PSYCHOLOGICAL ASSESSMENT – II

UNIT I

The Projective Hypothesis

Frank (1939, 1948) introduced the term projective method to describe a category of tests for studying personality with unstructured stimuli. In a projective test the examinee encounters vague, ambiguous stimuli and responds with his or her own constructions. Disciples of projective testing are heavily vested in psychoanalytic theory and its postulation of unconscious aspects of personality. These examiners believe that unstructured, vague, ambiguous stimuli provide the ideal circumstance for revelations about inner aspects of personality. The central assumption of projective testing is that responses to the test represent projections from the innermost unconscious mental processes of the examinee. We introduce this topic with some preliminary concepts and distinctions relevant to projective testing. The assumption that personal interpretations of ambiguous stimuli must necessarily reflect the unconscious needs, motives, and conflicts of the examinee is known as the projective hypothesis.

The challenge of projective testing is to decipher underlying personality processes (needs, motives, and conflicts) based on the individualized, unique, subjective responses of each examinee. In the sections that follow we will examine how well projective tests have met this portentous assignment.

A Classification of Projective Techniques

Lindzey (1959) has offered a classification of projective techniques into five categories:

- Association to inkblots or words
- Construction of stories or sequences
- Completions of sentences or stories
- Arrangement/selection of pictures or verbal choices
- Expression with drawings or play

Association techniques include the widely used Rorschach inkblot test and its psychometrically superior cousin the Holtzman Inkblot Technique, as well as word association tests. Construction techniques include the Thematic Apperception Test and the many variations upon this early instrument. Completion techniques consist mainly of sentence completion tests, discussed later. Arrangement/selection procedures such as the Szondi test (discussed in the first chapter) are currently seldom used. Finally, expression techniques such as the Draw-A-Person or House-Tree-Person test are very popular among clinicians in spite of dubious validity data
Association Techniques

The Rorschach :

The Rorschach consists of 10 inkblots devised by Herman Rorschach (1884–1922) in the early 1900s. He formed the inkblots by dribbling ink on a sheet of paper and folding the paper in half, producing relatively symmetrical bilateral designs. Five of the inkblots are black or shades of gray, while five contain color; each is displayed on a white background. An inkblot of the type employed by Rorschach is shown in Figure 8.1. The Rorschach is suited to persons age 5 and up but is most commonly used with adults.

Regrettably, Rorschach died before he could complete his scoring methods, so the systematization of Rorschach scoring was left to his followers. Five American psychologists produced overlapping but independent approaches to the test—Samuel Beck, Marguerite Hertz, Bruno Klopfer, Zygmunt Piotrowski, and David Rapaport (Erdberg, 1985). Predictably, the nuances of scoring varied from one scoring method to another. Beginning in the 1990s, John Exner and his colleagues began to codify and synthesize the scoring approaches into a single method known as the Rorschach Comprehensive System (Exner, 1991, 1993; Exner & Weiner, 1994). The Comprehensive System (CS) supplanted all previous methods and became the preferred scoring system because it was more clearly grounded in empirical research. Even so, reservations about the Rorschach in general and the CS in particular persisted in the trade (Lilienfield, Wood, & Garb, 2000, 2001).

Beginning in about 2010, a new system for administration, scoring, and interpretation of the Rorschach emerged. The Rorschach Performance Assessment System (R-PAS) represents an extension and improvement of the CS In using the R-PAS, the examiner first establishes rapport and then sits to the side of the client or patient to minimize body language communication. For each card, the examiner asks the respondent to look at the stimulus and to answer "What might this be?" Before the test, the examiner asks for "two, maybe three responses" per card. During the test, if only one reply is given, the examiner prompts for additional response(s), and pulls the card after four responses are provided. This is called response optimization, which elicits a typical range of 18 to 28 responses. This technique greatly reduces short and long records (protocols with upwards of 100 responses have been encountered), which affords a better fit with norms.

The R-PAS incorporates several laudable improvements (www.r-pas.org): • Evidence-based selection of scoring variables • Detailed guidelines for test administration • Methods to optimize the number of responses • Guidelines for clarifying coding uncertainties • Normative reference values for international samples • Form quality tables for accuracy and conventionality • Inexpensive scoring with a web-based program • Easy-to-read graphs with standard scores • Translations into several languages

Once the test is administered and the responses recorded, scoring begins. This is an intricate process that requires significant training. We can only refer to highlights here. Responses are scored for a number of variables such as location, content, form quality, thought processes, and determinants. Determinants are different aspects of the blot such as color, shading, and form, which appear to have influenced examinee responses.

Interrater reliability of R-PAS scores is excellent. Using a diverse sample of 50 Rorschach records randomly selected from ongoing research, the median intraclass correlation coefficient (an index of agreement between raters) for 60 variables was .92.

Another useful feature of this new approach to Rorschach scoring is the availability of an international reference sample for standardization of scoring variables. This sample of 1,396 protocols was obtained from 15 nations, including Australia, Brazil, Japan, Israel, and Spain—just to give a sense of the global distribution. The validity of the Rorschach as scored with the R-PAS (or any other scoring system) is difficult to summarize in any simple manner. Individual studies indicate good validity for some purposes, but limited validity for other applications.

Frank (1990) has emphasized that formal scoring of the Rorschach is insufficient for some purposes such as the diagnosis of schizophrenia. He stresses that an analysis of the patient's thinking for the presence of highly personal, illogical, and bizarre associations to the blots is essential for psychodiagnosis. In his approach, the Rorschach is really an adjunct to the interview, and not a test per se.

The meaningful use of color in the response also contributes to a positive score, whereas using color to depict explosions or diseases results in points being subtracted. Several categories are scored, yielding a total score that ranges from –12 to +17. The following interpretations are then assigned to different ranges of the RPRS score:

17 to 13: The person is almost able to help himself. A very promising case that just needs a little help.

12 to 7: Not quite so capable as the previous case to work out his problems himself but with some help is likely to do pretty well.

6 to 2: Better than 50–50 chance; any treatment will be of some help.

1 to −2: 50–50 chance.

−3 to −6: A difficult case that may be helped somewhat but is generally a poor treatment prospect.

−7 to −12: A hopeless case.

Another useful scoring system for the Rorschach is the Thought Disorder Index (TDI), which assesses formal thought disorder (Holtzman, Levy, & Johnston, 2005). Thought disorder exists on a continuum from mild slippage to bizarre disorganization and is especially characteristic of patients with schizophrenia. Thus, the assessment of thought disorder is pivotal in the diagnosis and treatment of individuals with schizophrenia or other serious mental illness.

The TDI is calculated by scoring each response for the severity level of thought disorder from none to extreme, with possible scores of 0, .25, .50, .75, and 1.0. Then the average score is computed across all responses. This number is multiplied by 100 to yield the final score on a range from 0 to 100. Thus, an overall score of 0 would mean that not one response revealed any thought disorder, whereas a score of 100 would signify that, without exception, every response was highly bizarre and disorganized.

The reliability of the TDI is reasonably good, with split-half correlations around .80 and interrater reliability coefficients of .90 and higher. Validity has been supported from a number of directions, such as huge improvements in scores when patients with schizophrenia are tested before and after comprehensive interventions including drug therapies. Mastering the TDI scoring criteria is far easier than learning the Comprehensive System. One liability is that learning the scoring system is an arduous and time consuming task that requires dozens of hours of practice and years of supervised experience. A second problem is that administering and scoring the Rorschach requires a few hours of professional time from a licensed psychologist. This time is a precious and expensive commodity. Someone has to pay for it. These practical issues are daunting.

Completion Techniques

Sentence Completion Tests : In a sentence completion test, the respondent is presented with a series of stems consisting of the first few words of a sentence, and the task is to provide an ending. As with any projective technique, the examiner assumes that the completed sentences reflect the underlying motivations, attitudes, conflicts, and fears of the respondent. Usually, sentence completion tests can be interpreted in two different ways: subjective-intuitive analysis of the underlying motivations projected in the subject's responses, or objective analysis by means of scores assigned to each completed sentence.

Of course, most sentence completion tests are much longer—anywhere from 40 to 100 stems—and contain more themes— anywhere from 4 to 15 topics. Dozens of sentence completion tests have been developed; most are unpublished and unstandardized instruments produced to meet a specific clinical need.

Loevinger's Washington University Sentence Completion Test is the most sophisticated and theory-bound. However, the Rotter Incomplete Sentences Blank has the strongest empirical underpinnings and is the most widely used in clinical settings.

Rotter Incomplete Sentences Blank: The Rotter Incomplete Sentences Blank (RISB) consists of three similar forms—high school, college, and adult—each containing 40 sentence stems written mostly in the first person (Rotter & Rafferty, 1950).

Although the test can be subjectively interpreted in the usual manner through qualitative analysis of needs projected in the subject's responses, it is the objective and quantitative scoring of the RISB that has drawn the most attention. In the objective scoring system each completed sentence receives an adjustment score from 0 (good adjustment) to 6 (very poor adjustment). These scores are based initially on the categorizing of each response as follows:

Omission—no response or response too short to be meaningful • Conflict response—indicative of hostility or unhappiness • Positive response—indicative of positive or hopeful attitude • Neutral response—declarative statement with neither positive nor negative affect

Conflict responses are scored 4, 5, or 6, from lowest to highest degree of the conflict expressed. Positive responses are scored 2, 1, or 0, from least to most positive response. Neutral responses and omissions receive no score. The manual gives examples of each scoring category. The overall adjustment score is obtained by adding the weighted ratings in the conflict and positive categories. The adjustment score can vary from 0 to 240, with higher scores indicating greater maladjustment.

The reliability of the adjustment score is exceptionally good, even when derived by assistants with minimal psychological expertise. Typically, interscorer reliabilities are in the .90s and split-half coefficients are in the .80s. The validity of this index has been investigated in numerous studies using the RISB as a screening device with a "maladjustment" cutoff score. These and similar findings support the construct validity of the adjustment index but also indicate that classification rates are much lower than needed for individual decision making or effective screening. It also appears that the norms for the adjustment index are outdated.

