CHAPTER 13

Personality Assessment Methods

Some people see the world as filled with love and goodness, where others see hate and evil. Some people equate *living* with behavioral excess, whereas others strive for moderation in all things. Some people have relatively realistic perceptions of themselves. Others labor under grossly distorted self-images and inaccurate perceptions of family, friends, and acquaintances. For psychologists and others interested in exploring differences among people with regard to these and other dimensions, many different tools are available. In this chapter, we survey some of the tools of personality assessment, including projective methods of assessment and behavioral assessment. We begin with objective methods.

JUST THINK . . .

How objective are objective methods of assessment?

Objective Methods

Typically associated with paper-and-pencil and computer-administered personality tests, **objective methods of personality assessment** characteristically contain short-answer items for which the assessee's task is to select one response from the two or more provided. The scoring is done according to set procedures involving little, if any, judgment on the part of the scorer. As with tests of ability, objective methods of personality assessment may include items written in a multiple-choice, true–false, or matching format.

Whereas a particular response on an objective ability test may be scored *correct* or *incorrect*, a response on an objective personality test is scored with reference to either the personality characteristic(s) being measured or the validity of the respondent's pattern of responses. For example, on a personality test where a *true* response is deemed indicative of the presence of a particular trait, a number of *true* responses to *true–false* items will be interpreted with reference to the presumed strength of that trait in the testtaker. Well, maybe.

If the respondent has also responded *true* to items indicative of the *absence* of the trait as well as to items rarely endorsed as such by testtakers, then the validity of the

protocol will be called into question. Scrutiny of the protocol may suggest an irregularity of some sort. For example, the items may have been responded to inconsistently, in random fashion, or with a *true* response to all questions. As we saw in the previous chapter, some objective personality tests are constructed with validity scales or other devices (such as a forced-choice format) designed to detect or deter response patterns that would call into question the meaningfulness of the scores.

Objective personality tests share many advantages with objective tests of ability. The items can be answered quickly, allowing the administration of many items covering varied aspects of the trait or traits the test is designed to assess. If the items on an objective test are well written then they require little explanation; this makes them well suited for both group and computerized administration. Objective items can usually be scored quickly and reliably by varied means, from hand scoring (usually with the aid of a template held over the test form) to computer scoring. Analysis and interpretation of such tests may be almost as fast as scoring, especially if conducted by computer and custom software.

Although objective personality test items share many characteristics with objective measures of ability, we hasten to add that the adjective *objective* is something of a misnomer when applied to personality testing and assessment. With reference to short-answer items on *ability* tests, the term *objective* gained favor because all items contained only one correct response. Well, that was not always true, either, but that's the way they were designed.

In contrast to the scoring of, say, essay tests, the scoring of objective, multiple-choice tests of ability left little room for emotion, bias, or favoritism on the part of the test scorer. Scoring was dispassionate and—for lack of a better term—objective. But unlike objective ability tests, objective personality tests typically contain no one correct answer. Rather, the selection of a particular choice from multiple-choice items provides information relevant to something about the testtaker—such as the presence, absence, or strength of a personality-related variable. Yes, the scoring of such tests can still be dispassionate and objective. However, the "objectivity" of the score derived from a so-called objective test of personality can be a matter of debate. Consider, for example, a personality test written in an objective test format designed to detect the existence of an unresolved oedipal conflict. The extent to which these test results will be viewed as "objective" is inextricably linked to one's views about the validity of psychoanalytic theory and, more specifically, the construct *oedipal conflict.*

Another issue related to the use of the adjective *objective* to modify *personality test* concerns self-report and the distinct lack of objectivity that can be associated with it. Testtakers' self-reports of what they like or dislike, what they agree or disagree with, what they do or do not do, and so forth can be anything but "objective" for many reasons. Some respondents may lack the insight to respond in what could reasonably be described as an objective manner. Some respondents respond in a manner they believe will place them in the best or worst possible light, depending on their goal. In other words, they can attempt to manage impressions by faking good or faking bad.

Ultimately, the term *objective* as applied to most personality tests may be best thought of as a shorthand description for a test format. Objective personality tests are objective in the sense that they employ a short-answer (typically multiple-choice) format, one that provides little, if any, room for discretion in terms of scoring. To describe a personality test as objective serves to distinguish it from projective and other measurement methods rather than to impart information about the reality, tangibility, or objectivity of scores derived from it.

Projective Methods

Suppose the lights in your classroom were dimmed and everyone was told to stare at the clean chalkboard for a minute or so. And suppose everyone was then asked to take out some paper and write down what they thought could be seen on the chalkboard (other than the chalkboard itself). If you examined what each of your fellow students wrote, you might find as many different things as there were students responding. You could assume that the students saw on the chalkboard—or, more accurately, *projected* onto the chalkboard—something that was not really there but rather was in (or on) their own minds. You might further assume that each student's response to the blank chalkboard reflected something very telling about that student's personality structure.

The **projective hypothesis** holds that an individual supplies structure to unstructured stimuli in a manner consistent with the individual's own unique pattern of conscious and unconscious needs, fears, desires, impulses, conflicts, and ways of perceiving and responding. In like manner, we may define the **projective method** as a technique of personality assessment in which some judgment of the assessee's personality is made on the basis of performance on a task that involves supplying some sort of structure to unstructured or incomplete stimuli. Almost any relatively unstructured stimulus will do for this purpose. In a scene in Shakespeare's play *Hamlet,* Polonius and Hamlet discuss what can be seen in clouds. Indeed, clouds could be used as a projective stimulus.[1] But psychologists, slaves to practicality (and scientific methods) as they are, have developed projective measures of personality that are more reliable than clouds and more portable than chalkboards. Inkblots, pictures, words, drawings, and other things have been used as projective stimuli.

JUST THINK . . .

Name something else that could be used as a projective stimulus for personality assessment purposes. Outline briefly how you might attempt to validate this new test.

Unlike self-report methods, projective tests are *indirect* methods of personality assessment. The examinee's task may be to talk about something or someone other than herself or himself, and inferences about the examinee's personality are made from the response. On such a task, the ability—and presumably the inclination—of examinees to fake is greatly minimized. Also minimized on some projective tasks is the testtaker's need for great proficiency in the English language. Minimal language skills are required to respond to or create a drawing. For that reason, and because some projective methods may be less linked to culture than are other measures of personality, proponents of projective testing believe that there is a promise of cross-cultural utility with these tests that has yet to be fulfilled. Proponents of projective measures also argue that a major advantage of such measures is that they tap unconscious as well as conscious material. In the words of the man who coined the term *projective methods,* "the most important things about an individual are what he cannot or will not say" (Frank, 1939, p. 395).[2]

Projective tests were born in the spirit of rebellion against normative data and through attempts by personality researchers to break down the study of personality into the study of specific traits of varying strengths. This orientation is exemplified by Frank (1939), who reflected: "It is interesting to see how the students of personality

1. In fact, clouds *have* been used as projective stimuli. Wilhelm Stern's Cloud Picture Test, in which subjects were asked to tell what they saw in pictures of clouds, was one of the earliest projective measures.
2. The first published use of the term *projective methods* that we are aware of was in an article entitled "Projective Methods in the Psychological Study of Children" by Ruth Horowitz and Lois Barclay Murphy (1938). However, these authors had read Lawrence K. Frank's (1939) as-yet-unpublished manuscript and credited him for having "applied the term 'projective methods'."

have attempted to meet the problem of individuality with methods and procedures designed for study of uniformities and norms that ignore or subordinate individuality, treating it as a troublesome deviation which derogates from the real, the superior, and only important central tendency, mode, average, etc." (pp. 392–393).

In contrast to methods of personality assessment that focused on the individual from a statistics-based, normative perspective, projective techniques were once the technique of choice for focusing on the individual from a purely clinical perspective—a perspective that examined the unique way an individual projects onto an ambiguous stimulus "his way of seeing life, his meanings, significances, patterns, and especially his feelings" (Frank, 1939, p. 403). Somewhat paradoxically, years of clinical experience with these tests and a mounting volume of research data have led the interpretation of responses to projective stimuli to become increasingly norm-referenced.

Inkblots as Projective Stimuli

Spill some ink in the center of a blank, white sheet of paper and fold it over. Allow to dry. There you have the recipe for an inkblot. Inkblots are not only used by assessment professionals as projective stimuli, they are very much associated with psychology itself in the public eye. The most famous inkblot test is, of course . . .

The Rorschach Hermann Rorschach (Figure 13–1) developed what he called a "form interpretation test" using inkblots as the forms to be interpreted. In 1921 he published his monograph on the technique, *Psychodiagnostics.* In the last section of that monograph, Rorschach proposed applications of his test to personality assessment. He provided 28 case studies employing normal (well, undiagnosed) subjects and people with various psychiatric diagnoses (including neurosis, psychosis, and manic-depressive illness) to illustrate his test. Rorschach died suddenly and unexpectedly at the age of 38, just a year after his book was published. A paper co-authored by Rorschach and Emil Oberholzer entitled "The Application of the Form Interpretation Test" was published posthumously in 1923.

Figure 13–1
Hermann Rorschach (1884–1922)

Rorschach was a Swiss psychiatrist whose father had been an art teacher and whose interests included art as well as psychoanalysis—particularly the work of Carl Jung, who had written extensively on methods of bringing unconscious material to light. In 1913, Rorschach published papers on how analysis of a patient's artwork could provide insights into personality. Rorschach's inkblot test was published in 1921, and it was not an immediate success. Rorschach died of peritonitis the following year at the age of 38, unaware of the great legacy he would leave. For more on Hermann Rorschach, read his Test Developer Profile on our companion Internet site at www.mhhe.com/cohentesting7.

Like Rorschach, we will refer to his test as just that—a test. However, students should be aware of controversy about whether it is properly a test, a method, a technique, or something else. For example, Goldfried et al. (1971) view the Rorschach as a structured interview, and Korchin and Schuldberg (1981) regard it as "less of a test" and more "an open and flexible arena for studying interpersonal transactions" (p. 1151). There has also been debate about whether or not the Rorschach is properly considered a projective instrument (Acklin, 1995; Aronow et al., 1995; Moreland et al., 1995; Ritzler, 1995). For example, John Exner, an authority on all things Rorschach, argued that the inkblots are "not completely ambiguous," that the task does not necessarily "force projection," and that "unfortunately, the Rorschach has been erroneously mislabeled a projective test for far too long" (1989, pp. 526–527; see also Exner, 1997). Regardless, *Rorschach* remains virtually synonymous with *projective test* among assessment professionals and, no matter how else referred to, it certainly qualifies as a "test."

The Rorschach consists of ten bilaterally symmetrical (that is, mirror-imaged if folded in half) inkblots printed on separate cards. Five inkblots are achromatic (meaning without color, or black-and-white). Two inkblots are black, white, and red. The remaining three inkblots are multicolored. The test comes with the cards only; there is no test manual or any administration, scoring, or interpretation instructions. There is no rationale for why some of the inkblots are achromatic and others are chromatic (with color). Unlike most psychological test kits, which today are published complete with test manual and optional, upgradable carrying case, this test contains 10 cards packaged in a cardboard box; that's it.

To fill the need for a test manual and instructions for administration, scoring, and interpretation, a number of manuals and handbooks set forth a variety of methods (such as Aronow & Reznikoff, 1976, 1983; Beck, 1944, 1945, 1952, 1960; Exner, 1974, 1978, 1986; Exner & Weiner, 1982; Klopfer & Davidson, 1962; Lerner, 1991, 1996a, 1996b; Piotrowski, 1957). The system most widely used is the "comprehensive system" devised by Exner. Before describing Exner's scoring system, however, here is a general overview of the process of administering, scoring, and interpreting the Rorschach.

Inkblot cards (similar in some respects to the one shown in Figure 13–2) are initially presented to the testtaker one at a time in numbered order from 1 to 10. The testtaker is instructed to tell what is on each of the cards with a question such as "What might this be?" Testtakers have a great deal of freedom with the Rorschach. They may, for example, rotate the cards and vary the number and length of their responses to each card. The examiner records all relevant information, including the testtaker's verbatim responses, nonverbal gestures, the length of time before the first response to each card, the position of the card, and so forth. The examiner does not engage in any discussion concerning the testtaker's responses during the initial administration of the cards. Every effort is made to provide the testtaker with the opportunity to *project,* free from any outside distractions.

After the entire set of cards has been administered once, a second administration, referred to as the **inquiry,** is conducted. During the inquiry, the examiner attempts to determine what features of the inkblot played a role in formulating the testtaker's **percept** (perception of an image). Questions such as "What made it look like (whatever)?" and "How do you see (whatever it is that the testtaker reported seeing)?" are asked in an attempt to clarify what was seen and which aspects of the inkblot were most influential in forming the perception. The inquiry provides information that is useful in scoring and interpreting the responses. The examiner also learns whether the testtaker remembers earlier responses, whether the original percept is still seen, and whether any new responses are now perceived.

Figure 13–2
A Rorschach-like Inkblot

A third component of the administration, referred to as **testing the limits,** may also be included. This procedure enables the examiner to restructure the situation by asking specific questions that provide additional information concerning personality functioning. If, for example, the testtaker has utilized the entire inkblot when forming percepts throughout the test, the examiner might want to determine if details within the inkblot could be elaborated on. Under those conditions, the examiner might say, "Sometimes people use a part of the blot to see something." Alternatively, the examiner might point to a specific area of the card and ask, "What does this look like?"

Other objectives of limit-testing procedures are (1) to identify any confusion or misunderstanding concerning the task, (2) to aid the examiner in determining if the testtaker is able to refocus percepts given a new frame of reference, and (3) to see if a testtaker made anxious by the ambiguous nature of the task is better able to perform given this added structure. At least one Rorschach researcher has advocated the technique of trying to elicit one last response from testtakers who think they have already given as many responses as they are going to give (Cerney, 1984). The rationale was that endings have many meanings, and the one last response may provide a source of questions and inferences applicable to treatment considerations.

JUST THINK . . .

The Rorschach is viewed by some as more of a structured interview than a test. What arguments could be made to support that point of view?

Hypotheses concerning personality functioning will be formed on the basis of all the variables we have outlined (such as the content of the response, the location of the response, the length of time to respond) as well as many additional ones. In general, Rorschach protocols are scored according to several categories, including location, determinants, content, popularity, and form. *Location* is the part of the inkblot that was utilized in forming the percept. Individuals may use the entire inkblot, a large section, a small section, a minute detail, or white spaces. *Determinants* are the qualities of the inkblot that determine what the individual perceives. Form, color, shading, or movement that the individual attributes to the inkblot are all considered determinants. *Content* is the content category of the response. Different scoring systems vary in some of the categories scored. Some typical content areas include human figures, animal figures, anatomical parts, blood, clouds, X-rays, and sexual responses. *Popularity* refers to the frequency with which a certain response has been found to correspond with a particular inkblot or section of an inkblot. A popular response is one that has frequently been obtained from the general population. A rare response is one that has been perceived infrequently by the general population. The *form* of a response is how accurately the individual's perception matches or fits the corresponding part of

the inkblot. Form level may be evaluated as being adequate or inadequate or as good or poor.

