed version.

Studies of the reliability and validity of the DIS have been both variable and controversial. Although much of this research was done on pre-DIS-IV versions, the similarity of format and content between the DIS and DIS-IV suggests that much of this earlier research is pertinent. The comparability of diagnosis by professionals and nonprofessionals using the DIS has generally been supported. This finding suggests that nonprofessionals can effectively use the DIS to help gather data for large epidemiological studies. For example, Robins et al. (1981) found diagnostic agreement between psychiatrists and nonprofessional interviewers to be .69. The sensitivity (percentage of interviewees correctly identified) of the DIS varied according to type of diagnosis but had a mean of 75%, with a mean specificity (percentage of noncases correctly identified) of 94%. More recent studies have similarly concluded that the specificity is stronger than its sensitivity (Eaton, Neufeld, Chen, & Cai, 2000; J. M. Murphy, Monson, Laird, Sobol, & Leighton, 2000). However, data on sensitivity and specificity were based on using psychiatrists' diagnoses as the true index of diagnostic accuracy. The difficulties in considering psychiatrists' ratings as the truly accurate or "gold standard" criterion for validity have already been noted; therefore, it is probably best to consider the preceding data on sensitivity and specificity as forms of interrater agreement rather than concurrent validity. In contrast to this study, Vandiver and Sheer (1991) found somewhat marginal median test-retest reliabilities, ranging between .37 and .46.

Although many of the DIS ratings between professional and lay interviewers were equivalent, Helzer et al. (1985) found that, when compared with psychiatrists, nonprofessional interviewers tended to overdiagnose major depression. In contrast to Helzer et al. (1987), Folstein et al. (1985) did not find a sufficiently high rate of

agreement between diagnoses by a panel of psychiatrists and diagnoses by the DIS to warrant its use in epidemiological studies. Specifically, it was found that the DIS generated more cases of depression and schizophrenia and fewer cases of alcoholism and antisocial personality (Cooney, Kadden, & Litt, 1990; Folstein et al., 1985). Eaton et al. (2000) noted that false-negative diagnoses for many cases could be attributed mainly to failure by patients to report symptoms based on life crises or medical conditions. In contrast, the DIS has been found to be comparable with other commonly used psychiatric rating devices, such as the Psychiatric Diagnostic Interview (Folstein et al., 1985; R. Weller et al., 1985). However, both diagnostic strategies may contain inaccuracies, and it is difficult to tell in which areas these inaccuracies occurred (R. Weller et al., 1985). The DIS has had the greatest difficulty accurately diagnosing borderline conditions and patients in remission, but this is to be expected because these are the most problematic diagnoses for many other assessment strategies (Robins & Helzer, 1994). In contrast, Swartz et al. (1989) were able to find quite respectable sensitivities (85.7%) and specificities (86.2%) for borderline conditions using a DIS borderline index.

Child Version

The Diagnostic Interview Schedule for Children (DISC; Costello, Edelbrock, Duncan, & Kalas, 1984; Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000) is similar to the adult version in that it is highly structured and designed for nonprofessional interviewers. It differs in that it is designed to be given as both a child interview (DISC-C) and parent interview (DISC-P). There have also been versions designed for teachers (Teacher DISC), screening (DISC Predictive Scales), young adults (Young Adult DISC), and administrations that can be given by computer or audio recording (Lucas et al., 2001; Shaffer et al., 2000). Ratings are coded as 0 (not true), 1 (somewhat true), or 2 (very often true). *DSM-IV* (and now *DSM-5*) diagnoses are generated based on the combined ratings for the child and parent interviews. Some of the more problematic diagnoses (Autism, Pervasive Developmental Disorder, Pica) are based on an interview with the parent only. The entire interview takes an average of 70 minutes per informant and 90 to 120 minutes per patient, but an explicit skip structure can enable some interviews to be somewhat shorter. The most recent modification of the DISC (DISC-IV; Robins et al., 1996; Shaffer et al., 2000) was designed to be compatible with *DSM-IV* and *ICD–10* criteria. The DISC-IV comprises six modules, each of which represents the major diagnostic clusters (Anxiety, Mood, Disruptive, Substance Use, Schizophrenia, Miscellaneous).

DISC test-retest reliability (1-year interval) for *DSM-IV* diagnoses in a clinical sample was good to adequate with parent ratings having higher reliabilities (.54–.79) than child interviews (.25–.92; Shaffer et al., 2000). However, test-retest reliabilities for a community sample were generally quite poor for child interviews (.27–.64), although adequate for parent interviews (.45–.68; Shaffer et al., 2000). Children's reliability increased with age, which is expected considering their increase in intellectual abilities, greater memory, and improved language comprehension and expression. In contrast, reliabilities based on ratings from interviews with the parents decreased with the child's age, probably because the parents have progressively less contact with their child.

Research on the validity of the DISC has found that discriminations between psychiatric and pediatric groups were good for children with severe diagnoses and severe symptoms but not for children with mild to moderate difficulties (Shaffer et al., 2000).