The simplicity of the single adjustment score is both the test's strength and weakness. True, the test provides a quick and efficient method for obtaining an overall index of how respondents are functioning on a day to-day basis. However, a single score cannot possibly

capture any nuances of personality functioning. In addition, the RISB is subject to the same types of bias as other self-report measures, namely, the information will reflect mainly what the respondent wants the examiner to know.

Construction Techniques

The Thematic Apperception Test (TAT): The TAT consists of 30 pictures that portray a variety of subject matters and themes in black-and white drawings and photographs; one card is blank. Most of the cards depict one or more persons engaged in ambiguous activities. Some cards are used for adult males (M), adult females (F), boys (B), or girls (G), or some combination (e.g., BM). As a consequence, exactly 20 cards are appropriate for every examinee.

In administering the TAT, the examiner requests the examinee to make up a dramatic story for each picture, telling what led up to the current scene, what is happening at the moment, how the characters are thinking and feeling, and what the outcome will be. The examiner writes down the story verbatim for later scoring and analysis.

The TAT was developed by Henry Murray and his colleagues at the Harvard Psychological Clinic (Morgan & Murray, 1935; Murray, 1938). The test was originally designed to assess constructs such as needs and press, elements central to Murray's personality theory. According to Murray, needs organize perception, thought, and action and energize behavior in the direction of their satisfaction. Examples of needs include the needs for achievement, affiliation, and dominance. In contrast, press refers to the power of environmental events to influence a person. Alpha press is objective or "real" external forces, whereas beta press concerns the subjective or perceived components of external forces. Murray (1938, 1943) developed an elaborate TAT scoring system for measuring 36 different needs and various aspects of press, as revealed by the examinee's stories.

Almost as soon as Murray released the TAT, other clinicians began to develop alternative scoring systems (e.g.,Dana, 1959; Tomkins, 1947). By the 1950s, there was no single preferred mode of administration, no single preferred system of scoring, and no single preferred method of interpretation. Clinicians even vary the wording of the instructions and commonly select an individualized subset of TAT cards for each client. Indeed, the absence of standardized procedures is such that we should rightly regard the TAT as a method, not a test.

Currently, clinicians downplay the emphasis on imagination and intelligence when giving instructions. Surely, this omission must influence the quality of the stories produced. Even though more than a dozen scoring systems have been proposed, interpretation of the TAT is usually based on a clinical-qualitative analysis of the story productions. A central consideration harks back to Murray's "hero" assumption. According to this viewpoint, the hero is the protagonist of the examinee's story. It is assumed that the examinee clearly identifies with this character and projects his or her own needs, strivings, and feelings onto the hero. Conversely, thoughts, feelings, or actions avoided by the hero may represent areas of conflict for the examinee.

The psychometric adequacy of the TAT is difficult to evaluate because of the abundance of scoring and interpretation methods. Clinicians defend the test on an anecdotal basis, pointing out remarkable and confirmatory findings such as illustrated here. However,

data-minded researchers are more cautious. One problem is that formally scored TAT protocols possess very low test–retest reliability, with a reported median value of r = .28.

Furthermore, an astonishing 97 percent of test users employ subjective and "personalized" procedures for interpreting the TAT; that is, only a tiny fraction of clinical practitioners rely on a standardized scoring system. This is troubling because a consistent theme in research on projective testing is that intuitive interpretations are likely to over-diagnose psychological disturbance.

In addition to clinical applications, the TAT has received considerable use for research purposes. For example, Turk, Brown, Symington, and Paul (2010) examined the content of TAT stories from 22 persons with agenesis of the corpus callosum (ACC), a congenital brain disorder in which the pathways connecting the two cerebral hemispheres are partially or completely absent. They used the linguistic inquiry software of James Pennebaker (Tauszcik & Pennebaker, 2010) to count words in psychologically meaningful categories. Compared to age- and IQ-matched controls, the ACC individuals used fewer words pertaining to emotionality, cognitive processes, and social processes, indicating that they experienced greater difficulty imagining and inferring the mental and emotional states of others. In this research application, the TAT proved helpful for enhancing our understanding of the unique qualities of persons with ACC.

The Picture Projective Test : The Picture Projective Test (PPT) is an attempt to construct a general-purpose instrument with improved psychometric qualities (Ritzler, Sharkey, & Chudy, 1980; Sharkey & Ritzler, 1985).

The developers of the PPT note that the majority of the TAT pictures exert a strong negative stimulus "pull" on storytelling. The TAT cards are cast in dark, shaded tones and most scenes portray persons in low-key or gloomy situations. It is not surprising, then, that projective responses to the TAT are strongly channeled toward negative, melancholic stories (Goldfried & Zax, 1965). In contrast, the PPT uses a set of pictures taken from the Family of Man photo essay published by the Museum of Modern Art (1955). The following criteria were used in selecting 30 pictures:

The pictures had to show promise of eliciting meaningful projective material. • Most but not all of the pictures had to include more than one human character, About half of the pictures had to depict humans showing positive affective expression (e.g., smiling, embracing, dancing). • About half of the pictures had to depict humans in active poses, not simply standing, sitting, or lying down.

Compared to the TAT productions, the PPT stories were of comparable length but were much more positive in thematic content and emotional tone. The PPT stories were also much more active, meaning that the central character had an active, self-determined effect on the situation in the story. Furthermore, the PPT stories placed greater emphasis on interpersonal rather than intrapersonal themes. In other words, the PPT stories placed more emphasis on "healthy," adaptive aspects of personality adjustment than did the TAT productions.

The PPT developers also compared their instrument against the TAT in a diagnostic validity study (Sharkey & Ritzler, 1985). PPT and TAT story productions of 50 subjects were compared: normals, nonhospitalized depressives, hospitalized depressives, hospitalized psychotics with good premorbid histories, and hospitalized psychotics with poor premorbid

histories (10 subjects in each group). Although the TAT and PPT were essentially equal in their capacity to discriminate normal from depressed subjects, the PPT was superior in differentiating psychotics from normals and depressives. On the PPT, depressives told stories with gloomier emotional tone and psychotics made more perceptual distortions, and thematic/interpretive deviations. The PPT appears to be a very promising instrument, although it is obvious that further research is needed on its psychometric qualities. One noteworthy feature is that anyone can purchase the PPT stimuli at their local bookstore. The requisite materials are found in the Family of Man photo collection (Museum of Modern Art, 1955).

Children's Apperception Test : Designed as a direct extension of the TAT, the Children's Apperception Test (CAT) consists of 10 pictures and is suitable for children 3 to 10 years of age. The preferred version for younger children (CAT-A) depicts animals in unmistakably human social settings (Bellak & Bellak, 1991). The test developers used animal drawings on the assumption that young children would identify better with animals than humans. A human figure version (CAT-H) is available for older children (Bellak & Bellak, 1994). No formal scoring system exists for the CAT and no statistical information is provided on reliability or validity. Instead, the examiner prepares a diagnosis or personality description based on a synthesis of 10 variables recorded for each story: (1) main theme; (2) main hero; (3) main needs and drives of hero; (4) conception of environment (or world); (5) perception of parental, contemporary, and junior figures; (6) conflicts; (7) anxieties; (8) defenses; (9) adequacy of superego; (10) integration of ego (including originality of story and nature of outcome) (Bellak, 1992). The lack of attention to psychometric issues of scoring, reliability, and validity of the CAT is troublesome to most testing specialists.

Other Variations on the TAT:

In addition, modifications and variations of the TAT have been developed for ethnic, racial, and linguistic minorities. One of the first was the Thompson TAT (T-TAT) in which 21 of the original TAT pictures were redrawn with African American figures (Thompson, 1949). This TAT modification incorporated certain unintended changes—for example, in facial expressions and the situations portrayed. As a result, the T-TAT should be considered a new test and not just a TAT translation suited to African American individuals (Aiken, 1989).

Another specialized TAT-like test is the TEMAS, which consists of 23 colorful drawings that depict Hispanic persons interacting in contemporary, inner-city settings (Aiken, 1989; Constantino, Malgady, & Rogler, 1988). TEMAS is Spanish for themes and an acronym for "tell me a story." The thematic content of TEMAS stories is scored for 18 cognitive functions, 9 personality (ego) functions, and 7 affective functions. The test can also be scored for various objective indices such as reaction time, fluency, unanswered inquiries, and stimulus transformations (e.g., a letter is transformed into a bomb). Hispanic children respond well to the TEMAS, even though they may be inarticulate in response to traditional projective tests.

The inconsistent reliability of the TEMAS is a source of concern, because reliability constrains validity. The manual reports that Cronbach's alpha for the 34 scoring functions ranged from .31 to .98 with half below .70. Test–retest reliabilities were even lower; the highest correlation was r = .53 and for 26 of the 34 functions the correlations were near zero!

Family Apperception Test : For children ages 6 and older, the Family Apperception Test consists of 21 cards depicting a family in various situations. For example, one card shows a

family sitting around a table with parents talking while the children eat. As with the TAT, the examinee is asked to describe what led up to the scene, what is happening now, what will happen next, and what the main characters are feeling. The test is based on family systems theory. The manual provides a scoring guide for categories such as limit-setting, conflict resolution, boundaries, quality of relationships, and emotional tone (Sotile, Julian, Henry, & Sotile, 1988).

Blacky Pictures : For children ages 5 and older, the Blacky Pictures test was also based on the premise that children identify more readily with animals than humans. The 11 cartoon stimuli depict the adventures of the dog Blacky and his family (Mama, Papa, and sibling Tippy). In addition to requesting a story for each card, the examiner also presents multiple-choice questions based on stages of psychosexual development derived from psychoanalytic theory (Blum, 1950). Although the test was originally developed with adults, children enjoy taking the Blacky and are quite responsive to the pictures. Problems with this test include the absence of norms, especially for children, and poor stability of scores (LaVoie, 1987).