The scoring categories are considered to correspond to various aspects of personality functioning. Hypotheses concerning aspects of personality are based both on the number of responses that fall within each category and on the interrelationships among the categories. For example, the number of whole responses (using the entire inkblot) in a Rorschach record is typically associated with conceptual thought process. Form level is associated with reality testing. Accordingly, psychotic patients would be expected to achieve low scores for form level. Human movement has been associated with creative imagination. Color responses have been associated with emotional reactivity.

Patterns of response, recurrent themes, and the interrelationships among the different scoring categories are all considered in arriving at a final description of the individual from a Rorschach protocol. Data concerning the responses of various clinical and nonclinical groups of adults, adolescents, and children have been compiled in various books and research publications.

Rorschach's form interpretation test was in its infancy at the time of its developer's death. The orphaned work-in-progress found a receptive home in the United States, where it was nurtured by several different schools, each with its own vision of how the test should be administered, scored, and interpreted. In this sense, the Rorschach is, as McDowell and Acklin (1996, p. 308) characterized it, "an anomaly in the field of psychological measurement when compared to objective and other projective techniques."

Although the test is widely called "the Rorschach" as though it were a standardized test, Rorschach practitioners and researchers have for many years employed a variety of Rorschach systems—on some occasions picking and choosing interpretive criteria from one or more systems. Consider in this context a study by Saunders (1991) that focused on Rorschach indicators of child abuse. Saunders wrote: "Rorschach protocols were scored using Rapaport et al.'s (1945–1946) system as the basic framework, but special scores of four different types were added. I borrowed two of these additional measures from other researchers . . . and developed the other two specifically for this study" (p. 55). Given the variation that existed in terminology and in administration and scoring practices, one readily appreciates how difficult it might be to muster consistent and credible evidence for the test's psychometric soundness.[3]

In a book that reviewed several Rorschach systems, John E. Exner, Jr. (Figure 13–3) wrote of the advisability of approaching "the Rorschach problem through a research integration of the systems" (1969, p. 251). Exner would subsequently develop such an integration—a "comprehensive system," as he called it (Exner 1974, 1978, 1986, 1990, 1991, 1993a, 1993b; Exner & Weiner, 1982, 1995; see also Handler, 1996)—for the test's administration, scoring, and interpretation. Exner's system has been well received by clinicians and is probably the system most used and most taught today. However, to inextricably link the fate of the Rorschach to Exner's system would be unfair, at least according to Bornstein and Masling (2005); Exner's system has much to recommend it but so do several other systems.

Prior to the development of Exner's system and its widespread adoption by clinicians and researchers, evaluations of the Rorschach's psychometric soundness tended to be mixed at best. Exner's system brought a degree of uniformity to Rorschach use and thus facilitated "apples-to-apples" (or "bats-to-bats") comparison of research studies.

---

3. A test called the Holtzman Inkblot Technique (HIT; Holtzman et al., 1961) was designed to be more psychometrically sound than any existing inkblot test. A description of the HIT, as well as speculation as to why it never achieved the popularity and acceptance of the Rorschach, can be found in the companion workbook to this text, *Exercises in Psychological Testing and Assessment* (Cohen, 2010).

Figure 13–3
John Ernest Exner Jr.

In their obituary of John E. Exner, Jr., Erdberg and Weiner (2007, p. 54) wrote: "Many psychologists bounce around a bit before they lock in on the specialty that becomes the focus of their professional life. That was not the case with John Exner. He first laid hands on a set of blots from the Rorschach Inkblot Test in 1953, and his fascination with the instrument anchored his career from then on. Through five decades, 14 books, more than 60 journal articles, and countless workshop and conference presentations, John Exner and the Rorschach became synonymous." Among other accomplishments, Exner was the founding curator of the Hermann Rorschach Museum and Archives in Bern, Switzerland. One of his last publications before his death at the age of 77 from leukemia was an article entitled "A New U.S. Adult Nonpatient Sample." In that article, Exner discussed implications for modifying Comprehensive System interpretive guidelines based on new data (Exner, 2007).

Yet, regardless of the scoring system employed, there were a number of reasons why the evaluation of the psychometric soundness of the Rorschach was a tricky business. For example, because each inkblot is considered to have a unique stimulus quality, evaluation of reliability by a split-half method would be inappropriate. Of historical interest in this regard is the work of Behn, who attempted to develop, under Sigmund Freud's direction, a similar but not alternate form of the test called the Behn-Rorschach (Buckle & Holt, 1951; Eichler, 1951; Swift, 1944).

Traditional test-retest reliability procedures were also inappropriate for use with the Rorschach. This is so because of the effect of familiarity on response to the cards and because responses may reflect transient states as opposed to enduring traits. Relevant to the discussion of the Rorschach's reliability is Exner's (1983) reflection that "some Comprehensive System scores defy the axiom that something cannot be valid unless it is also reliable" (p. 411).

The widespread acceptance of Exner's system has advanced the cause of Rorschach reliability—well, inter-scorer reliability, anyway. Exner, as well as others, have provided ample evidence that acceptable levels of inter-scorer reliability can be attained with the Rorschach. Using Exner's system, McDowell and Acklin (1996) reported an overall mean percentage agreement of 87% among Rorschach scorers. Still, as these researchers cautioned, "The complex types of data developed by the Rorschach introduce formidable obstacles to the application of standard procedures and canons of test development" (pp. 308–309). Far more pessimistic about such "formidable obstacles" and far less subtle in their conclusions were Hunsley and Bailey (1999). After reviewing the literature on the clinical utility of the Rorschach, they wrote of "meager support from thousands of publications" and expressed doubt that evidence would ever be developed that the Rorschach or Exner's comprehensive system could "contribute, in routine clinical practice, to scientifically informed psychological assessment" (p. 274).

JUST THINK . . .

Is it possible for scores on a test to defy the axiom that the score cannot be valid unless it is reliable?

Countering such pessimism are other reviews of the literature that are far more favorable to this test (Bornstein, 1998, 1999; Ganellen, 1996; 2007; Hughes et al., 2007; Meyer & Handler, 1997; Viglione, 1999). In their meta-analysis designed to compare the validity of the Rorschach with that of the MMPI, Hiller et al. (1999) concluded that "on average, both tests work about equally well when used for purposes deemed appropriate by experts" (p. 293). In a similar vein, Stricker and Gold (1999, p. 240) reflected that a test is not valid or invalid; rather, there are as many validity coefficients as there are purposes for which the test is used. The Rorschach can demonstrate its utility for several purposes and can be found wanting for several others.

Stricker and Gold went on to argue for an approach to assessment that incorporated many different types of methods:

> Arguably, Walt Whitman's greatest poem was entitled "Song of Myself." We believe that everything that is done by the person being assessed is a song of the self. The Rorschach is one instrument available to the clinician, who has the task of hearing all of the music. (1999, p. 249)

Decades ago, Jensen (1965, p. 509) opined that "the rate of scientific progress in clinical psychology might well be measured by the speed and thoroughness with which it gets over the Rorschach." If this statement were true, then the rate of scientific progress in clinical psychology could be characterized as a crawl. The Rorschach remains one of the most frequently used and frequently taught psychological tests. It is widely used in forensic work and generally accepted by the courts. As Weiner (1997) concluded in his evaluation of the status of the Rorschach at age 75, "Widely used and highly valued by clinicians and researchers in many countries of the world, it appears despite its fame not yet to have received the academic respect it deserves and, it can be hoped, will someday enjoy" (p. 17).

JUST THINK . . .

"If the Rorschach has anything at all going for it, it has great intuitive appeal." Explain.

Pictures as Projective Stimuli

Look at Figure 13–4. Now make up a story about it. Your story should have a beginning, a middle, and an end. Write it down, using as much paper as you need. Bring the story to class with you and compare it with some other student's story. What does the story reveal about your needs, fears, desires, impulse control, ways of viewing the world—your personality? What does the story written by your classmate reveal about her or him?

This exercise introduces you to the use of pictures as projective stimuli. Pictures used as projective stimuli may be photos of real people, animals, objects, or anything. They may be paintings, drawings, etchings, or any other variety of picture.

One of the earliest uses of pictures as projective stimuli came at the beginning of the twentieth century. Long before the "men are from Mars, women are from Venus" stuff, sex differences were reported in the stories that children gave in response to nine pictures (Brittain, 1907). The author reported that the girls were more interested in religious and moral themes than the boys were. Another early experiment using pictures and a storytelling technique investigated children's imagination. Differences in themes as a function of age were observed (Libby, 1908). In 1932, a psychiatrist working at the Clinic for Juvenile Research in Detroit developed the Social Situation Picture Test (Schwartz, 1932), a projective instrument that contained pictures relevant to juvenile delinquents. Working at the Harvard Psychological Clinic in 1935, Christiana D. Morgan (Figure 13–5) and Henry Murray (Figure 13–6) published the Thematic

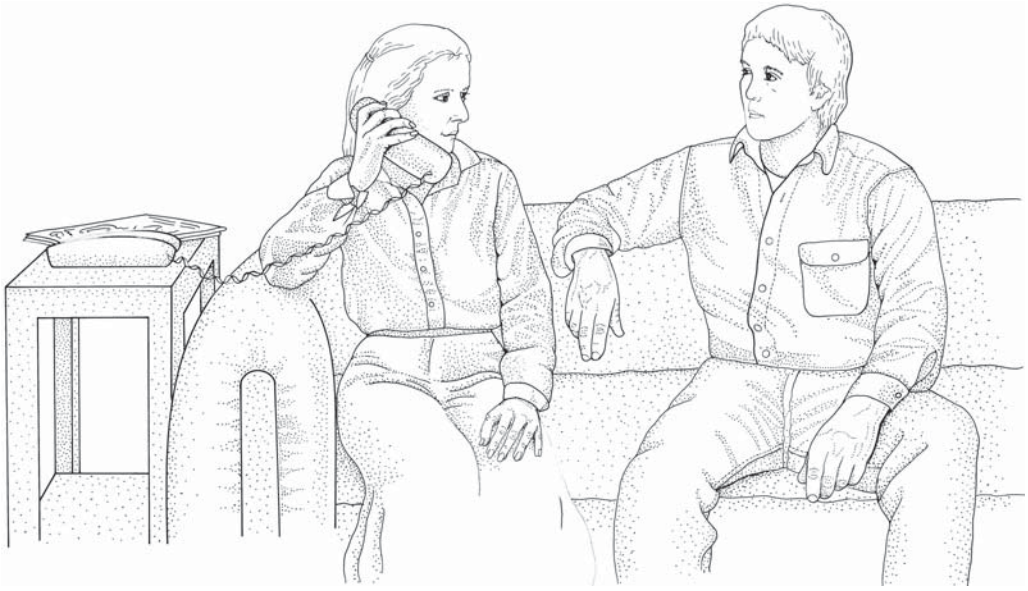

Figure 13–4
Ambiguous Picture for Use in Projective Storytelling Task

Apperception Test (TAT)—pronounced by saying the letters, not by rhyming with *cat*—the instrument that has come to be the most widely used of all the picture storytelling projective tests.

The Thematic Apperception Test (TAT) The TAT was originally designed as an aid to eliciting fantasy material from patients in psychoanalysis (Morgan & Murray, 1935). The stimulus materials consisted, as they do today, of 31 cards, one of which is blank. The 30 picture cards, all black-and-white, contain a variety of scenes designed to present the testtaker with "certain classical human situations" (Murray, 1943). Some of the pictures contain a lone individual, some contain a group of people, and some contain no people. Some of the pictures appear to be as real as a photo; others are surrealistic drawings. Testtakers are introduced to the examination with the cover story that it is a test of imagination in which it is their task to tell what events led up to the scene in the picture, what is happening at that moment, and what the outcome will be. Testtakers are also asked to describe what the people depicted in the cards are thinking and feeling. If the blank card is administered, examinees are instructed to imagine that there is a picture on the card and then proceed to tell a story about it.

In the TAT manual, Murray (1943) also advised examiners to attempt to find out the source of the examinee's story. It is noteworthy that the noun *apperception* is derived from the verb **apperceive,** which may be defined as *to perceive in terms of past perceptions.* The source of a story could be a personal experience, a dream, an imagined event, a book, an episode of *South Park*—really almost anything.

In clinical practice, examiners tend to take liberties with various elements pertaining to the administration, scoring, and interpretation of the TAT. For example, although 20 cards is the recommended number for presentation, in practice an examiner might administer as few as one or two cards or as many as all 31. If a clinician is assessing a patient who has a penchant for telling stories that fill reams of the clinician's notepad, it's a good bet that fewer cards will be administered. If, on the other hand, a patient tells

Figure 13–5
Christiana D. Morgan (1897–1967)

On the box cover of the widely used TAT and in numerous other measurement-related books and articles, the authorship of the TAT is listed as "Henry A. Murray, Ph.D., and the Staff of the Harvard Psychological Clinic." However, the first articles describing the TAT were written by Christiana D. Morgan (Morgan, 1938) or by Morgan and Murray with Morgan listed as senior author (Morgan & Murray, 1935, 1938). In a mimeographed manuscript in the Harvard University archives, an early version of the test was titled the "Morgan-Murray Thematic Apperception Test" (White et al., 1941). Wesley G. Morgan (1995) noted that, because Christiana Morgan "had been senior author of the earlier publications, a question is raised about why her name was omitted as an author of the 1943 version" (p. 238); Morgan took up that and related questions in a brief but fascinating account of the origin and history of the TAT images. More on the life of Christiana Morgan can be found in Translate This Darkness: The Life of Christiana Morgan *(Douglas, 1993). Her Test Developer Profile can be found on our Internet site at* www.mhhe.com/cohentesting7.

brief, one-or two-sentence stories, more cards may be administered in an attempt to collect more raw data with which to work. Some of the cards are suggested for use with adult males, adult females, or both, and some are suggested for use with children. This is so because certain pictorial representations lend themselves more than others to identification and projection by members of these groups. In one study involving 75 males (25 each of 11-, 14-, and 17-year-olds), Cooper (1981) identified the ten most productive cards for use with adolescent males. In practice, however, any card—be it one recommended for use with males, with females, or with children—may be administered to any subject. The administering clinician selects the cards that are believed likely to elicit responses pertinent to the objective of the testing.

JUST THINK . . .

And just imagine . . . describe a picture on a card that would really get *you* talking. What would you say?