Discriminations based on interviews with parents were generally more accurate than those based on children (Costello, Edelbrock, & Costello, 1985). Accuracy was also higher for externalizing than internalizing disorders (Friman et al., 2000). In addition, comparisons between psychiatric and pediatric referrals indicated that psychiatric referrals had more symptom scores and more psychiatric diagnoses than pediatric referrals (Costello et al., 1985). The DISC has also been found to identify risk factors for substance abuse (Greenbaum, Prange, Friedman, & Silver, 1991) and to predict behaviors related to conduct and oppositional disorders (Friman et al., 2000). Ratings between DISC and clinician-based diagnosis were moderate to good (.29–.74 for parent and .27–.79 for child; Shaffer et al., 2000) in research settings and followed strict diagnostic guidelines. However, there was very poor agreement between DISC and clinician-based diagnosis when the clinicians performed diagnosis in everyday clinical settings (A. L. Jensen & Weisz, 2002). This lack of agreement may reflect not so much a weakness of the DISC itself but more the fact that there are considerable differences between how diagnosis is achieved in research as opposed to practice contexts. In summary, the DISC has strengths in that it has good reliability and validity among clinical samples involving parent interviews, especially when the problems are related to externalizing disorders. However, the DISC is more problematic when ratings are based on child interviews, particularly among community samples and for internalizing disorders.

Diagnostic Interview for Children and Adolescents

The Renard Diagnostic Interview (Helzer et al., 1981) inspired both the DIS and the DICA (Herjanic & Campbell, 1977; Herjanic & Reich, 1982). The DICA has been through several revisions, which have incorporated the different editions of the *DSM* and elements of the DIS (W. Reich, 2000). Similar to the DIS, the DICA has been designed for administration by lay interviewers. The most recent version (DICA–IV) was published in 1997, aligns with the *DSM-IV*, and is available in child, adolescent, and parent versions (W. Reich, 2000). The DICA can be administered to children between ages 6 and 17 years. The format is semistructured and primarily organized around different themes, such as behavior at home, behavior at school, and interpersonal relationships with peers. Additional content areas are substance abuse and the presence of syndromes such as anxiety disorders, mania, and affective disorders. Elaborate instructions are given for skipping irrelevant items, and total administration time is between 1 and 2 hours. The administration begins with an interview of both the parent and child, which is designed to establish baseline behaviors and to obtain relevant chronological information. The parent is then questioned about the child to determine the possible appropriateness of common *DSM-IV* diagnostic categories. The final step is to administer the "Parent Questionnaire," which requests additional medical and developmental history and addresses possible diagnoses that have not been covered by previous questioning.

Reliability of the DICA has been quite variable. Test-retest reliability has been quite good, mostly ranging between .76 and .90 (Bartlett, Schleifer, Johnson, & Keller, 1991; Earls, Reich, Jung, & Cloninger, 1988). However, test-retest reliability for child (6 to 12) attention-deficit/hyperactivity disorder was low (.32) and oppositional disorder was

low to adequate (.46; W. Reich, 2000). Reliability has been found to be lowest for questions that were complex, related to time, and for children with the highest level of functional impairment. In contrast, questions with the highest reliability were related to frequency and to externalizing symptoms (Perez, Ascaso, Massons, & Chaparro, 1998). Most cross-informant (parent–child) agreement related to specific symptoms has been disappointingly low (.19 to .54; Herjanic & Reich, 1982). The highest level of agreement was for the oldest children and the lowest was for younger groups (W. Reich, 2000). Whereas parents reported more behavioral symptoms, children were more likely to report subjective complaints.

Validity studies on the DICA indicate that it can accurately make the somewhat gross distinction between middle- to older-age children who were referred to a general psychiatric clinic from those referred to a pediatric clinic (Herjanic & Campbell, 1977). However, there was considerable overlap for children between ages 6 and 8, thus suggesting that a greater possibility of misdiagnosis exists for children in this age range. The DICA was found to be most effective for assessing relationship problems, less effective for academic difficulties, and least effective for assessing school problems, somatic complaints, and neurotic symptoms (Herjanic & Campbell, 1977). In addition, adolescents diagnosed with depression on the DICA also had corresponding elevations on the Beck Depression Inventory (BDI; Martin, Churchard, Kutcher, & Korenblum, 1991). W. Reich (2000) reported that as the genetic similarity of persons diagnosed with Bipolar Disorder decreased, their level of psychopathology on the DISC correspondingly decreased. In summary, the psychometric properties of the DICA have been variable; more studies are needed to substantiate its validity, particularly concurrent validity (R. Rogers, 2001).

RECOMMENDED READING

Garb, H. N. (2007). Computer-administered interviews and rating scales. *Psychological Assessment*, *19*, 4–13.

Othmer, E., & Othmer, S. C. (2002). *The clinical interview using DSM-IV-TR: Vol. 1. Fundamentals*. Washington, DC: American Psychiatric Press.

Rogers, R. (2001). *Handbook of diagnostic and structured interviewing*. New York, NY: Guilford Press.

Sommers-Flanagan, R., & Sommers-Flanagan, J. (2013). *Clinical interviewing* (5th ed.). Hoboken, NJ: Wiley.

Summerfeldt, L. J., Kloosterman, P. H., & Antony, M. M. (2011). Structured and semistructured diagnostic interviews. In M. Antony & D. H. Barlow (Eds.), *Handbook of assessment and treatment planning for psychological disorders* (2nd ed., pp. 95–140). New York, NY: Guilford Press.