Michigan Picture Test-Revised : For older children ages 8 to 14 years, the MPT-R consists of 15 pictures and a blank card. Responses are scored for Tension Index (e.g., portrayal of personal adequacy), Direction of Force (whether the central figure acts or is acted upon), and Verb Tense (e.g., past, present, future). These three scores can be combined to yield a Maladjustment Index. Reliability and norms are adequate, although evidence of validity is unsatisfactory. A major problem with this test is that the cards portray interpersonal relationships so vividly that little is left to the child's imagination (Aiken, 1989).

Senior Apperception Test (SAT) : Although the 16 situations depicted on the SAT cards include some positive circumstances, the majority of pictures were designed to reflect themes of helplessness, abandonment, disability, family problems, loneliness, dependence, and low self-esteem (Bellak, 1992). Critics complain that the SAT stereotypes the elderly and therefore discourages active responding (Schaie, 1978).

Expression Techniques

The Draw-A-Person Test: Goodenough (1926) used the Draw-A-Man task as a basis for estimating intelligence. Subsequently, psychodynamically minded psychologists adapted the procedure to the projective assessment of personality. Karen Machover (1949, 1951) was the pioneer in this new field. Her procedure became known as the Draw-A-Person Test (DAP). Her test enjoyed early popularity and is still widely used as a clinical assessment tool. Watkins, Campbell, Nieberding, and Hallmark (1995) report that projective drawings such as the DAP rank eighth in popularity among clinicians in the United States.

The DAP is administered by presenting the examinee with a blank sheet of paper and a pencil with eraser, then asking the examinee to "draw a person." When the drawing is completed the examinee usually is directed to draw another person of the sex opposite that of the first figure. Finally, the examinee is asked to "make up a story about this person as if he [or she] were a character in a novel or a play" (Machover, 1949). Interpretation of the DAP proceeds in an entirely clinical-intuitive manner, guided by a number of tentative psychodynamically based hypotheses.

Machover maintained that examinees were likely to project acceptable impulses onto the same-sex figure and unacceptable impulses onto the opposite-sex figure. She also believed

that the relative sizes of the male and female figures revealed clues about the sexual identification of the examinee. These interpretive premises are colorful, interesting, and plausible. However, they are based entirely on psychodynamic theory and anecdotal observations. Machover made little effort to validate the interpretations. The empirical support for her hypotheses is somewhere between meager and nonexistent. In favor of the DAP, the overall quality of drawings does weakly predict psychological adjustment.

Rather than using the DAP to infer nuances of personality, a more appropriate application of this test is in the screening of children suspected of behavior disorder and emotional disturbance. For this purpose, Naglieri, McNeish, and Bardos (1991) developed the Draw A Person: Screening Procedure for Emotional Disturbance (DAP:SPED). In one study, diagnostic accuracy of problem children was significantly improved by application of the DAP:SPED scoring approach (Naglieri & Pfeiffer, 1992).

The House-Tree-Person Test (H-T-P):

The H-T-P is a projective test that uses freehand drawings of a house, tree, and person (Buck, 1948, 1981). The examinee is given almost complete freedom in sketching the three objects; separate pencil and crayon drawings are requested. Although the examiner can improvise an H-T-P Test with mere blank pieces of paper, Buck (1981) recommends the use of a four-page drawing form with identification information on the first page. Pages two, three, and four are titled House, Tree, and Person. Two drawing forms are needed for each examinee, one for pencil drawings and the other for crayon drawings. Buck (1981) also provides a separate four-page form for a postdrawing interrogation phase, which consists of 60 questions designed to elicit the examinee's opinions about elements of the drawings. Many practitioners feel the postdrawing interrogation phase is not worth the extended effort. Also, the value of separate crayon drawings is questioned.

The House-Tree-Person Test has much the same familial lineage as the Draw-A-Person Test. Like the DAP Test, the H-T-P Test was originally conceived as a measure of intelligence, complete with a quantitative scoring system to appraise an approximate level of ability (Buck, 1948). However, clinicians soon abandoned the use of the H-T-P as a measure of intelligence, and it is now used almost exclusively as a projective measure of personality.

The interpretation of the H-T-P rests on three general assumptions: the House drawing mirrors the examinee's home life and intrafamilial relationships; the Tree drawing reflects the manner in which the examinee experiences the environment; and the Person drawing echoes the examinee's interpersonal relationships. Buck (1981) provides numerous interpretive hypotheses for both quantitative and qualitative aspects of the three drawings. The H-T-P is an alluring test that has fascinated clinicians for more than 40 years. Unfortunately, Buck (1948, 1981) has never provided any evidence to support the reliability or validity of this instrument. In general, attempts to validate the H-T-P as a personality measure have failed miserably.

The popularity of the H-T-P has dropped off in recent years. A search of PsychINFO revealed only nine articles on the test since 2000, including four dissertations. Many clinicians do not use projective methods as tests at all but as auxiliary approaches to the clinical interview. These practitioners use projective techniques as clinical tools to derive tentative hypotheses about the examinee.

Self-Report and Behavioral Assessment of Psychopathology

There are many methods for the assessment of personality and related qualities. Broadly speaking two approaches have dominated the field: unstructured and structured. Unstructured methods such as the Rorschach, TAT, and sentence completion blanks permit broad latitude in the responses of the examinee. These approaches dominated personality testing in the early twentieth century but then slowly faded in standing. In contrast, structured approaches such as self-report inventories and behavior rating scales gained prominence in the mid-twentieth century and have continued to expand in popularity to the present time.

The more structured, objective methods for personality assessment favored by measurement minded psychologists include a wide variety of true–false, rating scale, and forced-choice instruments for assessing personality and other qualities.

The self-report approaches to testing are steeped in the details of psychometric methodology. These tests feature prominent references to reliability indices, criterion keying, factor analysis, construct validation, and other forms of technical craftsmanship. For this reason, the approaches are considered objective—as contrasted with projective. However, whether they are objective in any meaningful sense is really an empirical question that must be answered on the basis of research. Perhaps it is more accurate to call these methods structured. They are structured in the sense that highly specific rules are followed in the administration, scoring, interpretation, and narrative reporting of results. In fact, some of the approaches are so completely structured that an examinee can answer questions presented on a computer screen and observe a computer-generated narrative report spewed forth from the printer, literally seconds later.

Contemporary psychometricians have relied mainly upon three tactics for personality test development: theory-bounded approaches, factor-analytic approaches, and criterion-key methods.

Theory-Guided Inventories

The construction of several self-report inventories was guided closely by formal or informal theories of personality. Theory-guided inventories stand in contrast to factor-analytic approaches that often produce a retrospective theory based upon initial test findings. Theory-guided inventories also differ from the stark atheoretical empiricism found in criterion-key instruments such as the MMPI and MMPI-2. An example of a theory-guided inventory is the Personality Research Form (PRF), based on Murray's (1938) need-press theory of personality. Some theory-guided inventories such as the State-Trait Anxiety Inventory (STAI) attempt to measure very specific components of personality.

Personality Research Form:

The Personality Research Form (Jackson, 1999) is a true–false inventory based loosely on Murray's (1938) theory of manifest needs. The reader will recall from an earlier discussion that Murray posited 15 needs and developed a projective test, the Thematic Apperception Test, to tap those needs. Based on factor-analytic approaches, Jackson expanded the number of needs and produced several forms for assessment. The forms differ in the number of scales and number of items per scale. In addition to parallel short tests (forms A and B), the Personality Research Form (PRF) also exists as parallel long forms (forms AA and BB). These forms, used

primarily with college students, consist of 440 true–false items. The long forms yield 20 personality-scale scores and two validity scores, Infrequency and Desirability (Table 8.5). The most popular version of the PRF is form E, which consists of all 22 scales in a modified 352-item test.

In constructing the PRF form E, Jackson first formulated rigorous and theoretically based definitions of the traits to be measured, following Murray's (1938) system for personality description. Next, for each scale over 100 items were written to tape the traits underlying the hypothesized needs. After editorial review, these items were administered to large samples of college students. Item selection was based on simplicity of wording, high biserial correlations with total scale scores, low correlations with other scales (maximizing scale independence), and low correlations with the Desirability scale (minimizing social desirability bias). Convergent and discriminant validity was considered throughout. For the original long forms AA and BB, 20 items were selected for each scale, resulting in 20 × 22 or 440 items. For the PRF form E, about four items were dropped from each scale, yielding a 352-item test.

The rigorous scale construction procedures employed by Jackson (1970) yielded scales with good internal consistency, with a median coefficient alpha of .70. Test–retest reliabilities are exceptionally strong, ranging from .80 to .96 for a two-week interval, with a median of .91 (Jackson, 1999). Norms are based on thousands of college students from North America, and also include subgroup norms for psychiatric inpatients and criminal offenders. A desirable feature of the PRF is its readability: The test requires only a fifth- or sixth-grade reading level.

The validity of the PRF rests upon a substantial body of research over many decades. A lengthy bibliography citing more than 300 articles about the test can be found at www.sigmaassessmentsystems.com. Correlations between self and roommate ratings on the PRF constructs are reported to range from .27 to .74, with a median of .53. The construct validity of the PRF rests especially upon confirmatory factor analyses corroborating the grouping of the items into 20 scales (Jackson, 1970, 1984b). In addition, research indicates positive correlations with comparable scales on other inventories. Edwards and Abbott (1973) found exceptionally strong and confirmatory correlations between similar scales on the PRF and the Edwards Personality Inventory. The PRF outperformed the more widely used Sixteen Personality Factor Questionnaire (16PF, discussed later in this section) in predicting the job performance of 487 candidates for managerial positions.

State-Trait Anxiety Inventory

The State-Trait Anxiety Inventory (STAI) is a popular self-report measure of anxiety, used in research and clinical settings (Spielberger, 1983, 1989). The current version is called Form Y, a minor revision of the original Form X (Spielberger, Gorsuch, & Lushene, 1970). A similar scale for children also is available (Spielberger, 1973). The test has been translated into more than 40 languages. We limit our discussion here to the adult version.