The raw material used in deriving conclusions about the individual examined with the TAT are (1) the stories as they were told by the examinee, (2) the clinician's notes about the way or the manner in which the examinee responded to the cards, and (3) the clinician's notes about extra-test behavior and verbalizations. The last two categories of

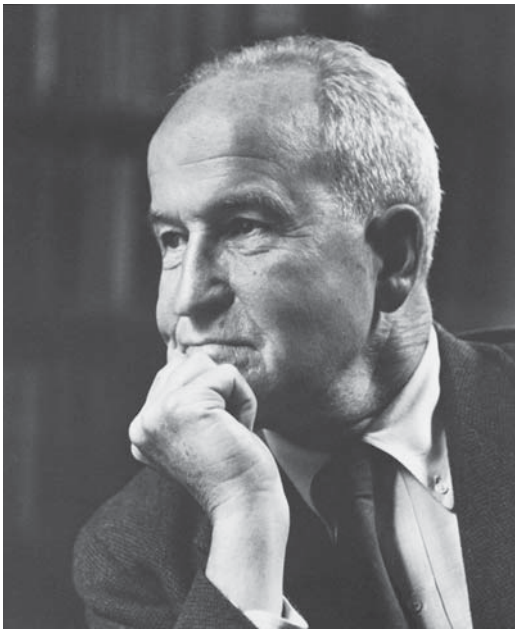

Figure 13–6
Henry A. Murray (1893–1988)

Henry Murray is perhaps best known for the influential theory of personality he developed, as well as for his role as author of the Thematic Apperception Test. Biographies of Murray have been written by Anderson (1990) and Robinson (1992). Murray's Test Developer Profile can be found on the Internet at www.mhhe.com/cohentesting7.

raw data (test and extra-test behavior) are sources of clinical interpretations for almost any individually administered test. Analysis of the story content requires special training. One illustration of how a testtaker's behavior during testing may influence the examiner's interpretations of the findings was provided by Sugarman (1991, p. 140), who told of a "highly narcissistic patient [who] demonstrated contempt and devaluation of the examiner (and presumably others) by dictating TAT stories complete with spelling and punctuation as though the examiner was a stenographer."

A number of systems for interpreting TAT data exist (for example, Thompson, 1986; Westen et al., 1988). Most of these interpretive systems incorporate, or are to some degree based on, Henry Murray's concepts of **need** (determinants of behavior arising from within the individual), **press** (determinants of behavior arising from within the environment), and **thema** (a unit of interaction between needs and press). In general, the guiding principle in interpreting TAT stories is that the testtaker is identifying with someone (the protagonist) in the story and that the needs, environmental demands, and conflicts of the protagonist in the story are in some way related to the concerns, hopes, fears, or desires of the examinee.

In his discussion of the TAT from the perspective of a clinician, William Henry (1956) examined each of the cards in the test with regard to such variables as *manifest stimulus demand, form demand, latent stimulus demand, frequent plots,* and *significant variations.* To get an idea of how some of these terms are used, look again at Figure 13–4—a picture that is *not* a TAT card—and then review Tables 13–1 and 13–2, which are descriptions of the card and some responses to the card from college-age respondents. Although a clinician may obtain bits of information from the stories told about every individual card, the clinician's final impressions will usually derive from a consideration of the overall patterns of themes that emerge.

As with the Rorschach and many other projective techniques, a debate between academics and practitioners regarding the psychometric soundness of the TAT has been unceasing through the years. Because of the general lack of standardization and uniformity with which administration, scoring, and interpretation procedures tend to

Table 13–1
A Description of the Sample TAT-like Picture

Author's Description

A male and a female are seated in close proximity on a sofa. The female is talking on the phone. There is an end table with a magazine on it next to the sofa.

Manifest Stimulus Demand

Some explanation of the nature of the relationship between these two persons and some reason the woman is on the phone are required. Less frequently noted is the magazine on the table and its role in this scene.

Form Demand

Two large details, the woman and the man, must be integrated. Small details include the magazine and the telephone.

Latent Stimulus Demand

This picture is likely to elicit attitudes toward heterosexuality and, within that context, material relevant to the examinee's "place" on optimism-pessimism, security-insecurity, dependence-independence, passivity-assertiveness, and related continuums. Alternatively, attitudes toward family and friends may be elicited, with the two primary figures being viewed as brother and sister, the female talking on the phone to a family member, and so on.

Frequent Plots

We haven't administered this card to enough people to make judgments about what constitutes "frequent plots." We have, however, provided a sampling of plots (Table 12–2).

Significant Variations

Just as we cannot provide information on frequent plots, we cannot report data on significant variations. We would guess, however, that most college students viewing this picture would perceive the two individuals in it as being involved in a heterosexual relationship. Were that to be the case, a significant variation would be a story in which the characters are not involved in a heterosexual relationship (for example, they are employer/employee). Close clinical attention will also be paid to the nature of the relationship of the characters to any "introduced figures" (persons not pictured in the card but introduced into the story by the examinee). The "pull" of this card is to introduce the figure to whom the woman is speaking. What is the phone call about? How will the story be resolved?

be applied in everyday clinical practice, concern on psychometric grounds is clearly justified. However, in experimental tests where trained examiners use the same procedures and scoring systems, inter-rater reliability coefficients can range from adequate to impressive (Stricker & Healy, 1990).

Research suggests that situational factors—including who the examiner is, how the test is administered, and the testtaker's experiences prior to and during the test's administration—may affect test responses. Additionally, transient internal need states such as hunger, thirst, fatigue, and higher-than-ordinary levels of sexual tension can affect a testtaker's responses. Different TAT cards have different stimulus "pulls" (Murstein & Mathes, 1996). Some pictures are more likely than others to elicit stories with themes of despair, for example. Given that the pictures have different stimulus "pulls" or, more technically stated, different latent stimulus demands, it becomes difficult if not impossible to determine the inter-item (read "inter-card") reliability of the test. Card 1 might reliably elicit themes of need for achievement, whereas card 16, for example, might not typically elicit any such themes. The possibility of widely variable story lengths in response to the cards presents yet another challenge to the documentation of inter-item reliability.

JUST THINK . . .

Why are split-half, test-retest, and alternate-form reliability measures inappropriate for use with the TAT?

Table 13–2
Some Responses to the Sample Picture

Respondent	Story
1. (Male)	This guy has been involved with this girl for a few months. Things haven't been going all that well. He's suspected that she's been seeing a lot of guys. This is just one scene in a whole evening where the phone hasn't stopped ringing. Pretty soon he is just going to get up and leave.
2. (Female)	This couple is dating. They haven't made any plans for the evening, and they are wondering what they should do. She is calling up another couple to ask if they want to get together. They will go out with the other couple and have a good time.
3. (Male)	This girl thinks she is pregnant and is calling the doctor for the results of her test. This guy is pretty worried because he has plans to finish college and go to graduate school. He is afraid she will want to get married, and he doesn't want to get trapped into anything. The doctor will tell her she isn't pregnant, and he'll be really relieved.
4. (Female)	This couple has been dating for about two years, and they're very much in love. She's on the phone firming up plans for a down payment on a hall that's going to cater the wedding. That's a bridal magazine on the table over there. They look like they're really in love. I think things will work out for them even though the odds are against it—the divorce rates and all.
5. (Male)	These are two very close friends. The guy has a real problem and needs to talk to someone. He is feeling really depressed and that he is all alone in the world. Every time he starts to tell her how he feels, the phone rings. Pretty soon he will leave feeling like no one has time for him and even more alone. I don't know what will happen to him, but it doesn't look good.

Conflicting opinions are presented in the scholarly literature concerning the validity of the TAT, including the validity of its assumptions and the validity of various applications (Barends et al., 1990; Cramer, 1996; Gluck, 1955; Hibbard et al., 1994; Kagan, 1956; Keiser & Prather, 1990; Mussen & Naylor, 1954; Ronan et al., 1995; Worchel & Dupree, 1990). Some have argued that as much motivational information could be obtained through much simpler, self-report methods. However, one meta-analysis of this literature concluded that there was little relation between TAT-derived data and that derived from self-report (Spangler, 1992). McClelland et al. (1989) distinguished the products of self-report and TAT-derived motivational information, arguing that self-report measures yielded "self-attributed motives" whereas the TAT was capable of yielding "implicit motives." Drawing partially on McClelland et al. (1989), we may define an **implicit motive** as a nonconscious influence on behavior typically acquired on the basis of experience.

JUST THINK . . .

If someone asked you about your "need to achieve," what would you say? How might what you say differ from the "implicit" measure of need for achievement that would emerge from your TAT protocol?

Although the relationship between expression of fantasy stories and real-life behavior is tentative at best, and although the TAT is highly susceptible to faking, the test is widely used by practitioners. Yet in contrast to the test's apparently widespread use are the findings of one survey of training directors of APA-approved programs in clinical psychology: The majority of these programs place little emphasis on the test and typically rely on psychoanalytic writings in their teaching of it (Rossini & Moretti, 1997).

A study by Peterson et al. (2008) provided partial support not only for the projective hypothesis but also for the value of the TAT in clinical assessment. The research subjects were 126 introductory psychology students (70 female, 56 male) with an average age of about 19½, They were pre-evaluated by self-report measures of personality and mood and also by a demographic questionnaire. Subjects were then exposed to rock music with suicide-related lyrics. The specific songs used were *Dirt, Desperate Now,* and *Fade to Black.* Subjects next completed a memory test for the music they had heard,

self-report measures of personality and mood, and a picture storytelling task using three TAT cards. Among the many findings, of particular interest here is that measured personality traits predicted the level of suicide-related responding in the TAT stories told. Participants who wrote stories with higher levels of suicide-related responding (a) tended to believe that suicidal thinking was valid, and that suicide-related lyrics in songs were potentially harmful, (b) felt more sad, angry, and isolated while listening to the music, and, (c) were more likely to report negative affect states after listening to the music. One unexpected finding from this study was that

> after listening to music with suicide lyrics, many participants wrote projective stories with altruistic themes. . . . There is a vast literature relating exposure to violence in music, video games, and movies to increased aggression but Meier [et al.] 2006 reported that this relationship does not occur for individuals who score high on measures of agreeableness. Indeed, such individuals respond to aggression-related cues by accessing pro-social thoughts. (Peterson et al., 2008, p. 167)

The rationale of the TAT, and of many similar published picture story tests (see Table 13–3), has great intuitive appeal. It does make sense that people would project their own motivation when asked to construct a story from an ambiguous stimulus. Another appeal for users of this test is that it is the clinician who tailors the test administration by selecting the cards and the nature of the inquiry—an undoubtedly welcome feature in the era of standardization, computer-adaptive testing, and computer-generated narrative summaries. But as with many projective tests, it seems that the TAT must ultimately be judged by a different standard—one more clinically than psychometrically oriented—if its contribution to personality assessment is to be fully appreciated.

JUST THINK . . .

Should all tests be measured by the same "psychometric yardstick"?

Other tests using pictures as projective stimuli A projective technique called the Hand Test (Wagner, 1983) consists of nine cards with pictures of hands on them and a tenth blank card. The testtaker is asked what the hands on each card might be doing. When presented with the blank card, the testtaker is instructed to imagine a pair of hands on the card and then describe what they might be doing. Testtakers may make several responses to each card, and all responses are recorded. Responses are interpreted according to 24 categories such as affection, dependence, and aggression.

Another projective technique, the Rosenzweig Picture-Frustration Study (Rosenzweig, 1945, 1978), employs cartoons depicting frustrating situations (Figure 13–7). The testtaker's task is to fill in the response of the cartoon figure being frustrated. The test, which is based on the assumption that the testtaker will identify with the person being frustrated, is available in forms for children, adolescents, and adults. Young children respond orally to the pictures, whereas older testtakers may respond either orally or in writing. An inquiry period is suggested after administration of all of the pictures in order to clarify the responses.

Test responses are scored in terms of the type of reaction elicited and the direction of the aggression expressed. The direction of the aggression may be *intropunitive* (aggression turned inward), *extrapunitive* (outwardly expressed), or *inpunitive* (aggression is evaded so as to avoid or gloss over the situation). Reactions are grouped into categories such as *obstacle dominance* (in which the response concentrates on the frustrating barrier), *ego defense* (in which attention is focused on protecting the frustrated person), and *need persistence* (in which attention is focused on solving the frustrating problem). For each scoring category, the percentage of responses is calculated and compared with normative data. A group conformity rating (GCR) is derived representing the degree to

Table 13–3
Some Picture Story Tests

Picture-Story Test	Description
Thompson (1949) modification of the original TAT	Designed specifically for use with African American testtakers, with pictures containing both Black and White protagonists.
TEMAS (Malgady et al., 1984)	Designed for use with urban Hispanic children, with drawings of scenes relevant to their experience.
Children's Apperception Test (CAT; Bellak, 1971) (first published in 1949)	Designed for use with ages 3 to 10 and based on the idea that animals engaged in various activities were useful in stimulating projective storytelling by children.
Children's Apperception Test-Human (CAT-H; Bellak & Bellak, 1965)	A version of the CAT based on the idea that depending on the maturity of the child, a more clinically valuable response might be obtained with humans instead of animals in the pictures.
Senior Apperception Technique (SAT; Bellak & Bellak, 1973)	Picture-story test depicting themes relevant to older adults.
The Picture Story Test (Symonds, 1949)	For use with adolescents, with pictures designed to elicit adolescent-related themes such as coming home late and leaving home.
Education Apperception Test (Thompson & Sones, 1973) and the School Apperception Method (Solomon & Starr, 1968)	Two independent tests, listed here together because both were designed to tap school-related themes.
The Michigan Picture Test (Andrew et al., 1953)	For ages 8 to 14, contains pictures designed to elicit various themes ranging from conflict with authority to feelings of personal inadequacy.
Roberts Apperception Test for Children (RATC; McArthur & Roberts, 1982)	Designed to elicit a variety of developmental themes such as family confrontation, parental conflict, parental affection, attitudes toward school, and peer action.
Children's Apperceptive Story-Telling Test (CAST; Schneider, 1989)	Theory-based test based on the work of Alfred Adler.
Blacky Pictures Test (Blum, 1950)	Psychoanalytically based, cartoon-like items featuring Blacky the Dog.
Make a Picture Story Method (Shneidman, 1952)	For ages 6 and up, respondents construct their own pictures from cutout materials included in the test kit and then tell a story.

which one's responses conform to or are typical of those of the standardization group. This test has captured the imagination of researchers for decades, although questions remain concerning how reactions to cartoons depicting frustrating situations are related to real-life situations.

One variation of the picture story method may appeal to "old school" clinicians as well as to clinicians who thrive on normative data with all of the companion statistics. The Apperceptive Personality Test (APT; Karp et al., 1990) represents an attempt to address some long-standing criticisms of the TAT as a projective instrument while introducing objectivity into the scoring system. The test consists of eight stimulus cards "depicting recognizable people in everyday settings" (Holmstrom et al., 1990, p. 252), including males and females of different ages as well as minority group members. This, by the way, is in contrast to the TAT stimulus cards, some of which depict fantastic or unreal types of scenes.[4] Another difference between the APT and the TAT is the emotional tone and draw of the stimulus cards. A long-standing criticism of the TAT cards has been their negative or gloomy tone, which may restrict the range of affect projected by a testtaker (Garfield & Eron, 1948; Ritzler et al., 1980). After telling a story about each of the APT pictures orally or in writing, testtakers respond to a series of multiple-choice

4. Murray et al. (1938) believed that fantastic or unreal types of stimuli might be particularly effective in tapping unconscious processes.