The purpose of the STAI is to differentiate between the temporary condition of state anxiety and the more long-standing quality of trait anxiety. State anxiety is defined as a "transitory emotional state or condition characterized by subjective feelings of tension and apprehension, and by activation of the autonomic nervous system." Trait anxiety refers to "relatively stable individual differences in anxiety proneness".

The state scale (A-State scale) consists of 20 items that evaluate how the respondent feels "right now, at this moment." Items are similar to I feel at peace and I am distressed. Responses are on a 4-point scale (Not At All, Somewhat, Moderately So, and Very Much So). The trait scale (A-Trait scale) consists of 20 items that assess how the respondent feels "generally." Items are similar to I am a stable person and I lack confidence. Reponses are on a 4-point scale (Almost Never, Sometimes, Often, and Almost Always). Of course, scoring is reversed for positively stated items. The range of scores for each scale is 20 to 80, with higher scores indicating greater anxiety. Extensive normative data are available, stratified by age and subdivided by setting (employed adults, college students, high school students, military recruits). The STAI has received extensive service in research, and also is used in health-related clinical applications such as gauging anxiety in pregnant women, monitoring improvement in psychotherapy patients.

State anxiety fluctuates in response to environmental circumstances and may change even from hour to hour. Therefore, we can expect that test– retest reliability will be lower for state anxiety than for trait anxiety. This is precisely what researchers find, with short-range reliability in the .40s and .50s for the A-State scale and in the high .80s for the A-Trait scale. Internal consistency of the scale is excellent, with Cronbach's alpha of .86 for the total score in a sample of medical patients. Individual alpha values for A-State and A-Trait are robust as well, with results of .95 and .93, respectively, in a sample of 567 patients treated at an anxiety disorders clinic.

The validity of the STAI is well established from dozens of studies demonstrating content validity, convergent/discriminant validity, and construct validity. In sum, the STAI is a brief, reliable, and valid measure of state and trait anxiety. The measure is a mainstay for clinicians and researchers.

Factor-Analytically Derived Inventories

Eysenck Personality Questionnaire

The Eysenck Personality Questionnaire (EPQ) was designed to measure the major dimensions of normal and abnormal personality (Eysenck & Eysenck, 1975). Based on a lifelong program of factor-analytic questionnaire research and laboratory experiments on learning and conditioning, Eysenck isolated three major dimensions of personality: Psychoticism (P), Extraversion (E), and Neuroticism (N). The EPQ consists of scales to measure these dimensions and also incorporates a Lie (L) scale to assess the validity of an examinee's responses. The EPQ contains 90 statements answered "yes" or "no" and is designed for persons aged 16 and older. A Junior EPQ containing 81 statements is suitable for children ages 7 to 15.

High scores on the P scale indicate aggressive and hostile traits, impulsivity, a preference for liking odd or unusual things, and empathy defects. Antisocial and schizoid patients often obtain high scores on this dimension. In contrast, low scores on P foretell more desirable characteristics such as empathy and interpersonal sensitivity.

High scores on the E scale indicate a loud, gregarious, outgoing, fun-loving person. Low scores on the E scale indicate introverted traits such as a preference for solitude and quiet activities.

The N scale reflects a dimension of emotionality that ranges from nervous, maladjusted, and overemotional (high scores) to stable and confident (low scores).

The reliability of the EPQ is excellent. For example, the one-month test–retest correlations were .78 (P), .89 (E), .86 (N), and .84 (L). Internal consistencies were in the .70s for P and the .80s for the other three scales. The construct validity of the EPQ is also well established through dozens of studies using behavioral, emotional, learning, attentional, and therapeutic criteria.

A major focus of research with the EPQ has been on the empirical correlates of extraversion and its polar opposite, introversion. In general, the technical characteristics of the EPQ are very strong, certainly stronger than found in most self-report inventories. The practical utility of the instrument is supported by voluminous research literature.

Comrey Personality Scales

For practitioners who desire a short self-report inventory suitable for college students and other adults, the Comrey Personality Scales (Comrey, 1970, 1980, 2008) would be a good choice. As a protégé of Guilford, Comrey pursued a factor-analytic strategy in developing his 180-item test. Comrey relied exclusively upon college students in the development and standardization of his test, so the CPS is well suited to assessment of personality in this subpopulation.

A special virtue of the CPS is its brevity. Consisting of 180 statements, the test is only onethird as long as competing instruments such as the MMPI-2. The eight CPS personality scales consist of 20 items each, divided equally between positively and negatively worded statements. Another 20 items are devoted to a validity check and the assessment of social desirability response bias.

(V) Validity Check. A score of 8 is the expected raw score. Any score on the V scale that gives a T-score equivalent below 70 is still within the normal range, however. Higher scores are suggestive of an invalid record.

(R) Response Bias. High scores indicate a tendency to answer questions in a socially desirable way, making the respondent look like a “nice” person.

(T) Trust versus Defensiveness. High scores indicate a belief in the basic honesty, trustworthiness, and good intentions of other people

(O) Orderliness versus Lack of Compulsion. High scores are characteristic of careful, meticulous, orderly, and highly organized individuals.

(C) Social Conformity versus Rebelliousness. Individuals with high scores accept society as it is, resent nonconformity in others, seek the approval of society, and respect the law.

(A) Activity versus Lack of Energy. High-scoring individuals have a great deal of energy and endurance, work hard, and strive to excel.

(S) Emotional Stability versus Neuroticism. High-scoring persons are free from depression, optimistic, relaxed, stable in mood, and confident.

(E) Extraversion versus Introversion. High scoring individuals meet people easily, seek new friends, feel comfortable with strangers, and do not suffer from stage fright.

(M) Mental Toughness versus Sensitivity. High-scoring individuals tend to be rather tough-minded people who are not bothered by blood, crawling creatures, vulgarity, and who do not cry easily or show much interest in love stories.

(P) Empathy versus Egocentrism. High-scoring individuals describe themselves as helpful, generous, sympathetic people who are interested in devoting their lives to the service of others.

Reflecting its careful factor-analytic derivation, the CPS scales possess exceptional internal consistencies, which range from .91 to .96. Cross-cultural studies with the CPS are highly supportive of its validity. Brief and Comrey (1993) report that the eightfactor solution to CPS item responses is found in factor analyses with Russian, U.S., Brazilian, Israeli, Italian, and New Zealand samples. Extreme scores on the CPS scales are strongly associated with psychological disturbance. This is particularly true for low scores on Trust versus Defensiveness, Activity versus Lack of Energy, Emotional Stability versus Neuroticism, Extraversion versus Introversion, and high scores on Orderliness versus Lack of Compulsion. The test is a reasonable predictor of clinical performance and personal suitability.

Criterion-Keyed Inventories

In a criterion-keyed approach, test items are assigned to a particular scale if, and only if, they discriminate between a well-defined criterion group and a relevant control group. For example, in devising a self-report scale for depression, items endorsed by depressed persons significantly more (or less) frequently than by normal controls would be assigned to the depression scale, keyed in the appropriate direction.

The test developer does not consult any theory of schizophrenia, depression, or anxiety reaction to determine which items belong on the respective scales. The essence of the criterion-keyed procedure is, so to speak, to let the items fall where they may.

Minnesota Multiphasic Personality Inventory-2 (MMPI-2)

First published in 1943, the MMPI was a 566-item true–false personality inventory designed originally as an aid in psychiatric diagnosis (Hathaway & McKinley, 1940, 1943). The test authors followed a strict empirical keying approach in the construction of the MMPI scales. The clinical scales were developed by contrasting item responses of carefully defined psychiatric patient groups (average N of about 50) with item responses of 724 control subjects. The result was a remarkable test useful both in psychiatric assessment and the description of normal personality. Within a few years, the MMPI became the most widely used personality test in the United States.

At first the MMPI aged gracefully; what appeared to be minor flaws were tolerated by practitioners. But as the MMPI reached middle age, the need for rejuvenation became increasingly obvious. The most serious problem was the original control group, which consisted primarily of relatives and visitors of medical patients at the University of Minnesota Hospital. The narrow choice of control subjects, tested mainly in the 1930s, proved to be a persistent source of criticism for the MMPI. All of the control subjects were white, and most

were young (average age about 35), married, and from a small town or rural area. This was a sample of convenience that was significantly unrepresentative of the population at large.

Several items used archaic and obsolete terminology, referring to "drop the handkerchief" (a parlor game from the 1930s), sleeping powders (sleeping pills), and streetcars (electric-powered buses). Other items used sexist language. Examinees found some items objectionable, especially those dealing with Christian religious beliefs. These items were the source of occasional lawsuits alleging invasion of privacy. Finally, a few items dealing with bowel functions and sexual behavior were just downright offensive.

From the standpoint of measurement, a more serious problem with item content was that of omission. The MMPI item pool was not broad enough to assess many important characteristics, including suicidal tendencies, drug abuse, and treatment related behaviors. An additional motive for MMPI revision was to extend the range of item coverage.

The MMPI-2 was released in 1989 after nearly a decade of revision and restandardization. The new, improved MMPI-2 incorporates a contemporary normative sample of 2,600 individuals who are loosely representative of the general population on major demographic variables (geographic location, race, age, occupational level, and income). Although higher educational levels are overrepresented, the MMPI-2 normative sample is still a vast improvement over the MMPI normative sample. The item pool has been significantly improved by revision of obsolete items, deletion of offensive items, and addition of new items to extend content coverage. The test developers retained the same titles and measurement objectives for the traditional validity and clinical scales.

In fact, when large samples of subjects complete the MMPI and the MMPI-2, scores on the individual validity and clinical scales typically correlate near .99. The MMPI-2 consists of 567 items carefully designed to assess a wide range of concerns. The examinee is asked to mark "true" or "false" for each statement as it applies to himself or herself. Most of the items are self-referential. The items encompass a wide variety of mainly pathological themes.