Figure 13–7
Sample Item from the Rosenzweig Picture-Frustration Study

questions. In addition to supplying quantitative information, the questionnaire segment of the test was designed to fill in information gaps from stories that are too brief or cryptic to otherwise score. Responses are thus subjected to both clinical and actuarial interpretation and may, in fact, be scored and interpreted with computer software.

Every picture tells a story—well, hopefully for the sake of the clinician or researcher trying to collect data. Otherwise, it may be time to introduce another type of test, one where words themselves are used as projective stimuli.

Words as Projective Stimuli

Projective techniques that employ words or open-ended phrases and sentences are referred to as *semi-structured* techniques because, although they allow for a variety of responses, they still provide a framework within which the subject must operate. Perhaps the two best-known examples of verbal projective techniques are *word association tests* and *sentence completion tests.*

Word association tests In general, a word association test may be defined as a semi-structured, individually administered, projective technique of personality assessment that involves the presentation of a list of stimulus words, to each of which an assessee responds verbally or in writing with whatever comes to mind first upon hearing the word. Responses are then analyzed on the basis of content and other variables. The first attempt to investigate word association was made by Galton (1879). Galton's method consisted of presenting a series of unrelated stimulus words and instructing the subject to respond with the first word that came to mind. Continued interest in the phenomenon of word association resulted in additional studies. Precise methods were developed for recording the responses given and the length of time elapsed before obtaining a response (Cattell, 1887; Trautscholdt, 1883). Cattell and Bryant (1889) were the first to use cards with stimulus words printed on them. Kraepelin (1895) studied the effect of physical states (such as hunger and fatigue) and of practice on word association. Mounting experimental evidence led psychologists to believe that the associations individuals made to words were not chance happenings but rather the

result of the interplay between one's life experiences, attitudes, and unique personality characteristics.

Jung (1910) maintained that, by selecting certain key words that represented possible areas of conflict, word association techniques could be employed for psychodiagnostic purposes. Jung's experiments served as an inspiration to creators of such tests as the Word Association Test developed by Rapaport, Gill, and Schafer (1945–1946) at the Menninger Clinic. This test consisted of three parts. In the first part, each stimulus word was administered to the examinee, who had been instructed to respond quickly with the first word that came to mind. The examiner recorded the length of time it took the subject to respond to each item. In the second part of the test, each stimulus word was again presented to the examinee. The examinee was instructed to reproduce the original responses. Any deviation between the original and this second response was recorded, as was the length of time before reacting. The third part of the test was the inquiry. Here the examiner asked questions to clarify the relationship that existed between the stimulus word and the response (for example, "What were you thinking about?" or "What was going through your mind?"). In some cases, the relationship may have been obvious; in others, however, the relationship between the two words may have been extremely idiosyncratic or even bizarre.

The test consisted of 60 words, some considered neutral by the test authors (for example, *chair, book, water, dance, taxi*) and some characterized as *traumatic.* In the latter category were "words that are likely to touch upon sensitive personal material according to clinical experience, and also words that attract associative disturbances" (Rapaport et al., 1968, p. 257). Examples of words so designated were *love, girlfriend, boyfriend, mother, father, suicide, fire, breast,* and *masturbation.*

Responses on the Word Association Test were evaluated with respect to variables such as popularity, reaction time, content, and test-retest responses. Normative data were provided regarding the percentage of occurrence of certain responses for college students and schizophrenic groups. For example, to the word *stomach,* 21% of the college group responded with "ache" and 13% with "ulcer." Ten percent of the schizophrenic group responded with "ulcer." To the word *mouth,* 20% of the college sample responded with "kiss," 13% with "nose," 11% with "tongue," 11% with "lips," and 11% with "eat." In the schizophrenic group, 19% responded with "teeth," and 10% responded with "eat." The test does not enjoy widespread clinical use today but is more apt to be found in the occasional research application.

JUST THINK . . .

As compared to the 1940s, how emotion-arousing do you think the "traumatic" stimuli on the Word Association Test are by contemporary standards? Why?

The Kent-Rosanoff Free Association Test (Kent & Rosanoff, 1910) represented one of the earliest attempts to develop a standardized test using words as projective stimuli.[5] The test consisted of 100 stimulus words, all commonly used and believed to be neutral with respect to emotional impact. The standardization sample consisted of 1,000 normal adults who varied in geographic location, educational level, occupation, age, and intellectual capacity. Frequency tables based on the responses of these 1,000 cases were developed. These tables were used to evaluate examinees' responses according to the clinical judgment of psychopathology. Psychiatric patients were found to have a lower frequency of popular responses than the normal subjects in

5. The term **free association** refers to the technique of having subjects relate all their thoughts as they are occurring and is most frequently used in psychoanalysis; the only structure imposed is provided by the subjects themselves. The technique employed in the Kent-Rosanoff is that of **word association** (not free association), in which the examinee relates the first word that comes to mind in response to a stimulus word. The term *free association* in the test's title is, therefore, a misnomer.

the standardization group. However, as it became apparent that the individuality of responses may be influenced by many variables other than psychopathology (such as creativity, age, education, and socioeconomic factors), the popularity of the Kent-Rosanoff as a differential diagnostic instrument diminished. Damaging, too, was research indicating that scores on the Kent-Rosanoff were unrelated to other measures of psychotic thought (Ward et al., 1991). Still, the test endures as a standardized instrument of word association responses and, more than ninety years after its publication, continues to be used in experimental research and clinical practice.

JUST THINK . . .

Quick . . . the first thought that comes into your mind. . . . Ready? *Word association.*

Sentence completion tests Other projective techniques that use verbal material as projective stimuli are sentence completion tests. How might you complete the following sentences?

I like to _____.

Someday, I will _____.

I will always remember the time _____.

I worry about _____.

I am most frightened when _____.

My feelings are hurt _____.

My mother _____.

I wish my parents _____.

Sentence completion tests may contain items that, like those listed here, are quite general and appropriate for administration in a wide variety of settings. Alternatively, **sentence completion stems** (the first part of the item) may be developed for use in specific types of settings (such as school or business) or for specific purposes. Sentence completion tests may be relatively atheoretical or linked very closely to some theory. As an example of the latter, the Washington University Sentence Completion Test (Loevinger et al., 1970) was based on the writings of Loevinger and her colleagues in the area of ego development.

Loevinger (1966; Loevinger & Ossorio, 1958) believes that maturity brings a transformation of one's self-image from an essentially stereotypic and socially acceptable one to a more personalized and realistic one. The Washington University Sentence Completion Test was constructed to assess self-concept according to Loevinger's theory. Some evidence for the validity of this test comes from its ability to predict social attitudes in a manner consistent with Loevinger's theory (Browning, 1987). It is possible to obtain other traditional psychometric indices on this test. For example, inter-rater reliability for this test has been estimated to range from .74 to .88, internal consistency is in the high .80s, and test-retest reliability ranges from .67 to .76 or from .88 to .92, depending upon how the test is scored (Weiss et al., 1989).

A number of standardized sentence completion tests are available to the clinician. One such test, the Rotter[6] Incomplete Sentences Blank (Rotter & Rafferty, 1950) is the most popular of all. The Rotter was developed for use with populations from grade 9 through adulthood and is available in three levels: high school (grades 9 through 12), college (grades 13 through 16), and adult. Testtakers are instructed to respond to each of the 40 incomplete sentence items in a way that expresses their "real feelings." The manual suggests that responses on the test be interpreted according to several

6. The *o* in *Rotter* is long, as in *rote*.

categories: family attitudes, social and sexual attitudes, general attitudes, and character traits. Each response is evaluated on a seven-point scale that ranges from *need for therapy* to *extremely good adjustment.*

The manual contains normative data for a sample of 85 female and 214 male college freshmen but no norms for high-school and adult populations. Also presented in the test manual are sample responses of several subjects along with background information about the subjects. According to the psychometric studies quoted in the test manual, the Rotter is a reliable and valid instrument. Estimates of inter-scorer reliability were reported to be in the .90s. Independently from the original validity studies, sociometric techniques have been used to demonstrate the validity of the Rotter as a measure of adjustment (Lah, 1989).

In general, a sentence completion test may be useful for obtaining diverse information about an individual's interests, educational aspirations, future goals, fears, conflicts, needs, and so forth. The tests have a high degree of face validity. However, with this high degree of face validity comes a certain degree of transparency about the objective of the test. For this reason, sentence completion tests are perhaps the most vulnerable of all the projective methods to faking on the part of an examinee intent on making a good—or a bad—impression.

Sounds as Projective Stimuli

Let's state at the outset that this section is included more as a fascinating footnote in the history of projectives than as a description of widely used tests. The history of the use of sound as a projective stimulus is fascinating because of its origins in the laboratory of a then-junior fellow of Harvard University. You may be surprised to learn that it was a behaviorist whose name has seldom been uttered in the same sentence as the term *projective test* by any contemporary psychologist: B. F. Skinner (Figure 13–8). The device was something "like auditory inkblots" (Skinner, 1979, p. 175).

The time was the mid-1930s. Skinner's colleagues, Henry Murray and Christiana Morgan, were working on the TAT in the Harvard Psychological Clinic. Psychoanalytic theory was very much in vogue. Even behaviorists were curious about Freud's approach, and some were even undergoing psychoanalysis themselves. Switching on the equipment in his laboratory in the biology building, the rhythmic noise served as a stimulus for Skinner to create words that went along with it. This inspired Skinner to think of an application for sound, not only in behavioral terms but in the elicitation of "latent" verbal behavior that was significant "in the Freudian sense" (Skinner, 1979, p. 175). Skinner created a series of recorded sounds much like muffled, spoken vowels, to which people would be instructed to associate. The sounds, packaged as a device he called a *verbal summator,* presumably would act as a stimulus for the person to verbalize certain unconscious material. Henry Murray, by the way, liked the idea and supplied Skinner with a room at the clinic in which to test subjects. Saul Rosenzweig also liked the idea; he and David Shakow renamed the instrument the *tautophone* (from the Greek *tauto,* meaning "repeating the same") and did research with it (Rutherford, 2003). Their instructions to subjects were as follows:

> Here is a phonograph. On it is a record of a man's voice saying different things. He speaks rather unclearly, so I'll play over what he says a number of times. You'll have to listen carefully. As soon as you have some idea of what he's saying, tell me at once. (Shakow & Rosenzweig, 1940, p. 217)

As recounted in detail by Rutherford (2003), there was little compelling evidence to show that the instrument could differentiate between members of clinical and nonclinical groups. Still, a number of other auditory projective techniques were developed. There

Figure 13–8
Projective Test Pioneer B. F. Skinner . . . *What?!*

Working at the Harvard Psychological Clinic with the blessing of (and even some financial support from) Henry Murray, B. F. Skinner (who today is an icon of behaviorism) evinced great enthusiasm for an auditory projective test he had developed. He believed the technique had potential as "a device for snaring out complexes" (Skinner, 1979, p. 176). A number of well-known psychologists of the day apparently agreed. For example, Joseph Zubin, in correspondence with Skinner, wrote that Skinner's technique had promise "as a means for throwing light on the less objective aspects of the Rorschach experiment" (Zubin, 1939). Of course, if the test really had that much promise, Skinner would probably be getting equal billing in this chapter with Murray and Rorschach.

was the Auditory Apperception Test (Stone, 1950), in which the subject's task was to respond by creating a story based on three sounds played on a phonograph record. Other researchers produced similar tests, one called an auditory sound association test (Wilmer & Husni, 1951) and the other referred to as an auditory apperception test (Ball & Bernardoni, 1953). Henry Murray also got into the act with his Azzageddi test (Davids & Murray, 1955), named for a Herman Melville character. Unlike other auditory projectives, the Azzageddi presented subjects with spoken paragraphs.

So why aren't test publishers today punching out CDs with projective sounds at a pace to match the publication of inkblots and pictures? Rutherford (2003) speculated that a combination of factors conspired to cause the demise of auditory projective methods. The tests proved not to differentiate between different groups of subjects who took it. Responses to the auditory stimuli lacked the complexity and richness of responses to inkblots, pictures, and other projective stimuli. None of the available scoring systems was very satisfactory. Except for use with the blind, auditory projective tests were seen as redundant and not as good as the TAT.

The Production of Figure Drawings

A relatively quick, easily administered projective technique is the analysis of drawings. Drawings can provide the psychodiagnostician with a wealth of clinical hypotheses to be confirmed or discarded as the result of other findings (Figure 13–9). The use of drawings in clinical and research settings has extended beyond the area of personality assessment. Attempts have been made to use artistic productions as a source of information about intelligence, neurological intactness, visual-motor coordination, cognitive development, and even learning disabilities (Neale & Rosal, 1993). Figure drawings are an appealing source of diagnostic data because the instructions for them can be administered individually or in a group by nonclinicians such as teachers, and no materials other than a pencil and paper are required.

Drawing by a 25-year-old schoolteacher after becoming engaged. Previously, she had entered psychotherapy because of problems relating to men and a block against getting married. The position of the hands was interpreted as indicating a fear of sexual intercourse.

Drawing by a male with a "Don Juan" complex—a man who pursued one affair after another. The collar pulled up to guard the neck and the excessive shading of the buttocks suggests a fear of being attacked from the rear. It is possible that this man's Don Juanism is an outward defense against a lack of masculinity—even feelings of effeminacy—with which he may be struggling inside.

Drawing by an authoritarian and sadistic male who had been head disciplinarian of a reformatory for boys before he was suspended for child abuse. His description of this picture was that it "looked like a Prussian or a Nazi general."

The manacled hands, tied feet, exposed buttocks, and large foot drawn to the side of the drawing taken together are reflective, according to Hammer, of masochistic, homosexual, and exhibitionistic needs.

This drawing by an acutely paranoid, psychotic man was described by Hammer (1981, p. 170) as follows: "The savage mouth expresses the rage-filled projections loose within him. The emphasized eyes and ears with the eyes almost emanating magical rays reflect the visual and auditory hallucinations the patient actually experiences. The snake in the stomach points up his delusion of a reptile within, eating away and generating venom and evil."

Figure 13–9
Some Sample Interpretations Made from Figure Drawings

Source: Hammer (1981)

Figure-drawing tests In general, a **figure drawing test** may be defined as a projective method of personality assessment whereby the assessee produces a drawing that is analyzed on the basis of its content and related variables. The classic work on the use of figure drawings as a projective stimulus is a book entitled *Personality Projection in the Drawing of the Human Figure* by Karen Machover (1949). Machover wrote that

> the human figure drawn by an individual who is directed to "draw a person" [is] related intimately to the impulses, anxieties, conflicts, and compensations characteristic of that individual. In some sense, the figure drawn is the person, and the paper corresponds to the environment. (p. 35)

The instructions for administering the Draw A Person (DAP) test are quite straightforward. The examinee is given a pencil and a blank sheet of 8½-by-11-inch white paper and told to draw a person. Inquiries on the part of the examinee concerning how the picture is to be drawn are met with statements such as "Make it the way you think it should be" or "Do the best you can." Immediately after the first drawing is completed, the examinee is handed a second sheet of paper and instructed to draw a picture of a person of the sex opposite that of the person just drawn.[7] Subsequently, many clinicians will ask questions about the drawings, such as "Tell me a story about that figure," "Tell me about that boy/girl, man/lady," "What is the person doing?" "How is the person feeling?" "What is nice or not nice about the person?" Responses to these questions are used in forming various hypotheses and interpretations about personality functioning.