The MMPI requires a sixth-grade reading level and is completed by most persons in 1 to 1½ hours. The original MMPI scales were developed by contrasting item responses of carefully defined psychiatric patient groups (average N of about 50) with item responses of about 700 controls. The psychiatric patient groups included the following diagnostic categories: hypochondriasis, depression, hysteria, psychopathy, male homosexuality, paranoia, psychasthenia,5 schizophrenia, and the early phase of mania (hypomania). In addition, samples of socially introverted and socially extraverted college students were used to construct a scale for social introversion. The MMPI-2 retains the basic clinical scales with only minor item deletions and revisions.

The MMPI-2 can be scored for four validity scales, 10 standard clinical scales, and dozens of supplementary scales. In practice, clinicians place the greatest emphasis upon the validity and standard clinical scales. The supplementary scales are just that—supplementary. They provide information helpful in fine-tuning the interpretation of the traditional validity and clinical scales. MMPI-2 scale raw scores are converted to T scores, with a mean of 50 and a standard deviation of 10. Scores that exceed T of 65 merit special consideration.

The four validity scales are the Cannot Say (or ?), L, F, and K SCALES.

The Cannot Say score is simply the total number of items omitted or double-marked in completion of the answer sheet. The instructions for the test encourage examinees to mark all items, but omissions or double-marked items will occur. However, this is rare—the modal number of items omitted is zero. Omission of up to 10 items appears to have little effect on the overall test results—one of the benefits of having a huge pool of statements in the MMPI-2. A very high score on this scale may indicate a reading problem, opposition to authority, defensiveness, or indecisiveness caused by depression.

The L Scale is composed of 15 items all scored in the false direction. By answering "false" to L Scale items, the examinee asserts that he or she possesses a degree of personal virtue that is rarely observed in our culture (e.g., never gets angry, likes everyone, never lies, reads every newspaper editorial, and would rather lose than win). The L Scale was designed to identify a general, deliberate, evasive test-taking attitude. A high score on the L Scale indicates that the examinee is not only defensive, but naively so. Persons with any degree of psychological sophistication can adopt a defensive test-taking attitude and still score in the normal range on the L Scale..

The F Scale consists of 60 items answered by normal subjects in the scored direction no more than 10 percent of the time. These items reflect a broad spectrum of serious maladjustment, including peculiar thoughts, apathy, and social alienation. Even though F Scale items seem to indicate psychiatric pathology, they are seldom endorsed by patients. Fewer than 50 percent of these items appear on the clinical scales. Many persons with significant psychiatric disturbance do produce elevated scores in the range of T =70 or 80 on the F Scale. On the other hand, exceptionally high scores suggest additional hypotheses: insufficient reading ability, random or uncooperative responding, a motivated attempt to "fake bad" on the test, or an exaggerated "cry for help" in a distressed client.

The K Scale was designed to help detect a subtle form of defensiveness. The 30-item scale is composed, in part, of 22 items that differentiated normal profiles produced by defensive hospitalized psychiatric patients from those produced by normal controls. Additionally, eight items that improved discrimination of depressive and schizophrenic symptoms were added (McKinley, Hathaway & Meehl, 1948). An elevated score on the K Scale may indicate a defensive test-taking attitude. Normal range elevations on the K Scale suggest good ego strength—the presence of useful psychological defenses that allow the person to function well in spite of internal conflict.

In addition to the validity scales, the MMPI-2 is always scored for 10 clinical scales. With the exception of Social Introversion, these clinical scales were constructed in the usual criterion-keyed manner by contrasting responses of clinical subjects and normal controls. Social Introversion was developed by contrasting the responses of college students high and low in social introversion.

Hypochondriasis, depression, Hysteria, Psychopathic Deviate, Masculinity-Feminity, Paranoia, Psychasthenia, schizophrenia, Hypomania and social Introversion are the 10 clinical scales.

Dozens of supplementary scales can also be scored on the MMPI-2. Some of the supplementary scales are based upon rational identification of symptom clusters and subsequent scale

purification by empirical means. Fifteen useful MMPI-2 Content Scales were developed in this manner.

Many of the supplementary scales were developed by independent investigators; these scales vary widely in quality. In practice, only about 30 of the additional scales are routinely scored. Examples of the supplementary scales include Anxiety, Repression, Ego Strength, and the MacAndrew Alcoholism Scale-Revised. Anxiety (A) and Repression (R) are the first two major factors that always emerge from factor analysis of MMPI-2 responses. An interesting supplementary scale is Barron's (1953) Ego Strength (Es) Scale, which purports to predict positive response to psychotherapy.

The MacAndrew Alcoholism Scale-Revised (MAC-R; MacAndrew, 1965) is a useful index of alcohol or other substance abuse. The MAC-R is not only useful in assessment of alcoholism but is also helpful in the identification of heavy drinkers and drug-dependent individuals.

MMPI-2 Interpretation:

The interpretation of an MMPI-2 profile can proceed along two different paths: scale by scale or configural. In the simplest possible approach, scale by scale, the examiner determines the validity of the test, as discussed previously, by inspecting the four validity scales. If the test appears reasonably valid by these criteria, the examiner consults a relevant resource book and proceeds scale by scale to produce a series of hypotheses.

The configural approach to MMPI-2 interpretation is somewhat more complicated and consists of classifying the profile as belonging to one or another loosely defined code type that has been studied extensively. Code types are usually defined by a combination of elevation (two or more clinical scales elevated beyond a certain criterion) and definition (two or more clinical scales clearly standing out from the others). For example, in its full-blown manifestation, the 4–9 code type can be defined by a valid profile in which scale 4 (Psychopathic Deviate) and scale 9 (Hypomania) are the high-point elevations, both exceed T of 65 (elevation), and both exceed the next highest clinical scale by at least 5 T-score points (definition).

Several computerized interpretation systems are available for the MMPI and the MMPI-2. The Minnesota Report™ (Butcher, 1993) is the best. This system generates a very cautious and methodical 16-page report that includes discussion of profile validity, symptomatic patterns, interpersonal relations, diagnostic considerations, and treatment considerations. The Minnesota Report™ also provides a variety of figures and tables to illustrate test results. The adequacy of computerized MMPI-2 narrative reports is generally good, but There is a danger that computer-generated test reports will be erroneous.

Technical Properties of the MMPI-2:

From the standpoint of traditional psychometric criteria, the MMPI-2 presents a mixed picture. Reliability data are generally positive, with median internal consistency coefficients (alpha) typically in the .70s and .80s, but as low as the .30s for some scales in some samples. One-week test–retest coefficients range from the high .50s to the low .90s, with a median in the .80s.

A shortcoming of the MMPI-2 is that intercorrelations among the clinical scales are extremely high. For example, in the case of scales 7 and 8, the Psychasthenia and Schizophrenia scales, the correlation is commonly in the .70s. In part, this reflects the item overlap between MMPI scales—scales 7 and 8 share 17 items in common.

The validity of the MMPI-2 is difficult to summarize, owing to the sheer volume of research on this instrument and its predecessor, the MMPI. Graham (1993) provides a brief but excellent review of validity studies on the MMPI/MMPI-2. He notes that the average validity coefficient for MMPI studies conducted between 1970 and 1981 was a healthy .46. He also points out the confirming pattern of extratest correlates in dozens of studies of identified patient groups. Research also indicates that the MMPI-2 is highly comparable to the MMPI, for which a substantial body of validity data has been compiled. . The MMPI-2 likely will maintain its status as the premiere instrument for assessment of psychopathology in adulthood for many years to come.

In 2008, a new version of the MMPI-2 with reduced length and restructured scales was released (Ben-Porath & Tellegen, 2008; Tellegen & Ben-Porath, 2008). Because it embodies a restructured format (RF), the recent entry is called the MMPI-2-RF. This innovative test comprises 338 items carefully selected from the original 567 items of the MMPI-2, using modern psychometric methods for scale construction. Certainly the reduced length is a potential advantage. Patients often tire when completing the MMPI-2, and some find the experience tedious and onerous. Even so, the MMPI-2-RF constitutes a dramatic departure from the parent instrument and is therefore really a new test.

Millon Clinical Multiaxial Inventory-III (MCMI-III):

The MCMI-III is a personality inventory designed for the same purposes as the MMPI-2, namely, to provide useful information for psychiatric diagnosis (Millon, 1983, 1987, 1994). The MCMI-III has two advantages over the MMPI-2. First, it is much shorter (175 true–false items) and, therefore, more palatable to clinical referrals; second, it is planned and organized to identify clinical patterns in a manner that is compatible with the Diagnostic and Statistical Manual (DSM-IV) of the American Psychiatric Association.

The MCMI-III is a highly theory-driven test, incorporating Millon's elaborate theoretical formulations on the nature of psychopathology and personality disorder. The test includes 27 scales. The first 11 scales measure personality styles or traits such as narcissism and antisocial tendencies; the next three assess more severe personality pathology (schizotypal, borderline, and paranoid disorders); the following seven scales assess clinical syndromes such as anxiety and depression; the next three scales assess severe clinical syndromes such as thought disorder; the last three scales are validity (response style) indices. Scores on these scales (Disclosure, Desirability, and Debasement) are used to adjust the other scale scores upward or downward, based on defensiveness or exaggeration of symptoms, respectively.

Scale development for the MCMI-III and its precursors was careful and methodical. 3,500 initial items were culled to 175 statements in three stages of test development: a theoretical-substantive stage (theory-guided item writing), an internal-structural stage (item-scale correlations), and an external criterion stage (contrast of diagnostic groups with the reference group). A special feature of the last stage was Millon's use of general psychiatric patients instead of normal controls as the reference group. The purpose of this strategy was to

enhance the capacity of MCMI scales to differentiate specific diagnostic groups from one another. Unfortunately, one side effect of this particular criterion-keyed approach was a rather substantial degree of item overlap for the clinical scales. Millon planned for and expected the item overlap but probably did not anticipate that some pairs of scales on the MCMI would share the majority of their items in common.