Traditionally, DAP productions have been formally evaluated through analysis of various characteristics of the drawing. Attention has been given to such factors as the length of time required to complete the picture, placement of the figures, the size of the figure, pencil pressure used, symmetry, line quality, shading, the presence of erasures, facial expressions, posture, clothing, and overall appearance. Various hypotheses have been generated based on these factors (Knoff, 1990). For example, the *placement* of the figure on the paper is seen as representing how the individual functions within the environment. The person who draws a tiny figure at the bottom of the paper might have a poor self-concept or might be insecure or depressed. The individual who draws a picture that cannot be contained on one sheet of paper and goes off the page is considered to be impulsive. Unusually light pressure suggests character disturbance (Exner, 1962). According to Buck (1948, 1950), placement of drawing on the right of the page suggests orientation to the future; placement to the left suggests an orientation to the past. Placement at the upper right suggests a desire to suppress an unpleasant past as well as excessive optimism about the future. Placement to the lower left suggests depression with a desire to flee into the past.

Another variable of interest to those who analyze figure drawings is the *characteristics* of the individual drawn. For example, unusually large eyes or large ears suggest suspiciousness, ideas of reference, or other paranoid characteristics (Machover, 1949; Shneidman, 1958). Unusually large breasts drawn by a male may be interpreted as unresolved oedipal problems with maternal dependence (Jolles, 1952). Long and conspicuous ties suggest sexual aggressiveness, perhaps overcompensating for fear of impotence (Machover, 1949). Button emphasis suggests dependent, infantile, inadequate personality (Halpern, 1958).

JUST THINK . . .

Draw a person. Contemplate what that drawing tells about you.

The House-Tree-Person test (HTP; Buck, 1948) is another projective figure-drawing test. As the name of the test implies, the testtaker's task is to draw a picture of a house, a tree, and a person. In much the same way that different aspects of the human figure are presumed to be reflective of psychological functioning, the ways in which

7. When instructed simply to "draw a person," most people will draw a person of the same sex, so it is deemed clinically significant if the assessee draws a person of the opposite sex when given this instruction. Rierdan and Koff (1981) found that, in some cases, children are uncertain of the sex of the figure drawn. They hypothesized that in such cases "the child has an indefinite or ill-defined notion of sexual identity" (p. 257).

an individual represents a house and a tree are considered symbolically significant. Another test, this one thought to be of particular value in learning about the examinee in relation to her or his family, is the Kinetic Family Drawing (KFD). Derived from Hulse's (1951, 1952) Family Drawing Test, an administration of the KFD (Burns & Kaufman, 1970, 1972) begins with the presentation of an 8½-by-11-inch sheet of paper and a pencil with an eraser. The examinee, usually though not necessarily a child, is instructed as follows:

> Draw a picture of everyone in your family, including you, DOING something. Try to draw whole people, not cartoons or stick people. Remember, make everyone DOING something—some kind of actions. (Burns & Kaufman, 1972, p. 5)

In addition to yielding graphic representations of each family member for analysis, this procedure may yield a wealth of information in the form of examinee verbalizations while the drawing is being executed. After the examinee has completed the drawing, a rather detailed inquiry follows. The examinee is asked to identify each of the figures, talk about their relationship, and detail what they are doing in the picture and why. A number of formal scoring systems for the KFD are available. Related techniques include a school adaptation called the Kinetic School Drawing (KSD; Prout & Phillips, 1974); a test that combines aspects of the KFD and the KSD called the Kinetic Drawing System (KDS; Knoff & Prout, 1985); and the Collaborative Drawing Technique (D. K. Smith, 1985), a test that provides an occasion for family members to collaborate on the creation of a drawing—presumably all the better to "draw together."

The Draw A Person: Screening Procedure for Emotional Disturbance (DAP:SPED; Naglieri et al., 1991) features a standardized test administration and quantitative scoring system designed to screen testtakers (ages 6–17) for emotional problems. Based on the assumption that the rendering of unusual features in figure drawings signals emotional problems, one point is scored for each such feature. With age and normative information taken into account, high scores signal the need for more detailed evaluation. Validity data are presented in the test manual, but both an independent evaluation of the test (Motta et al., 1993a, 1993b) and a study by two of the test's authors (McNeish & Naglieri, 1993) raised concerns about the number of misidentifications (both false positives and false negatives) that might result from the test's use even as a screening tool.

Like other projective techniques, figure-drawing tests, although thought to be clinically useful, have had a rather embattled history in relation to their psychometric soundness (Joiner & Schmidt, 1997). In general, the techniques are vulnerable with regard to the assumptions that drawings are essentially self-representations (Tharinger & Stark, 1990) and represent something far more than drawing ability (Swensen, 1968). Although a number of systems have been devised to score figure drawings, solid support for the validity of such approaches has been elusive (Watson et al., 1967). Experience and expertise do not necessarily correlate with greater clinical accuracy in drawing interpretation. Karen Machover (cited in Watson, 1967) herself reportedly had "grave misgivings" (p. 145) about the misuse of her test for diagnostic purposes.

To be sure, the clinical use of figure drawings has its academic defenders (Riethmiller & Handler, 1997a, 1997b). Waehler (1997), for example, cautioned that tests are not foolproof and that a person who comes across as rife with pathology in an interview might well seem benign on a psychological test. He went on to advise that figure drawings "can be considered more than 'tests'; they involve tasks that can also serve as stepping-off points for clients and examiners to discuss and clarify the picture" (p. 486).

Just before taking a step back and reviewing projective methods in perspective, let's mention efforts to combine inkblot methodology with storytelling and drawing to come

MEET AN ASSESSMENT PROFESSIONAL

Meet Dr. Tonia Caselman

When I meet with children and their parents I always complete a genogram. A **genogram** is a graphic presentation of a person's family relationships. In some ways it is a therapist's version of a family tree. However, it displays more information than a simple listing of family members. It is used to identify themes and patterns of behavior in a family history. It assists both the client and the therapist in quickly seeing the impact of the family environment. It also is a convenient and effective communication tool. In completing a genogram, I like to include at least three–four generations.

I have found that by using genograms clients feel more comfortable talking about sensitive family dynamics, problems, and relationships. By using joint attention on the drawing, it relieves the tension that clients often feel by too much intense eye contact. This is particularly true of clients from cultures where less eye contact is considered respectful. A genogram is also a useful instrument for joining with or engaging a family in treatment. It is a more relaxed, informal way to gather information.

Tonia Caselman, Ph.D., School of Social Work, University of Oklahoma

Read more of what Dr. Caselman had to say—her complete essay—at www.mhhe.com/cohentesting7.

up with a therapeutic method called "Color Inkblot Therapeutic Storytelling" (Sakaki et al., 2007). Originally developed as a culturally responsive assessment method for use with Japanese clients, this therapeutic assessment method is designed to provide an indirect and nonthreatening approach to clients' problems. Tangentially, our guest test user also uses an indirect and nonthreatening approach in her efforts at therapeutic assessment (see *Meet an Assessment Professional*).

Projective Methods in Perspective

Used enthusiastically by many clinicians and criticized harshly by many academics, projective methods continue to occupy a rather unique habitat in the psychological landscape. Lilienfeld et al. (2000) raised serious questions regarding whether that habitat is worth maintaining. These authors focused their criticism on scoring systems for the Rorschach, the TAT, and figure drawings. They concluded that there was empirical support for only a relatively small number of Rorschach and TAT indices. They found even fewer compelling reasons to justify the continued use of figure drawings. Some of their assertions with regard to the Rorschach and the TAT—as well as the response of a projective test user and advocate, Stephen Hibbard (2003)—are presented in Table 13–4. Hibbard commented only on the Rorschach and the TAT because of his greater experience with these tests as opposed to figure drawings.

Table 13–4
The Cons and Pros (or Cons Rebutted) of Projective Methods

Lilienfeld et al. (2000) on the Cons	Hibbard (2003) in Rebuttal
Projective techniques tend not to provide incremental validity above more structured measures, as is the argument of proponents of the projective hypothesis as stated by Dosajh (1996).	Lilienfeld et al. presented an outmoded caricature of projection and then proceeded to attack it. Dosajh has not published on any of the coding systems targeted for criticism. None of the authors who developed coding systems that were attacked espouse a view of projection similar to Dosajh's. Some of the criticized authors have even positioned their systems as nonprojective.
The norms for Exner's Comprehensive System (CS) are in error. They may overpathologize normal individuals and may even harm clients.	Evidence is inconclusive as to error in the norms. Observed discrepancies may have many explanations. Overpathologization may be a result of "drift" similar to that observed in the measurement of intelligence (Flynn effect).
There is limited support for the generalizability of the CS across different cultures.	More cross-cultural studies do need to be done, but the same could be said for most major tests.
Four studies are cited to support the deficiency of the test-retest reliability of the CS.	Only three of the four studies cited are in *refereed journals* (for which submitted manuscripts undergo critical review and may be selected or rejected for publication), and none of these three studies are bona fide test-retest reliability studies.
With regard to the TAT, there is no point in aggregating scores into a scale in the absence of applying internal consistency reliability criteria.	This assertion is incorrect because "each subunit of an aggregated group of predictors of a construct could be unrelated to the other, but when found in combination, they might well predict important variance in the construct" (p. 264).
TAT test-retest reliability estimates have been "notoriously problematic" (p. 41).	"… higher retest reliability would accrue to motive measures if the retest instructions permitted participants to tell stories with the same content as previously" (p. 265).
Various validity studies with different TAT scoring systems can be faulted on methodological grounds.	Lilienfeld et al. (2000) misinterpreted some studies they cited and did not cite other studies. For example, a number of relevant validity studies in support of Cramer's (1991) Defense Mechanism Manual coding system for the TAT were not cited.

Note: Interested readers are encouraged to read the full text of Lilienfeld et al. (2000) and Hibbard (2003), as the arguments made by each are far more detailed than the brief samples presented here.

In general, critics have attacked projective methods on grounds related to the *assumptions* inherent in their use, the *situational variables* that attend their use, and several *psychometric considerations*—most notably, a paucity of data to support their reliability and validity.

Assumptions Murstein (1961) examined ten assumptions of projective techniques and argued that none of them was scientifically compelling. Several assumptions concern the stimulus material. For example, it is assumed that the more ambiguous the stimuli, the more subjects reveal about their personality. However, Murstein describes the stimulus material as only one aspect of the "total stimulus situation." Environmental variables, response style, reactions to the examiner, and related factors all contribute to response patterns. In addition, in situations where the stimulus properties of the projective material were designed to be unclear or hazy or are presented with uncompleted lines—thereby increasing ambiguity—projection on the part of the subject was not found to increase.

Another assumption concerns the supposedly idiosyncratic nature of the responses evoked by projective stimuli. In fact, similarities in the response themes of different subjects to the same stimulus cards suggest that the stimulus material may not be as ambiguous and amenable to projection as previously assumed. Some consideration of the stimulus properties and the ways they affect the subject's

JUST THINK . . .

Suppose a Rorschach card or a TAT card elicited much the same response from *most* people. Would that be an argument for or against the use of the card?

responses is therefore indicated. The assumption that projection is greater onto stimulus material that is similar to the subject (in physical appearance, gender, occupation, and so on) has also been found questionable.

Murstein (1961) also raised questions about the way projective tests are interpreted. He questioned numerous assumptions, including the assumptions that:

- every response provides meaning for personality analysis
- a relationship exists between the strength of a need and its manifestation on projective instruments
- testtakers are unaware of what they are disclosing about themselves
- a projective protocol reflects sufficient data concerning personality functioning for formulation of judgments
- there is a parallel between behavior obtained on a projective instrument and behavior displayed in social situations

Murstein dismissed such contentions as "cherished beliefs" that are accepted "without the support of sufficient research validation" (p. 343). Still, proponents of projectives continue to be convinced, for example, that the ambiguous nature of a task such as inkblot interpretation make for test results that are less subject (as compared to nonprojective tasks) to faking, especially "faking good," on the part of testtakers. This assumption is evident in the writings of advocates for the use of the Rorschach in forensic and related applications (see, for example, Gacono et al., 2008). So, for example, Weiss et al. (2008) listed, among the compelling reasons to use the test in pre-employment screening of police personnel, the test's utility in bypassing "volitional controls." Supporting the assumption that the Rorschach test frustrates testtakers' efforts to fake good or manage favorable impressions is a study conducted in China (Cai & Shen, 2007). These researchers found differences in the self-concept of 61 college students as obtained through Rorschach protocols and scores on the Tennessee Self-Concept Scale. The Rorschach was seen as a superior measure of self-concept because respondents were unable to manage favorable impressions.

In a study that compared Rorschach responses of sex-offending Roman Catholic clergy to a control group of non-offending clergy, the offenders clearly exhibited higher distortion in thinking styles (Ryan et al., 2008). Although such studies could be cited to support assumptions in Rorschach use relevant to impression management, a number of studies focusing directly on this issue have yielded mixed results ranging from supportive to equivocal (Conti, 2007; Fahs, 2004; Gregg, 1998; Whittington, 1998; Yell, 2008). At the very least, it can be observed that as a measurement method, the Rorschach provides a stimulus that is less susceptible than others to learned, rehearsed, and/or socially conventional responding. It may also be useful in obtaining insights into the respondent's unique way of perceiving and organizing novel stimuli.

To Murstein's list of questionable assumptions underlying the use of projective tests we might add one that is basic to projective assessment: something called "the unconscious" exists. Though the term *unconscious* is widely used as if its existence were a given, some academicians have questioned whether in fact the unconscious exists in the same way that, say, the liver exists. The scientific studies typically cited to support the existence of the unconscious (or, perhaps more accurately, the efficacy of the construct *unconscious*) have used a wide array of methodologies; see, for example, Diven (1937), Erdelyi (1974), Greenspoon (1955), and Razran (1961). The conclusions of each of these types of studies are subject to alternative explanations. Also subject to alternative explanation are conclusions about the existence of the unconscious based on experimental

testing of predictions derived from hypnotic phenomena, from signal detection theory, and from specific personality theories (Brody, 1972). More generally, many interpretive systems for the Rorschach and other projective instruments are based on psychodynamic theory, which itself has no shortage of critics.

Situational variables Proponents of projective techniques have claimed that such tests are capable of illuminating the mind's recesses much like X-rays illuminate the body. Frank (1939) conceptualized projective tests as tapping personality patterns without disturbing the pattern being tapped. If that were true, then variables related to the test situation should have no effect on the data obtained. However, situational variables such as the examiner's presence or absence have significantly affected the responses of experimental subjects. For example, TAT stories written in private are likely to be less guarded, less optimistic, and more affectively involved than those written in the presence of the examiner (Bernstein, 1956). The age of the examiner is likely to affect projective protocols (Mussen & Scodel, 1955), as are the specific instructions (Henry & Rotter, 1956) and the subtle reinforcement cues provided by the examiner (Wickes, 1956).

Masling (1960) reviewed the literature on the influence of situational and interpersonal variables in projective testing and concluded that there was strong evidence for a role of situational and interpersonal influences in projection. Masling concluded that subjects utilized every available cue in the testing situation, including cues related to the actions or the appearance of the examiner. Moreover, Masling argued that examiners also relied on situational cues, in some instances over and above what they were taught. Examiners appeared to interpret projective data with regard to their own needs and expectations, their own subjective feelings about the person being tested, and their own constructions regarding the total test situation. Masling (1965) experimentally demonstrated that Rorschach examiners—through postural, gestural, and facial cues—are capable of unwittingly eliciting the responses they expect.

In any given clinical situation, many variables may be placed in the mix. The interaction of these variables may influence clinical judgments. So it is that research has suggested that even in situations involving objective (not projective) tests or simple history taking, the effect of the clinician's training (Chapman & Chapman, 1967; Fitzgibbons & Shearn, 1972) and role perspective (Snyder et al., 1976) as well as the patient's social class (Hollingshead & Redlich, 1958; Lee, 1968; Routh & King, 1972) and motivation to manage a desired impression (Edwards & Walsh, 1964; Wilcox & Krasnoff, 1967) are capable of influencing ratings of pathology (Langer & Abelson, 1974) and related conclusions (Batson, 1975). These and other variables are given wider latitude in the projective test situation, where the examiner may be at liberty to choose not only the test and extra-test data on which interpretation will be focused but also the scoring system that will be used to arrive at that interpretation.

Psychometric considerations The psychometric soundness of many widely used projective instruments has yet to be demonstrated. Critics of projective techniques have called attention to variables such as uncontrolled variations in protocol length, inappropriate subject samples, inadequate control groups, and poor external criteria as factors contributing to spuriously increased ratings of validity. There are methodological obstacles in researching projectives because many test-retest or split-half methods are inappropriate. It is, to say the least, a challenge to design and execute validity studies that effectively rule out, limit, or

JUST THINK . . .

Projective tests have been around for a long time because of their appeal to many clinicians. What are their advantages? Why should they be around for a long time to come?

statistically take into account the unique situational variables that attend the administration of such tests.

The debate between academicians who argue that projective tests are not technically sound instruments and clinicians who find such tests useful has been raging ever since projectives came into widespread use. Frank (1939) responded to those who would reject projective methods because of their lack of technical rigor:

> These leads to the study of personality have been rejected by many psychologists because they do not meet psychometric requirements for validity and reliability, but they are being employed in association with clinical and other studies of personality where they are finding increasing validation in the consistency of results for the same subject when independently assayed by each of these procedures. . . .
>
> If we face the problem of personality, in its full complexity, as an active dynamic process to be studied as a *process* rather than as entity or aggregate of traits, factors, or as static organization, then these projective methods offer many advantages for obtaining data on the process of organizing experience which is peculiar to each personality and has a life career. (Frank, 1939, p. 408; emphasis in the original)

Behavioral Assessment Methods

Traits, states, motives, needs, drives, defenses, and related psychological constructs have no tangible existence. They are constructs whose existence must be inferred from behavior. In the traditional approach to clinical assessment, tests as well as other tools are employed to gather data. From these data, diagnoses and inferences are made concerning the existence and strength of psychological constructs. The traditional approach to assessment might therefore be labeled a *sign* approach because test responses are deemed to be signs or clues to underlying personality or ability. In contrast to this traditional approach is an alternative philosophy of assessment that may be termed the *sample* approach. The sample approach focuses on the behavior itself. Emitted behavior is viewed not as a sign of something but rather as a sample to be interpreted in its own right.

The emphasis in **behavioral assessment** is on "what a person *does* in situations rather than on inferences about what attributes he *has* more globally" (Mischel, 1968, p. 10). Predicting what a person will do is thought to entail an understanding of the assessee with respect to both antecedent conditions and consequences of a particular situation (Smith & Iwata, 1997). Upon close scrutiny, however, the trait concept is still present in many behavioral measures, though more narrowly defined and more closely linked to specific situations (Zuckerman, 1979).

To illustrate behavioral observation as an assessment strategy, consider the plight of the single female client who presents herself at the university counseling center. She complains that even though all her friends tell her how attractive she is, she has great difficulty meeting men—so much so that she doesn't even want to try anymore. A counselor confronted with such a client might, among other things, (1) interview the client about this problem, (2) administer an appropriate test to the client, (3) ask the client to keep a detailed diary of her thoughts and behaviors related to various aspects of her efforts to meet men, including her expectations, and (4) accompany the client on a typical night out to a singles bar or similar venue and observe her behavior. The latter two strategies come under the heading of behavioral observation. With regard to the diary, the client is engaging in self-observation. In the scenario of the night out, the counselor is doing the actual observation.

The more traditional administration of a psychological test or test battery to a client such as this single woman might yield signs that then could be inferred to relate to the problem. For example, if a number of the client's TAT stories involved themes of demeaning, hostile, or otherwise unsatisfactory heterosexual encounters as a result of venturing out into the street, a counselor might make an interpretation at a deeper or second level of inference. For example, a counselor, especially one with a psychoanalytic orientation, might reach a conclusion something like this:

> The client's expressed fear of going outdoors, and ultimately her fear of meeting men, might in some way be related to an unconscious fear of promiscuity—a fear of becoming a streetwalker.

Such a conclusion in turn would have implications for treatment. Many hours of treatment might be devoted to uncovering the "real" fear so that it is apparent to the client herself and ultimately dealt with effectively.

In contrast to the sign approach, the clinician employing the sample or behavioral approach to assessment might examine the behavioral diary that the client kept and design an appropriate therapy program on the basis of those records. Thus, for example, the antecedent conditions under which the client would feel most distraught and unmotivated to do anything about the problem might be delineated and worked on in counseling sessions.

An advantage of the sign approach over the sample approach is that—in the hands of a skillful, perceptive clinician—the client might be put in touch with feelings that even she was not really aware of before the assessment. The client may have been consciously (or unconsciously) avoiding certain thoughts and images (those attendant on the expression of her sexuality, for example), and this inability to deal with those thoughts and images may indeed have been a factor contributing to her ambivalence about meeting men.

Behavioral assessors seldom make such deeper-level inferences. For example, if sexuality is not raised as an area of difficulty by the client (in an interview, a diary, a checklist, or by some other behavioral assessment technique), this problem area may well be ignored or given short shrift. Behavioral assessors do, however, tend to be more empirical in their approach, as they systematically assess the client's presenting problem both from the perspective of the client and from the perspective of one observing the client in social situations and the environment in general. The behavioral assessor does not search the Rorschach or other protocols for clues to treatment. Rather, the behaviorally oriented counselor or clinician relies much more on what the client *does* and *has done* for guideposts to treatment. In a sense, the behavioral approach does not require as much clinical creativity as the sign approach. Perhaps for that reason, the behavioral approach may be considered less an art than a science (at least as compared to some other clinical approaches). It is certainly science-based in that it relies on relatively precise methods of proven validity (Haynes & Kaholokula, 2008).

Early on, the shift away from traditional psychological tests by behaviorally oriented clinicians compelled some to call for a way to integrate such tests in behavioral evaluations. This view is typified by the wish that "psychological tests should be able to provide the behavior therapist with information that should be of value in doing behavior therapy. This contention is based on the assumption that the behavior on any psychological test should be lawful" (Greenspoon & Gersten, 1967, p. 849). Accordingly, psychological tests could be useful, for example, in helping the behavior therapist identify the kinds of contingent stimuli that would be most effective

JUST THINK . . .

Is there a way to integrate traditional psychological testing and assessment and behavioral assessment?

with a given patient. For example, patients with high percentages of color or color/form responses on the Rorschach and with IQs over 90 might be most responsive to positive verbal contingencies (such as *good, excellent,* and so forth), whereas patients with high percentages of movement or vista (three-dimensional) responses on the Rorschach and IQs over 90 might be most responsive to negative verbal contingencies (such as *no* or *wrong*). Such innovative efforts to narrow a widening schism in the field of clinical assessment have failed to ignite experimental enthusiasm, perhaps because more direct ways exist to assess responsiveness to various contingencies.

Differences between traditional and behavioral approaches to assessment have to do with varying assumptions about the nature of personality and the causes of behavior. The data from traditional assessment are used primarily to describe, classify, or diagnose, whereas the data from a behavioral assessment are typically more directly related to the formulation of a specific treatment program. Some of the other differences between the two approaches are summarized in Table 13–5.

The Who, What, When, Where, Why, and How of It

The name says it all: *Behavior* is the focus of assessment in behavioral assessment—not traits, states, or other constructs presumed to be present in various strengths—just behavior. This will become clear as we survey the *who, what, when, where, why,* and *how* of behavioral assessment.

Who? *Who* is assessed? The person being assessed may be, for example, a patient on a closed psychiatric ward, a client seeking help at a counseling center, or a subject in an academic experiment. Regardless of whether the assessment is for research, clinical, or other purposes, the hallmark of behavioral assessment is intensive study of individuals. This is in contrast to mass testing of groups of people to obtain normative data with respect to some hypothesized trait or state.

Who is the assessor? Depending on the circumstances, the assessor may be a highly qualified professional or a technician/assistant trained to conduct a particular assessment. Technicians are frequently employed to record the number of times a targeted behavior is exhibited. In this context, the assessor may also be a classroom teacher recording, for example, the number of times a child leaves her or his seat. An assessor in behavioral assessment may also be the assessee. Assessees are frequently directed to maintain behavioral diaries, complete behavioral checklists, or engage in other activities designed to monitor their own behavior.

What? *What* is measured in behavioral assessment? Perhaps not surprisingly, the behavior or behaviors targeted for assessment will vary as a function of the objectives of the assessment. What constitutes a targeted behavior will typically be described in sufficient detail prior to any assessment. For the purposes of assessment, the targeted behavior must be measurable—that is, quantifiable in some way. Examples of such measurable behaviors can range from the number of seconds elapsed before a child calls out in class to the number of degrees body temperature is altered. Note that descriptions of targeted behaviors in behavioral assessment typically begin with the phrase *the number of.* In studies that focus on physiological variables such as muscle tension or autonomic responding, special equipment is required to obtain the behavioral measurements (see, for example, Conrad et al., 2008).

When? *When* is an assessment of behavior made? One response to this question is that assessment of behavior is typically made at times when the problem behavior is most

Table 13–5
Differences between Behavioral and Traditional Approaches to Psychological Assessment

	Behavioral	Traditional
Assumptions		
Conception of personality	Personality constructs mainly employed to summarize specific behavior patterns, if at all	Personality as a reflection of enduring, underlying states or traits
Causes of behavior	Maintaining conditions sought in current environment	Intrapsychic, or within the individual
Implications		
Role of behavior	Important as a sample of person's repertoire in specific situation	Behavior assumes importance only insofar as it indexes underlying causes
Role of history	Relatively unimportant except, for example, to provide a retrospective baseline	Crucial in that present conditions seen as products of the past
Consistency of behavior	Behavior thought to be specific to the situation	Behavior expected to be consistent across time and settings
Uses of data	To describe target behaviors and maintain conditions To select the appropriate treatment To evaluate and revise treatment	To describe personality functioning and etiology To diagnose or classify To make prognosis; to predict
Other characteristics		
Level of inferences	Low	Medium to high
Comparisons	More emphasis on intraindividual or idiographic	More emphasis on interindividual or nomothetic
Methods of assessment	More emphasis on direct methods (e.g., observations of behavior in natural environment)	More emphasis on indirect methods (e.g., interviews and self-report)
Timing of assessment	More ongoing; prior, during, and after treatment	Pre- and perhaps posttreatment, or strictly to diagnose
Scope of assessment	Specific measures and of more variables (e.g., of target behaviors in various situations, of side effects, context, strengths as well as deficiencies)	More global measures (e.g., of cure, or improvement) but only of the individual

Source: Hartmann, Roper, and Bradford (1979)

likely to be elicited. So, for example, if Valeria is most likely to get into verbal and physical altercations during lunch, a behavioral assessor would focus on lunch hour as a time to assess her behavior.

Another way to address the *when* question has to do with the various schedules with which behavioral assessments may be made. For example, one schedule of assessment is referred to as *frequency* or *event recording*. Each time the targeted behavior occurs, it is recorded. Another schedule of assessment is referred to as *interval recording*. Assessment according to this schedule occurs only during predefined intervals of time (for example, every other minute, every 48 hours, every third week). Beyond merely tallying the number of times a particular behavior occurs, the assessor may also maintain a record of the *intensity* of the behavior. Intensity of a behavior may be gauged by observable and quantifiable events such as the *duration* of the behavior, stated in number of seconds, minutes, hours, days, weeks, months, or years. Alternatively, it may be stated in terms of some ratio or percentage of time that the behavior occurs during a specified interval of time. One method of recording the frequency and intensity of target behavior is **timeline followback (TLFB) methodology** (Sobell & Sobell, 1992, 2000). An illustration of the application of TLFB to gambling behavior can be found in the research of Jeremiah Weinstock and colleagues (Weinstock et al., 2004, 2007a, 2007b).

JUST THINK . . .

You are a behavior therapist with a client who is a compulsive gambler. You advise the client to keep a record of his behavior. Do you advise that this self-monitoring be kept on a frequency basis or an interval schedule?

Where? *Where* does the assessment take place? In contrast to the administration of psychological tests, behavioral assessment may take place just about anywhere—preferably in the environment where the targeted behavior is most likely to occur naturally. For example, a behavioral assessor studying the obsessive-compulsive habits of a patient might wish to visit the patient at home to see firsthand the variety and intensity of the behaviors exhibited. Does the patient check the oven for gas left on, for example? If so, how many times per hour? Does the patient engage in excessive hand-washing? If so, to what extent? These and related questions may be raised and answered effectively through firsthand observation in the patient's home. In some instances, when virtual reality is deemed preferable to reality, the assessment may involve stimuli created in a laboratory setting, rather than a "real life" setting (see, for example, Bordnick et al., 2008).