The revised instrument also incorporates an item-weighting procedure. In this approach, individual questions are weighted 2 or 1 to reflect their importance in discriminating the prototype for each scale. The normative sample for the MCMI-III consisted of about a thousand men and women patients from across the United States. More typically, population-proportionate sampling of reasonably normal individuals is used. Millon offers the arguable justification that a patient sample is adequate for the normative sample because the base rates (in the general population) for specific personality and clinical disorders were consulted to calibrate the cutting points on the individual scales.

But this approach is complex, experimental, and difficult to understand. The reliability of the individual scales is good: Internal consistency coefficients average .82 to .90, and test–retest coefficients for one week range from .81 to .87. Support for the validity of the MCMI-III is mixed.

Personality Inventory for Children-2 (PIC-2)

The PIC-2 (Lachar & Gruber, 2001) is a substantial revision of the PIC-R, a popular instrument that dates back to the late 1950s. The current version, suitable for children 5 through 19 years of age, consists of 275 true–false statements that are completed by a parent or parental surrogate. The PIC-2 is one corner of a triad of instruments developed by David Lachar and colleagues to provide a comprehensive, multiview perspective on children's emotional and behavioral adjustment in the home, school, and community. The complementary instruments are the Personality Inventory for Youth (PIY), which is filled out by the child, and the Student Behavior Survey (SBS), which is filled out by the teacher.

The instrument also provides a shorter 96-item version known as the Behavioral Summary, suitable for screening and research purposes. The test developers of the PIC-2 followed a complex multistage methodology to assign individual items to scales and subscales. The goal was to minimize content overlap between scales and subscales by examining preliminary item × subscale correlations and then retaining only those items for each specific subscale that showed high correlations. As a consequence of this test development strategy, each subscale possesses homogeneous content and the individual statements correlate substantially with one another. The resulting instrument consists of three response validity scales (Inconsistency, Dissimulation, Defensiveness) and nine adjustment scales.

Scale raw scores are converted to T scores with a mean of 50 and standard deviation of 10. Higher T scores indicated increased probability of psychopathology or deficit. Norms for children ages 5 through 19 years of age are based on a nationally representative sample of 2,306 parents of boys and girls in kindergarten through 12th grade.

With the possible exception of the three validity scales (Inconsistency, Dissimulation, and Defensiveness), the PIC-2 scale and subscale names are self-explanatory. The validity scales are (1) Inconsistency, which includes 35 similar pairs of items to determine consistency of responding; (2) Dissimulation, a 35-item scale designed to identify deliberate exaggeration

(fake bad) about symptoms or random responding; and (3) Defensiveness, a 24-item scale consisting of improbable virtues (e.g., "my child never has any problems") and therefore an index of naive defensiveness.

The reliability of PIC-2 scales and subscales is good, with test–retest values in the range of .82 to .92 and internal consistency coefficients in the range of .81 to .92. The test manual summarizes a huge body of criterion-related validity studies such as correlations with independent ratings from clinicians. These correlations are very strong for similar behavioral dimensions (and weak for dissimilar behavioral dimensions), thus supporting the validity of individual scales and subscales.

The Adjustment Scales of the PIC-2:

- Cognitive Impairment Scale: Inadequate Abilities Poor Achievement Developmental Delay Impulsivity and Distractibility Scale: Disruptive Behavior Fearlessness
- Delinquency Scale: Antisocial Behavior Dyscontrol Noncompliance
- Family Dysfunction Scale: Conflict among Members Parent Maladjustment Reality Distortion Scale: Developmental Deviation Hallucinations and Delusions
- Somatic Concern Scale: Psychosomatic Preoccupation Muscular Tension and Anxiety
- Psychological Discomfort Scale: Fear and Worry Depression Sleep Disturbance/Death Preoccupation Social Withdrawal Scale: Social Introversion Isolation
- Social Skills Deficits Scale: Limited Peer Status Conflict with Peers

In like manner, PIC-2 subscale scores show theory-consistent relationships with the DSM-IV diagnostic categories of clinic-referred children. For example, 63 children independently diagnosed with Oppositional Defiant Disorder showed highly elevated scores (average T scores of 75 to 80) on the following PIC-2 subscales: Disruptive Behavior, Fearlessness, Dyscontrol, and Noncompliance. This is a perfect match to the major clinical features of this DSM-IV diagnostic category.

BEHAVIORAL ASSESSMENT

Behavioral assessment concentrates on behavior itself rather than on underlying traits, hypothetical causes, or presumed dimensions of personality. The many methods of behavioral assessment offer a practical alternative to projective tests, self-report inventories, and other unwieldy techniques aimed at global personality assessment.

Typically, behavioral assessment is designed to meet the needs of therapists and their clients in a quick and uncomplicated manner. But behavioral assessment differs from traditional assessment in more than its simplicity. The basic assumptions, practical aspects, and essential goals of behavioral and traditional approaches are as different as night and day. Traditional assessment strategies tend to be complex, indirect, psychodynamic, and often extraneous to treatment. In contrast, behavioral assessment strategies tend to be simple, direct, behavior-analytic, and continuous with treatment.

Behavior therapists use a wide range of modalities to evaluate their clients, patients, and subjects. The methods of behavioral assessment include, but are not limited to, behavioral observations, self-reports, parent ratings, staff ratings, sibling ratings, judges' ratings, teacher ratings, therapist ratings, nurses' ratings, physiological assessment, biochemical assessment, biological assessment, structured interviews, semistructured interviews, and analogue tests.

In recent years, a new form of behavioral assessment known as ecological momentary assessment has become increasingly popular. In ecological momentary assessment, the client carries a wireless handheld device similar to a personal digital assistant and responds in real time to preplanned inquiries from the researcher. This approach is designed to circumvent a number of limitations of traditional self-report techniques.

Behavioral assessment is often an integral part of behavior therapy designed to change the duration, frequency, or intensity of a well-defined target behavior. Behavioral assessment often exists in service of behavior therapy. In many cases, the nature of behavioral assessment is dictated by the procedures and goals of behavior therapy. Behavior therapy, also called behavior modification, is the application of the methods and findings of experimental psychology to the modification of maladaptive behavior.

The roots of behavior therapy can be traced to Skinner's (1953) seminal book, Science and Human Behavior, which detailed the application of operant conditioning to the problems of human behavior. Skinner shunned any reference to private, nonobservable events such as thoughts or feelings; he emphasized the importance of identifying observable behaviors and methodically altering the environmental consequences of those behaviors.

Research by Wolpe (1958) on the systematic behavioral treatment of phobias also was influential in founding the methods of behavior therapy. Wolpe's clinical procedures were derived from his laboratory work on the conditioning and counterconditioning of fear in cats. Like Skinner, Wolpe deemphasized the significance of thoughts and beliefs. He viewed fear as a learned phenomenon that could be unlearned by following a strict protocol of graduated exposure to the feared object or situation.

After Skinner, Bandura (1977), Mahoney and Arnkoff (1978), and Meichenbaum (1977) reintroduced cognitive factors into the ever-changing behavioral framework. For example, Bandura (1977) demonstrated that persons are perfectly capable of cognitively based learning. In particular, he showed that individuals can learn from mere observation of the response contingencies experienced by models. Since this learning occurs in the absence of personal consequences, it must be cognitively mediated. As a consequence of this paradigm shift, practically all modern-day behavior therapists concern themselves—at least to some extent—with the thoughts and beliefs of their clients. This new emphasis is reflected in a family of very popular treatment procedures known collectively as cognitive behavior therapy.

Behavior Therapy and Behavioral Assessment:

At present, the specific techniques of behavior therapy can be classified into four overlapping categories (Johnston, 1986): exposure-based methods, cognitive behavior therapies, self-control procedures, and social skills training.

Exposure-Based Methods:

Exposure-based methods of behavioral therapy are well suited to the treatment of phobias, which include intense and unreasonable fears (e.g., of spiders, blood, public speaking). One approach to phobic avoidance is systematic exposure of the client to the feared situation or object. Wolpe (1973) favored gradual exposure with minimal anxiety in a procedure known as systematic desensitization. In this therapeutic approach, the client first learns total relaxation

and then proceeds from imagined exposure to actual or in vivo exposure to the feared stimulus. Another exposure-based method is flooding or implosion in which the client is immediately and totally immersed in the anxiety-inducing situation.

The therapist needs some type of behavioral assessment to gauge the continuing progress of a client undergoing an exposure-based treatment for a phobia. In the simplest possible assessment approach, known as a **behavioral avoidance test (BAT),** the therapist measures how long the client can tolerate the anxiety-inducing stimulus. The researchers discovered that the avoidance anxiety score from the BAT technique was strongly related to self-reports of catastrophic thoughts (e.g., choking to death, having a heart attack, acting foolish, becoming helpless). This finding illustrates that behavioral assessment approaches often encompass a cognitive component as well.

The BAT approach is predicated on the reasonable assumption that the client's fear is the main determinant of behavior in the testing situation. The client's tolerance of the anxietyinducing stimulus will bear some relationship to experienced fear but also has much to do with the situational context of assessment.

A **fear survey schedule** is another type of behavioral assessment useful in the identification and quantification of fears. Fear survey schedules are face valid devices that require respondents to indicate the presence and intensity of their fears in relation to various stimuli, typically on a 5- or 7-point Likert scale. Dozens of these instruments have been published, including versions by Wolpe (1973), Ollendick (1983), and Cautela (1977). Tasto, Hickson, and Rubin (1971) used factor analysis to develop a 40-item survey that yields a profile of fear scores in five categories.

Fear survey schedules are often used in research projects to screen large samples of persons in search of subjects who share a common fear. Another use of these schedules is to monitor changes in fears, including those that have been targeted for clinical intervention.

Cognitive Behavior Therapies:

The one factor common to all cognitive behavior therapies is an emphasis on changing the belief structure of the client. The three best-known variants of cognitive behavior therapy are Ellis's (1962) rational emotive therapy (RET), Meichenbaum's (1977) self-instructional training, and Beck's (1976) cognitive therapy.

Ellis postulates that most disturbed behavior is caused by irrational beliefs, such as the widespread belief that one must have the love and approval of all significant persons at all times. Ellis attempts to alter such core irrational beliefs, primarily by logical argument and forceful exhortation.

Meichenbaum's self-instructional technique consists of teaching the client to use coping self-statements to combat stressful situations. For example, a college student suffering from intense test-taking anxiety might be taught to use the following self-talk during examinations: "You have a strategy this time. . . . Take a deep breath and relax. . . . Just answer one question at a time. . . ."

Beck's cognitive therapy concentrates mainly on the role of cognitive distortions in the maintenance of depression and other emotional disturbances. Beck (1983) regards depression as primarily a cognitive disorder characterized by the negative cognitive triad: a pessimistic

view of the world, a pessimistic self-concept, and a pessimistic view of the future. In therapy, he uses a gentle form of cognitive restructuring to help the client perceive his or her problems in alternative, solvable terms.

Cognitive behavior therapists need not use formal assessment tools in their clinical practice. Typically, these therapists monitor the belief structure of their clients on an informal session-to-session basis. Irrational and distorted thoughts are challenged as they arise during therapy. In the end, the client's self-report of improvement may constitute the main index of therapeutic success. Nonetheless, several straightforward measures of cognitive distortion are available. These instruments are mainly research questionnaires suitable to the testing of group differences, but not sufficiently validated for individual assessment.

Questionnaires to measure cognitive distortion:

1. **Anxious Self-Statements Questionnaire** (ASSQ) (Kendall & Hollon, 1989) : Examinee rates how often specific anxious thoughts occurred over the last week. Items are of the form: I can't stand it anymore. What's going to happen to me now? I'm not going to make it. A psychometrically sound instrument, the ASSQ can be used to assess changes in the frequency of anxious self-talk.
2. **Automatic Thoughts Questionnaire** (ATQ) (Hollon & Kendall, 1980; Kazdin, 1990) : The ATQ is a frequency measure of depression-related cognitions that assesses personal maladjustment, negative self-concept and expectations, low self-esteem, and giving up/ helplessness. The 30-item ATQ correlates very well with the MMPI Depression scale and the Beck Depression Inventory (Ross, Gottfredson, Christensen, & Weaver, 1986).
3. **Cognitive Errors Questionnaire** (CEQ) (Lefebvre, 1981) : The CEQ assesses the degree of maladaptive thinking in general situations and also situations related to chronic low back pain. Discrete vignettes concerning chronic back pain and general scenes are each followed by an illogical dysphoric cognition. The respodent indicates on a 5-point scale how similar the cognition is to the thought he or she would have in the same situation.
4. **Attribution Styles Questionnaire** (ASQ) (Seligman, Abramson, Semmel, & Von Baeyer, 1979) : The ASQ measures three attributional dimensions relevant to Seligman's learned helplessness model of depression: internal-external, stable-unstable, and global-specific. Depressed persons attribute bad outcomes to internal, stable, and global causes; they attribute good outcomes to external, unstable causes. The questionnaire consists of 12 hypothetical situations, 6 describing good outcomes, 6 describing bad outcomes (e.g., "You have been looking for a job unsuccessfully for some time"). The respondents rate each vignette on a 7-point scale for degree of internality, stability, and globality.
5. **Hopelessness Scale** (HS) (Beck, 1987; Dyce, 1996) : A 20-item true/false scale, the HS is designed to quantify hopelessness, one component of the negative cognitive triad found in depressed persons. (The triad consists of negative views of self, world, and future.) The scale is sensitive to changes in the patient's state of depression. In a validational study, Beck, Riskind, Brown, and Steer (1988) found that HS scores had a negligible relationship to anxiety or general psychopathology when the influence of coexisting depression was partialed out. Thus, the HS appears to measure a specific attribute of depression rather than general psychopathology.

Beck Depression Inventory (BDI):

The BDI is a short, simple, self-report questionnaire that focuses, in part, on the cognitive distortions that underlie depression. One reason for its popularity is that most patients can complete the 21 items on the BDI in 10 minutes or less. The test has been widely used: More than 1,900 articles using the BDI have been published. A second edition of the inventory was released in 1996 (Beck, Steer, & Brown, 1996). On the BDI-II, several items were revised so as to bring the inventory into closer conformity with prevailing diagnostic criteria for depression.

Thirteen items cover cognitive and affective components of depression such as pessimism, guilt, crying, indecision, and self-accusations; eight items assess somatic and performance variables such as sleep problems, body image, work difficulties, and loss of interest in sex. The examinee receives a score of 0 to 3 for each item; the total raw score is the sum of the endorsements for the 21 items; the highest possible score is 63.

In a meta-analysis of BDI research studies, the internal consistency of the scale (coefficient alpha) ranged from .73 to .95, with a mean of .86 in nine psychiatric populations. The BDI-II possesses excellent internal consistency with a coefficient alpha of .92. Test–retest reliability of the BDI is modest, with a range of .60 to .83 in nonpsychiatric samples and .48 to .86 in psychiatric samples. However, the test–retest methodology is not well suited to phenomena such as depression that are naturally unstable. Subjective depression fluctuates dramatically from week to week, day to day, even hour to hour.

A variety of normative results are available, with BDI data for samples of patients with major depression, dysthymia, alcoholism, heroin addiction, and mixed problems. The manual also provides guidelines for degree of depression based upon BDI score (0 to 9, normal; 10 to 19, mild to moderate; 20 to 29, moderate to severe; 30 and above, extremely severe). These ratings are based upon clinical evaluations of patients.

The BDI-II is particularly useful in primary care medical settings, where the presence of significant depression can be overlooked. Overall, the BDI-II was 92 percent accurate in identifying patients meeting the formal criteria for Major Depressive Disorder. The only shortcoming of the BDI-II is its transparency. Patients who wish to hide their despair or exaggerate their depression can do so easily. However, for patients who are motivated to accurately report their cognitive and emotional status, the BDI-II ranks among the best instruments for indexing the presence and degree of depression.

Self-Monitoring Procedures:

In self-monitoring, the client chooses the goals and actively participates in supervising, charting, and recording progress toward the end point(s) of therapy. According to this model, the therapist is relegated to the status of expert consultant. Self-monitoring procedures are especially useful in the treatment of depression, a prevalent behavior disorder consisting of sad mood, low activity level, feelings of worthlessness, concentration problems, and physical symptoms (sleep loss, appetite disturbance, reduced interest in sex).

Lewinsohn observed that depression goes hand in hand with a marked reduction in the experiencing of pleasant events. Depressed persons retreat from engaging in pleasant activities; the behavioral withdrawal only contributes further to their depression, inciting a continuous

downward spiral. Fortunately, it is possible to replace the downward spiral with an upward one. To help reverse the downward spiral of depression, Lewinsohn and his colleagues devised the **Pleasant Events Schedule** (PES; MacPhillamy & Lewinsohn, 1982). The purpose of the PES is twofold. First, in the baseline assessment phase, the PES is used to self-monitor the frequency (F) and pleasantness (P) of 320 largely ordinary, everyday events.

Examples of the kinds of events listed on the PES include the following: reading magazines, going for a walk, being with pets, playing a musical instrument, making food for charity, listening to the radio, reading poetry, attending a church service, watching a sports event, playing catch with a friend and working on my job.

The frequency and pleasantness of these everyday events are both rated 0 to 2.6 The mean rate of pleasant activities is then calculated from the sum of the F × P scores; that is, mean rate = F × P/320. Normative findings for mean F, mean P, and mean F × P are reported in Lewinsohn, Munoz, Youngren, and Zeiss (1986) and serve as a basis for treatment planning.

The second use of the PES is to self-monitor therapeutic progress. Based on the initial PES results, clients identify 100 or so potentially pleasant events and strive to increase the frequency of these events, monitoring daily mood along the way. Clients who increase the frequency of pleasant events generally show an improvement in mood and other depressive symptoms. The PES is a highly useful tool for clinicians who wish to implement a self-monitoring approach to the assessment and treatment of depression. The instrument has fair to good test–retest reliability (one-month correlations in the range of .69 to .86), excellent concurrent validity with trained observers, and promising construct validity. In general, the subscales behave as one would predict on the basis of the constructs they purport to measure.

Structured Interview Schedules

An important responsibility for many mental health practitioners is to determine a proper psychiatric diagnosis for their patients, within prevailing guidelines. Almost without exception, practitioners utilize the Diagnostic and Statistical Manual of Mental Disorders, now in its fourth edition (DSM-IV; APA, 2000). The latest version includes a “Text Revision” and for this reason is known technically as DSMIV-TR.

Five axes are included in the DSM-IV classification. Axis I concerns clinical disorders such as Alcohol Use Disorder, Panic Disorder, Major Depressive Disorder, or Schizophrenia. Axis II pertains to personality disorders such as Borderline Personality Disorder, Avoidant Personality Disorder, or Dependent Personality Disorder. Axis III is employed to identify general medical conditions (e.g., hypothyroidism, heart disease) that may bear upon psychological adjustment. Axis IV is for reporting psychosocial and environmental problems (e.g., loss of friends, unemployment, litigation, no health insurance) that may impact personal functioning. Axis V consists of an anchored rating scale, the Global Assessment of Function (GAF) Scale, used to assign a summary score of functioning from 1 (e.g., immobilized, suicidal) to 100 (e.g., thriving, sought out). Of course, intermediate scores are available and clearly operationalized.

Diagnosis is construed by some people as a form of pointless, overconfident, pigeonholing. In truth, it serves a number of indispensable functions. As outlined by Andreasen and Black (1995), these key purposes include: • Reducing the complexity of clinical phenomena • Facilitating communication between clinicians • Predicting the outcome of the

disorder • Deciding on an appropriate treatment • Assisting in the search for etiology • Determining the prevalence of diseases worldwide • Making decisions about insurance coverage.