Why? *Why* conduct behavioral assessment? In general, data derived from behavioral assessment may have several advantages over data derived by other means. Data derived from behavioral assessment can be used:

- to provide behavioral baseline data with which other behavioral data (accumulated after the passage of time, after intervention, or after some other event) may be compared
- to provide a record of the assessee's behavioral strengths and weaknesses across a variety of situations
- to pinpoint environmental conditions that are acting to trigger, maintain, or extinguish certain behaviors
- to target specific behavioral patterns for modification through interventions
- to create graphic displays useful in stimulating innovative or more effective treatment approaches

In the era of managed care and frugal third-party payers, let's also note that insurance companies tend to favor behavioral assessments over more traditional assessments. This is because behavioral assessment is typically not linked to any particular theory of personality, and patient progress tends to be gauged on the basis of documented behavioral events.

How? *How* is behavioral assessment conducted? The answer to this question will vary, of course, according to the purpose of the assessment. In some situations, the only special equipment required will be a trained observer with pad and pencil. In other types of situations, highly sophisticated recording equipment may be necessary.

Another key *how* question relates to the analysis of data from behavioral assessment. The extent to which traditional psychometric standards are deemed applicable to behavioral assessment is a controversial issue, with two opposing camps. One camp may be characterized as accepting traditional psychometric assumptions about behavioral assessment, including assumptions about the measurement of reliability (Russo et al., 1980) and validity (Haynes et al., 1979; Haynes et al., 1981). Representative of this position are statements such as that made by Bellack and Hersen (1988) that "the reliability, validity, and utility of any procedure should be paramount, regardless of its behavioral or nonbehavioral development" (p. 614).

JUST THINK . . .

Imagine that you are a NASA psychologist studying the psychological and behavioral effects of space travel on astronauts. What types of behavioral measures might you employ, and what special equipment would you need—or design—to obtain those measures?

Cone (1977) championed the traditionalist approach to behavioral assessment in an article entitled "The Relevance of Reliability and Validity for Behavioral Assessment." However, as the years passed, Cone (1986, 1987) became a leading proponent of an alternative position, one in which traditional psychometric standards are rejected as inappropriate yardsticks for behavioral assessment. Cone (1981) wrote, for example, that "a truly behavioral view of assessment is based on an approach to the study of behavior so radically different from the customary individual differences model that a correspondingly different approach must be taken in evaluating the adequacy of behavioral assessment procedures" (p. 51).

JUST THINK . . .

Do traditional psychometric standards apply to behavioral assessment?

Others, too, have questioned the utility of traditional approaches to test reliability in behavioral assessment, noting that "the assessment tool may be precise, but the behavior being measured may have changed" (Nelson et al., 1977, p. 428). Based on the conceptualization of each behavioral assessment as an experiment unto itself, Dickson (1975) wrote: "If one assumes that each target for assessment represents a single experiment, then what is needed is the scientific method of experimentation and research, rather than a formalized schedule for assessment. . . . Within this framework, each situation is seen as unique, and the reliability of the approach is not a function of standardization techniques . . . but rather is a function of following the experimental method in evaluation" (pp. 376–377).

Approaches to Behavioral Assessment

Behavioral assessment may be accomplished through various means, including behavioral observation and behavior rating scales, analogue studies, self-monitoring, and situational performance methods. Let's briefly take a closer look at each of these as well as related methods.

Behavioral observation and rating scales *A child psychologist observes a client in a playroom through a one-way mirror. A family therapist views a videotape of a troubled family attempting to resolve a conflict. A school psychologist observes a child interacting with peers in the school cafeteria.* These are all examples of the use of an assessment technique termed **behavioral observation.** As its name implies, this technique involves watching the activities of targeted clients or research subjects and, typically, maintaining some kind of record of those activities. Researchers, clinicians, or counselors may themselves serve as observers, or they may designate trained assistants or other people (such as parents, siblings, teachers, and supervisors) as the observers. Even the observed person can be the behavior observer, although in such cases the term *self-observation* is more appropriate than *behavioral observation.*

In some instances, behavioral observation employs mechanical means, such as a video recording of an event. Recording behavioral events relieves the clinician, the researcher, or any other observer of the need to be physically present when the behavior occurs and allows for detailed analysis of it at a more convenient time. Factors noted in behavioral observation will typically include the presence or absence of specific, targeted behaviors, behavioral excesses, behavioral deficits, behavioral assets, and the situational antecedents and consequences of the observed behaviors. Of course, because the people doing the observing and rating are human themselves, behavioral observation isn't always as cut and dried as it may appear (see *Everyday Psychometrics*).

Confessions of a Behavior Rater

In discussions of behavioral assessment, the focus is often placed squarely on the individual being evaluated. Only infrequently, if ever, is reference made to the thoughts and feelings of the person responsible for evaluating the behavior of another. What follows are the hypothetical thoughts of one behavior rater. We say hypothetical because these ideas are not really one person's thoughts but instead a compilation of thoughts of many people responsible for conducting behavioral evaluations.

The behavior raters interviewed for this feature were all on the staff at a community-based inpatient/outpatient facility in Brewster, New York. One objective of this facility is to prepare its adolescent and adult members for a constructive, independent life. Members live in residences with varying degrees of supervision, and their behavior is monitored on a 24-hour basis. Each day, members are issued an eight-page behavior rating sheet referred to as a CDR (clinical data recorder), which is circulated to supervising staff for rating through the course of the day. The staff records behavioral information on variables such as activities, social skills, support needed, and dysfunctional behavior.

On the basis of behavioral data, certain medical or other interventions may be recommended. Because behavioral monitoring is daily and consistent, changes in patient behavior as a function of medication, activities, or other variables are quickly noted and intervention strategies adjusted. In short, the behavioral data may significantly affect the course of a patient's institutional stay—everything from amount of daily supervision to privileges to date of discharge is influenced by the behavioral data. Both patients and staff are aware of this fact of institutional life; therefore, both patients and staff take the completion of the CDR very seriously. With that as background, here are some private thoughts of a behavior rater.

A member receives training in kitchen skills for independent living as a staff member monitors behavior on the CDR.

> I record behavioral data in the presence of patients, and the patients are usually keenly aware of what I am doing. After I am through coding patients' CDRs for the time they are with me, other staff members will code them with respect to the time they spend with the patient. And so it goes. It is as if each patient is keeping a detailed diary of his or her life; only, it is we, the staff, who are keeping that diary for them.
>
> Sometimes, especially for new staff, it feels odd to be rating the behavior of fellow human beings. One morning, perhaps out of empathy for a patient, I tossed a blank CDR to a patient and jokingly offered to let him rate my behavior. By dinner, long after I had forgotten that incident in the morning, I realized the patient was coding me for poor table manners. Outwardly, I laughed. Inwardly, I was really a bit offended. Subsequently, I told a joke to the assembled company that in retrospect probably was not in the best of taste. The patient coded me for being socially offensive. Now, I was genuinely becoming self-conscious. Later that evening, we drove to a local video store to return a tape we had rented, and the patient coded me for reckless driving. My discomfort level rose to the point where I thought it was time to end the joke. In retrospect, I had experienced firsthand the self-consciousness and discomfort some of our patients had experienced as their every move was monitored on a daily basis by staff members.

Even though patients are not always comfortable having their behavior rated—and indeed many patients have outbursts with staff members that are in one way or another related to the rating system—it is also true that the system seems to work. Sometimes, self-consciousness is what is needed for people to get better. Here, I think of Sandy, a bright young man who gradually became fascinated by the CDR and soon spent much of the day asking staff members various questions about it. Before long, Sandy asked if he could be allowed to code his own CDR. No one had ever asked to do that before, and a staff meeting was held to mull over the consequences of such an action. As an experiment, it was decided that this patient would be allowed to code his own CDR. The experiment paid off. Sandy's self-coding kept him relatively "on track" with regard to his behavioral goals, and he found himself trying even harder to get better as he showed signs of improvement. Upon discharge, Sandy said he would miss tracking his progress with the CDR.

Instruments such as the CDR can and probably have been used as weapons or rewards by staff. Staff may threaten patients with a poor behavioral evaluation. Overly negative evaluations in response to dysfunctional behavior that is particularly upsetting to the staff is also an ever-present possibility. Yet all the time you are keenly aware that the system works best when staff code patients' behavior consistently and fairly.

Behavioral observation may take many forms. The observer may, in the tradition of the naturalist, record a running narrative of events using tools such as pencil and paper, a video, film, or still camera, or a cassette recorder. Mehl and Pennebaker (2003), for example, used such a naturalistic approach in their study of student social life. They tracked the conversations of 52 undergraduates across two two-day periods by means of a computerized tape recorder.

Another form of behavioral observation employs what is called a *behavior rating scale*—a preprinted sheet on which the observer notes the presence or intensity of targeted behaviors, usually by checking boxes or filling in coded terms. Sometimes the user of a behavior rating form writes in coded descriptions of various behaviors. The code is preferable to a running narrative because it takes far less time to enter the data and thus frees the observer to enter data relating to any of hundreds of possible behaviors, not just the ones printed on the sheets. For example, a number of coding systems for observing the behavior of couples and families are available. Two such systems are the Marital Interaction Coding System (Weiss & Summers, 1983) and the Couples Interaction Scoring System (Notarius & Markman, 1981). Handheld data entry devices are frequently used today to facilitate the work of the observer.

As approaches to behavioral assessment in general, behavior rating scales and systems may be categorized in different ways. A continuum of *direct* to *indirect* applies to the setting in which the observed behavior occurs and how closely that setting approximates the setting in which the behavior naturally occurs. The more natural the setting, the more direct the measure; the more removed from the natural setting, the less direct the measure (Shapiro & Skinner, 1990). According to this categorization, for example, assessing a firefighter's actions and reactions while fighting a real fire would provide a *direct* measure of firefighting ability. Asking the firefighter to demonstrate reactions to events that occur during a fire would constitute an *indirect* measure of firefighting ability. Shapiro and Skinner (1990) also distinguished between *broad-band* instruments, designed to measure a wide variety of behaviors, and *narrow-band instruments,* which may focus on behaviors related to single, specific constructs such as hyperactivity, shyness, or depression.

Self-monitoring **Self-monitoring** may be defined as the act of systematically observing and recording aspects of one's own behavior and/or events related to that behavior. Self-monitoring is different from self-report. As noted by Cone (1999, p. 411), self-monitoring

> relies on observations of *the* behavior of clinical interest . . . at the *time* . . . and *place* . . . of its actual occurrence. In contrast, self-report uses stand-ins or surrogates (verbal descriptions, reports) of the behavior of interest that are obtained at a time and place different from the time and place of the behavior's actual occurrence. (emphasis in the original)

Self-monitoring may be used to record specific thoughts, feelings, or behaviors. The utility of self-monitoring depends in large part on the competence, diligence, and motivation of the assessee, although a number of ingenious methods have been devised to assist in the process or to ensure compliance (Barton et al., 1999; Bornstein et al., 1986; Wilson & Vitousek, 1999). For example, handheld computers have been programmed to beep as a cue to observe and record behavior (Shiffman et al., 1997).

Self-monitoring is both a tool of assessment and a tool of intervention. In some instances, the very act of self-monitoring (of smoking, eating, anxiety, and panic, for example) may be therapeutic. Practical issues that must be considered include the methodology employed, the targeting of specific thoughts, feelings, or behaviors, the sampling procedures put in place, the actual self-monitoring devices and procedures, and the training and preparation (Foster et al., 1999).

Psychometric issues also must be considered (Jackson, 1999), including the potential problem of *reactivity.* **Reactivity** refers to the possible changes in an assessee's behavior, thinking, or performance that may arise in response to being observed, assessed, or evaluated. For example, if you are on a weight-loss program and are self-monitoring your food intake, you may be more inclined to forgo the cheesecake than to consume it. In this case, reactivity has a positive effect on the assessee's behavior. There are many instances in which reactivity may have a negative effect on an assessee's behavior or performance. For example, we have previously noted how the presence of third parties during an evaluation may adversely effect an assessee's performance on tasks that require memory or attention (Gavett et al., 2005). Education, training, and adequate preparation are some of the tools used to counter the effects of reactivity in self-monitoring. In addition, post-self-monitoring interviews on the effects of reactivity can provide additional insights about the occurrence of the targeted thoughts or behaviors as well as any reactivity effects.

JUST THINK . . .

Create an original example to illustrate how self-monitoring can be a tool of assessment as well as an intervention.

Analogue studies The behavioral approach to clinical assessment and treatment has been likened to a researcher's approach to experimentation. The behavioral assessor proceeds in many ways like a researcher; the client's problem is the dependent variable, and the factor (or factors) responsible for causing or maintaining the problem behavior is the independent variable. Behavioral assessors typically use the phrase **functional analysis of behavior** to convey the process of identifying the dependent and independent variables with respect to the presenting problem. However, just as experimenters must frequently employ independent and dependent variables that imitate those variables in the real world, so must behavioral assessors.

An **analogue study** is a research investigation in which one or more variables are similar or analogous to the real variable that the investigator wishes to examine. This definition is admittedly very broad, and the term *analogue study* has been used

in various ways. It has been used, for example, to describe research conducted with white rats when the experimenter really wishes to learn about humans. It has been used to describe research conducted with full-time students when the experimenter really wishes to learn about people employed full-time in business settings. It has been used to describe research on aggression defined as the laboratory administration of electric shock when the experimenter really wishes to learn about real-world aggression outside the laboratory.

More specific than the term *analogue study* is **analogue behavioral observation,** which, after Haynes (2001a), may be defined as the observation of a person or persons in an environment designed to increase the chance that the assessor can observe targeted behaviors and interactions. The person or persons in this definition may be clients (including individual children and adults, families, or couples) or research subjects (including students, co-workers, or any other research sample). The targeted behavior, of course, depends on the objective of the research. For a client who avoids hiking because of a fear of snakes, the behavior targeted for assessment (and change) is the fear reaction to snakes, most typically elicited while hiking. This behavior may be assessed (and treated) in analogue fashion within the confines of a clinician's office, using photos of snakes, videos of snakes, live snakes that are caged, and live snakes that are not caged.

JUST THINK . . .

As a result of a car accident, a client of a behavior therapist claims not to be able to get into a car and drive again. The therapist wishes to assess this complaint by means of analogue behavioral observation. How should the therapist proceed?

A variety of environments have been designed to increase the assessor's chances of observing the targeted behavior (see, for example, Heyman, 2001; Mori & Armendariz, 2001; Norton & Hope, 2001; and Roberts, 2001). Questions about how analogous some analogue studies really are have been raised, along with questions regarding their ultimate utility (Haynes, 2001b).

Situational performance measures and role-play measures both may be thought of as analogue approaches to assessment.

Situational performance measures If you have ever applied for a part-time clerical job and been required to take a typing test, you have had firsthand experience with *situational performance measures.* Broadly stated, a **situational performance measure** is a procedure that allows for observation and evaluation of an individual under a standard set of circumstances. A situational performance measure typically involves performance of some specific task under actual or simulated conditions. The road test you took to obtain your driver's license was a situational performance measure that involved an evaluation of your driving skills in a real car on a real road in real traffic. On the other hand, situational performance measures used to assess the skills of prospective space-traveling astronauts are done in rocket simulators in laboratories firmly planted on Mother Earth. Common to all situational performance measures is that the construct they measure is thought to be more accurately assessed by examining behavior directly than by asking subjects to describe their behavior. In some cases, subjects may be motivated to misrepresent themselves, as when asked about moral behavior. In other situations, subjects may simply not know how they will respond under particular circumstances, as in a stress test.