Several interview schedules have been developed to reduce the time needed for diagnosis and also to improve the reliability of the enterprise by standardizing the procedures. Broadly speaking, these instruments are of two types: semistructured approaches that allow for some clinician leeway in follow-up questioning, and structured approaches that mandate a completely scripted approach.

The Schedule for Affective Disorders and Schizophrenia (SADS; Spitzer & Endicott, 1978):

It is a highly respected diagnostic interview for evaluating Axis I mood and psychotic disorders. The SADS is a semistructured inquiry that includes standard questions asked of all patients and optional probes used to clarify patient responses. Additional unstructured questions can be asked to augment the optional probes. Part I of the SADS methodically examines Axis I symptoms for the current episode, including the worst period and the current week, whereas Part II provides a survey of past episodes. Through a progression of questions and criteria, the interviewer solicits sufficient information to assess the severity of disturbance and also to elucidate the diagnosis. For example, one item on the SADS addresses prominent signs of depression: pessimism and hopelessness.

The consensus from over 21 studies is that the interrater reliability for specific diagnoses is typically strong, with median kappa coefficients of greater than .85. Kappa is the index of interrater agreement, corrected for chance. Validity for the SADS also is robust with moderate predictive validity (e.g., results moderately predict the course and outcome of mood disorders) and strong concurrent validity (e.g., results correlate with other similar schedules). A child's version of the schedule, known as the "kiddie" SADS or K-SADS, also is available (Ambrosini, 2000).

SCID, the Structured Clinical Interview for DSM-IV (First & Gibbon, 2004):

SCID comes in numerous editions and variations, including SCID-I for Axis I diagnoses, SCID-II for Axis II diagnoses, SCID-P for determining the differential diagnosis of psychotic symptoms, and SCID-NP for nonpatient settings in which a current psychiatric disorder is unlikely. All of the forms follow the same format in which the interviewer reads the SCID questions to the client in sequence, the objective being to elicit sufficient information to determine whether individual DSM-IV criteria are met. The interviewer has the leeway to ask for specific examples of affirmative answers. Thus, SCID is a semistructured interview. A logical flow sheet is followed to determine the appropriate diagnosis. The SCID reveals generally good interrater agreement for DSM-IV diagnosis, but this is variable from one diagnosis to the other. In Table 8.12, we have summarized the average kappas from multiple studies of SCID reliability. Kappa values above .70 are considered good agreement, values from .50 to .69 are deemed fair, and values below .50 indicate poor agreement.

Assessment by Systematic Direct Observation

Systematic and direct observation is widely used in the evaluation of children, especially by psychologists who work in school systems. Systematic observation is the single most commonly used assessment method among school-based practitioners.

It is essential to distinguish systematic, direct observation from more casual approaches such as naturalistic observation. Anyone can engage in the informal and anecdotal methods that characterize naturalistic observation—and most people do so every day. These methods typically culminate in formless conclusions such as "Johnny seems to be out of his seat a lot during the school day." In contrast, systematic and direct observation is highly structured and set apart by five characteristics:

1. The goal of observation is to measure specific behaviors.
2. The target behaviors have been operationally defined beforehand.
3. Observations are conducted under objective, standardized procedures.
4. The times and places for observation are carefully specified
5. Scoring is standardized and does not vary from one observer to another.

This form of assessment is appealing because of its direct link to intervention. In fact, it is common to employ observational assessment before, during, and after an intervention to determine the impact on the individual student.

Commonly, systematic and direct observation is executed by means of an objective, structured coding system. Many different styles of coding systems have been proposed.

a. One straightforward approach is **simple frequency counting of target behaviors**. Typically, the target behaviors are undesirable behaviors such as a student leaving his or her seat, calling out, or being off task. Of course, the characteristics of these behaviors would be carefully specified in advance. Then an observer sits off to the side and unobtrusively records the frequency of each behavior within discrete time periods. The purpose of this kind of assessment is to objectify the extent of troublesome actions. This information serves as a baseline for later comparison to determine the effectiveness of any interventions.
b. Another approach to systematic, direct observation is to record **the duration of target behaviors**. Typically, target behaviors are undesirable actions such as temper tantrums, social isolation, or aggressive outbursts, but the focus of assessment also may include desirable behaviors such as staying on task during a designated reading period or vigilantly working on a homework assignment. For some behaviors, duration may be more important than frequency.
c. In addition to the individualized forms of direct observation, dozens of published forms also are available. For these instruments, the categories of observation and the operational definitions are prespecified, which saves time for the practitioner.
d. Shapiro (1996) has issued the **Behavior Observation of Students in Schools (BOSS),** a straightforward form that consists of six categories of classroom behavior—five designed for students and one for the teacher. The BOSS classifies behaviors as active engagement, passive engagement, off-task motor, off-task verbal, and off-task passive. Of course, these categories are thoroughly defined in operational terms. Direct instruction by the teacher also is recorded. The BOSS is rated in 15-second intervals for a 15-minute interval. The instrument also allows for the collection of behavioral norms for classmates to determine normative patterns in each category

Sattler (2002) has catalogued the sources of unreliability, which include personal qualities of the observer, poor design of instruments, and problems in obtaining a representative sample of behavior. For example, observer drift occurs when an observer becomes fatigued and less

vigilant over time, thus failing to notice target behaviors when they occur. Expectations also can influence ratings such as when the observer has been told that a child is aggressive—and then records questionably aggressive acts as aggressive. With regard to poor design of instruments, the most common error is coding complexity, in which there are too many categories or ill-defined categories. Attention to design of rating scales and pretesting of instruments will avert this problem. Problems also can arise in the suitable sampling of behavior.

Analogue Behavioral Assessment

The methods of analogue behavioral assessment are closely related to the methods of systematic, direct observation. The main difference has to do with the settings in which the observations occur. In systematic, direct observation, the assessment of clients takes place in a natural setting such as a classroom. In analogue behavioral assessment, clients are observed in a contrived but plausible setting and also are instructed to engage in relevant tasks designed to elicit behaviors of interest (Haynes, 2001). The goal is to create a state of affairs analogous to pivotal situations in real life—hence, the use of the word analogue.

One application of analogue behavioral assessment is the evaluation of children referred for assessment of behavior or school problems. A specialist who works with these children could dedicate a separate room in his or her clinic to analogue behavioral assessment. The room might resemble a small classroom, complete with blackboard, a few student desks, and bookcases. The referred child would be given a realistic homework assignment and told to work on it for 30 minutes while waiting for the interview. The psychologist then observes through a one-way window and records relevant behaviors using a suitable rating scale.

Analogue behavioral assessment also can be used to evaluate parent–child interactions. For example, in evaluating a 3-year-old referred for behavior problems, the clinician might place the parent and child in a room full of toys with instructions to play for 10 minutes. The psychologist then instructs the parent to tell the child, “Okay, it’s time to go. You have to pick up the toys just like you do at home.” The clinician observes through a one-way window and codes both the parental management style and the nature and degree of child compliance.

In like manner, analogue behavioral assessment has been used in the assessment of adult couples, including husbands and wives seeking marital therapy. In a standard paradigm, the clinician asks the couple to discuss two conflict areas for 5 to 7 minutes each. The clinician sits to the side observing the interactions and recording communication patterns with a standard form such as the **Rapid Couples Interaction Scoring System**. The RCISS consists of 22 codes that address speaker and listener behaviors, both verbal and nonverbal, in such categories as criticism, disagreement, compromise, positive solution, questioning, humor, and smiling. Instruments of this genre typically do not reveal strong interrater agreement for specific constructs (e.g., put-downs), but the more inclusive constructs such as positive affect versus negative affect fare better and provide information that is helpful in characterizing communication patterns.

Ecological Momentary Assessment

Recent advances in wireless connectivity have spawned an entirely new approach to assessment known as ecological momentary assessment (EMA). Ecological momentary

assessment is defined as the “real-time measurement of patient experience in the real world, at the point of experience” (Shiffman, Hufford, & Paty, 2001).

An EMA approach instead would consist of patients reporting their instantaneous experiences on a handheld device, with responses immediately transmitted (via the same wireless technology used by cell phones) to a central computer for ultimate analysis with sophisticated software. For example, the handheld device might “beep” to signal that the patient should immediately respond (on a touch-sensitive screen) to a series of rating scales for pain, mood, fatigue, and other relevant dimensions. The entire selfrating procedure might take less than a minute. The ratings would be requested several times a day on a randomized schedule.

Because EMA responses of clients are immediate and based on a schedule determined by the researcher, several biases of human recall are avoided. For instance, a very brief episode of severe migraine pain may be recalled as lasting much longer than the actual experience because of the emotional valence of the incident. Whereas a retrospective questionnaire report of this pain would be affected by the salience of the event, an EMA analysis, with periodic real-time sampling of the actual pain experiences, would provide a more accurate portrayal of the episode. Recency is another recall bias that is circumvented by EMA. The recency bias refers to the fact that people are more likely to recall recent events than remote events.

In general, EMA provides a more accurate and reliable approach to the assessment of patient experience than traditional approaches such as retrospective questionnaires. One advantage is that compliance cannot be faked (as when patients fill out a week’s worth of daily questionnaires minutes before handing them in to the researcher). In fact, because EMA approaches are highly user-friendly, researchers report an astonishing overall compliance of 93 to 99 percent averaged across many studies. EMA has been used in research into treatments for acute pain, alcoholism, arthritis, asthma, depression, eating disorders, headaches, hypertension, gastrointestinal disorders, schizophrenia, smoking, and urinary incontinence. EMA methodology also can be used to test psychological theories.

Of course, the added advantage of the EMA approach is that data are collected in naturalistic settings in real time, and, therefore, not prone to biases in recall. In some cases, EMA provides for insights that would be difficult to achieve with any other research methodology.