The **leaderless group technique** is a situational assessment procedure wherein several people are organized into a group for the purpose of carrying out a task as an observer records information related to individual group members' initiative, cooperation, leadership, and related variables. Usually, all group members know they are being evaluated and that their behavior is being observed and recorded. Purposely vague

instructions are typically provided to the group, and no one is placed in the position of leadership or authority. The group determines how it will accomplish the task and who will be responsible for what duties. The leaderless group situation provides an opportunity to observe the degree of cooperation exhibited by each individual group member and the extent to which each is able to function as part of a team.

The leaderless group technique has been employed in military and industrial settings. Its use in the military developed out of attempts by the U.S. Office of Strategic Services (OSS Assessment Staff, 1948) to assess leadership as well as other personality traits. The procedure was designed to aid in the establishment of cohesive military units—cockpit crews, tank crews, and so forth—in which members would work together well and could each make a significant contribution. Similarly, the procedure is used in industrial and organizational settings to identify people who work well together and those with superior managerial skills and "executive potential."

JUST THINK . . .

You are a management consultant to a major corporation with an assignment: Create a situational performance measure designed to identify an *unleader.* Briefly outline your plan.

The self-managed work-group approach challenges traditional conceptions of manager and worker. How does one manage a group that is supposed to manage itself? One approach is to try to identify *unleaders,* who act primarily as facilitators in the workplace and are able to balance a hands-off management style with a style that is more directive when necessary (Manz & Simms, 1984).

Role play The technique of **role play,** or acting an improvised or partially improvised part in a simulated situation, can be used in teaching, therapy, and assessment. Police departments, for example, routinely prepare rookies for emergencies by having them play roles, such as an officer confronted by a criminal holding a hostage at gunpoint. Part of the prospective police officer's final exam may be successful performance on a role-playing task. A therapist might use role play to help a feuding couple avoid harmful shouting matches and learn more effective methods of conflict resolution. That same couple's successful resolution of role-played issues may be one of a therapist's criteria for terminating therapy.

A large and growing literature exists on role play as a method of assessment. In general, role play can provide a relatively inexpensive and highly adaptable means of assessing various behavior "potentials." We cautiously say "potentials" because of the uncertainty that role-played behavior will then be elicited in a naturalistic situation (Kern et al., 1983; Kolotkin & Wielkiewicz, 1984). Bellack et al. (1990) employed role play for both evaluative and instructional purposes with psychiatric inpatients who were being prepared for independent living. While acknowledging the benefits of role play in assessing patients' readiness to return to the community, these authors cautioned that "the ultimate validity criterion for any laboratory- or clinic-based assessment is unobtrusive observation of the target behavior in the community" (p. 253).

JUST THINK . . .

Describe a referral for evaluation that would ideally lend itself to the use of role play as a tool of assessment.

Psychophysiological methods The search for clues to understanding and predicting human behavior has led researchers to the study of physiological indices such as heart rate and blood pressure. These and other indices are known to be influenced by psychological factors—hence the term **psychophysiological** to describe these variables as well as the methods used to study them. Whether these methods are properly regarded as *behavioral* in nature is debatable. Still, these techniques do tend to be associated with behaviorally oriented clinicians and researchers.

Perhaps the best known of all psychophysiological methods used by psychologists is *biofeedback.* **Biofeedback** is a generic term that may be defined broadly as a class of psychophysiological assessment techniques designed to gauge, display, and record a continuous monitoring of selected biological processes such as pulse and blood pressure. Depending on how biofeedback instrumentation is designed, many different biological processes—such as respiration rate, electrical resistance of the skin, and brain waves—may be monitored and "fed back" to the assessee via visual displays, such as lights and scales, or auditory stimuli, such as bells and buzzers.

The use of biofeedback with humans was inspired by reports that animals given rewards (and hence feedback) for exhibiting certain involuntary responses (such as heart rate) could successfully modify those responses (Miller, 1969). Early experimentation with humans demonstrated a capacity to produce certain types of brain waves on command (Kamiya, 1962, 1968). Since that time, biofeedback has been used in a wide range of therapeutic and assessment-related applications (French et al., 1997; Hazlett et al., 1997; Hermann et al., 1997; Zhang et al., 1997).

The **plethysmograph** is an instrument that records changes in the volume of a part of the body arising from variations in blood supply. Investigators have used this device to explore changes in blood flow as a dependent variable. For example, Kelly (1966) found significant differences in the blood supplies of normal, anxiety-ridden, and psychoneurotic groups (the anxiety group having the highest mean) by using a plethysmograph to measure blood supply in the forearm.

A **penile plethysmograph** is also an instrument designed to measure changes in blood flow, but more specifically blood flow to the penis. Because the volume of blood in the penis increases with male sexual arousal, the penile plethysmograph has found application in the assessment of male sexual offenders. In one study, subjects who were convicted rapists demonstrated more sexual arousal to descriptions of rape and less arousal to consenting-sex stories than did control subjects (Quinsey et al., 1984). Offenders who continue to deny deviant sexual object choices may be confronted with the findings from such studies as a means of compelling them to speak more openly about their thoughts and behavior (Abel et al., 1986). **Phallometric data,** as it is referred to, also has treatment and program evaluation applications. In one such type of application, the offender—a rapist, a child molester, an exhibitionist, or some other sexual offender—is exposed to visual and/or auditory stimuli depicting scenes of normal and deviant behavior while penile tumescence is simultaneously gauged.

In the public eye, the best-known of all psychophysiological measurement tools is what is commonly referred to as a *lie detector* or **polygraph** (literally, "more than one graph"). Although not commonly associated with psychological assessment, the lie detection industry—given the frequency with which such tests are administered and the potential consequences of the tests—may be characterized as "one of the most important branches of applied psychology" (Lykken, 1981, p. 4). Based on the assumption that detectable physical changes occur when an individual lies, the polygraph provides a continuous written record (variously referred to as a *tracing,* a *graph,* a *chart,* or a *polygram*) of several physiological indices (typically respiration, galvanic skin response, and blood volume/pulse rate) as an interviewer and instrument operator (known as a *polygrapher* or *polygraphist*) asks the assessee a series of yes–no questions. Judgments of the truthfulness of the responses are made either informally by surveying the charts or more formally by means of a scoring system.

JUST THINK . . .

Polygraph evidence is not admissable in most courts, yet law enforcement agencies and the military continue to use it as a tool of evaluation. Your thoughts?

The reliability of judgments made by polygraphers is a matter of controversy (Iacono & Lykken, 1997). Different methods of conducting polygraphic examinations

exist (Lykken, 1981), and polygraphic equipment is not standardized (Abrams, 1977; Skolnick, 1961). A problem with the method is a high false-positive rate for lying. The procedure "may label more than 50% of the innocent subjects as guilty" (Kleinmuntz & Szucko, 1984, p. 774). In light of the judgments that polygraphers are called upon to make, their education, training, and background requirements seem minimal: One may qualify as a polygrapher after as few as six weeks of training. From the available psychometric and related data, it seems reasonable to conclude that the promise of a machine purporting to detect dishonesty remains unfulfilled (Alpher & Blanton, 1985).

JUST THINK . . .

Webb et al. (1966) argued that unobtrusive measures can usefully complement other research techniques such as interviews and questionnaires. What unobtrusive measure could conceivably be used to complement a questionnaire on student study habits?

Unobtrusive measures A type of measure quite different from any we have discussed so far is the *nonreactive* or *unobtrusive* variety (Webb et al., 1966). In many instances, an **unobtrusive measure** is a telling physical trace or record. In one study, it was garbage—literally (Cote et al., 1985). Because of their nature, unobtrusive measures do not necessarily require the presence or cooperation of respondents when measurements are being conducted. In a now-classic book that was almost entitled *The Bullfighter's Beard*,[8] Webb et al. (1966) listed numerous examples of unobtrusive measures, including the following:

- The popularity of a museum exhibit can be measured by examination of the erosion of the floor around it relative to the erosion around other exhibits.
- The amount of whiskey consumption in a town can be measured by counting the number of empty bottles in trashcans.
- The degree of fear induced by a ghost-story-telling session can be measured by noting the shrinking diameter of a circle of seated children.

JUST THINK . . .

Stice et al. (2004) devised several unobtrusive measures to estimate the caloric intake of dieters; however, they were unable to devise an ethically acceptable way to gauge caloric intake in the home. Can you think of a way to accomplish this objective?

More recently, wrappers left on trays at fast-food restaurants were used to estimate the caloric intake of restaurant patrons (Stice et al., 2004). In another innovative use of a "telling record," researchers used college yearbook photos to study the relationship between positive emotional expression and other variables, such as personality and life outcome (see this chapter's *Close-up*).

Issues in Behavioral Assessment

The psychometric soundness of tools of behavioral assessment can be evaluated, but how best to do that is debatable. More specifically, questions arise about the appropriateness of various models of measurement. You may recall from Chapter 5 that classical test theory and generalizability theory conceptualize test-score variation in somewhat different ways. In generalizability theory, rather than trying to estimate a single true score, consideration is given to how test scores would be expected to shift across

8. Webb et al. (1966) explained that the provocative, if uncommunicative, title *The Bullfighter's Beard* was a "title drawn from the observation that toreadors' beards are longer on the day of the fight than on any other day. No one seems to know if the toreador's beard really grows faster that day because of anxiety or if he simply stands further away from the blade, shaking razor in hand. Either way, there were not enough American aficionados to get the point" (p. v). The title they finally settled on was *Unobtrusive Measures: Nonreactive Research in the Social Sciences.*

CLOSE-UP

Personality, Life Outcomes, and College Yearbook Photos

Few people would be shocked to learn that individual differences in emotion are associated with differences in personality. Yet it will probably surprise many to learn that interpersonal differences in emotion may well have a pervasive effect on the course of one's life. In one study, it was observed that a tendency to express uncontrolled anger in early childhood was associated with ill temper across the lifespan and with several negative life outcomes, such as lower educational attainment, lower-status jobs, erratic work patterns, lower military rank, and divorce (Caspi et al., 1987). Suggestive findings such as these have prompted other investigators to wonder about the possible effects of positive emotions on personality and life outcomes.

Positive emotions have many beneficial effects, ranging from the broadening of thoughts and action repertoires (Cunningham, 1988; Frederickson, 1998; Isen, 1987) to the facilitation of the approach of other people (Berry & Hansen, 1996; Frijda & Mesquita, 1994; Ruch, 1993). A smile may send the message that one is friendly and nonthreatening (Henley & LaFrance, 1984; Keating et al., 1981) and may lead to positive attributions about one's sociability, friendliness, likeability, and stability (Borkenau & Liebler, 1992; Frank et al., 1993; Matsumoto & Kudoh, 1993). On the basis of such findings and related research, Harker and Keltner (2001) hypothesized that positive emotional expression would predict higher levels of well-being across adulthood. They tested the hypothesis by examining the relationship of individual differences in positive emotional expression to personality and other variables.

A measure of positive emotional expression was obtained by coding judges' ratings of college yearbook photographs of women who participated in a longitudinal research project (Helson, 1967; Helson et al., 1984). These coded judgments were analyzed with respect to personality data on file (such as the subjects' responses to the Adjective Check List at ages 21, 27, 43, and 52) and life outcome data (including well-being as measured by the California Psychological Inventory, marital status, and the Marital Tensions Checklist).

Consistent with the researchers' hypothesis, positive emotional expression as evidenced in the college yearbook photos was found to correlate positively with life outcomes such as marital satisfaction and sense of personal well-being. This was the case even when the possible confounding influences of physical attractiveness or social desirability in responding were controlled for in the analysis of the data.

Is there a relationship between emotion expressed in college yearbook photos and personality and life outcomes? According to one study, the answer is yes. Researchers found that positive emotional expression in women's college photos predicted favorable outcomes in marriage and personal well-being up to 30 years later.

The researchers cautioned, however, that the measure of emotional expression used in the study (the yearbook photo) consisted of a single instance of very limited behavior. They urged future researchers to consider the use of different measures of emotional expression obtained in different contexts. The researchers also cautioned that their findings are limited to research with women. Smiling may have different implications for the lives of men (Stoppard & Gruchy, 1993). In fact, smiling was negatively correlated with positive outcomes for a sample of male cadets at West Point (Mueller & Mazur, 1996).

This thought-provoking study was, according to Harker and Keltner (2001), "one of the first to document that individual differences in expression relate to personality and may be stable aspects of personality" (p. 121).

situations as a result of changes in the characteristic being measured. It is for this and related reasons that generalizability theory seems more applicable to behavioral assessment than to the measurement of personality traits. Behavior changes across situations, necessitating an approach to reliability that can account for those changes. In contrast, personality traits are assumed by many to be relatively stable across situations. Personality traits are therefore presumed to be more appropriately measured by instruments with assumptions that are consistent with the true score model.

Regardless of whether behavioral measures are evaluated in accordance with classical test theory, generalizability theory, or something else (such as a Skinnerian experimental analysis), it would seem there are some things on which everyone can agree. One is that there must be an acceptable level of inter-rater reliability among behavior observers or raters. A potential source of error in behavioral ratings may arise when a dissimilarity in two or more of the observed behaviors (or other things being rated) leads to a more favorable or unfavorable rating than would have been made had the dissimilarity not existed (Maurer & Alexander, 1991). A behavioral rating may be excessively positive (or negative) because a prior rating was excessively negative (or positive). This source of error is referred to as a **contrast effect** (Figure 13–10).

Contrast effects have been observed in interviews (Schuh, 1978), in behavioral diaries and checklists (Maurer et al., 1993), in laboratory-based performance evaluations (Smither et al., 1988), and in field performance evaluations (Ivancevich, 1983). In one study of employment interviews, as much as 80% of the total variance was thought to be due to contrast effects (Wexley et al., 1972).

To combat potential contrast effects and other types of rating error, rigorous training of raters is necessary. However, such training may be costly in terms of time and labor. For example, teaching professionals how to use the behavior observation and coding system of the Marital Interaction Coding System took "two to three months of weekly instruction and practice to learn how to use its 32 codes" (Fredman & Sherman, 1987, p. 28). Another approach to minimizing error and improving inter-rater reliability among behavioral raters is to employ a **composite judgment,** which is, in essence, an averaging of multiple judgments.

Some types of observer bias cannot practically or readily be remedied. For example, in behavioral observation involving the use of video equipment, it would on many occasions be advantageous if multiple cameras and recorders could be used to cover

Figure 13–10
The Contrast Effect at the Rink

Figure skating judges, like other behavior raters, are only human. Skaters who give performances worthy of extremely high marks may not always get what they deserve, simply because the skater who performed just before they did excelled by contrast. Ratings may be more favorable when the performance just prior to theirs was very poor. Because of this contrast effect, the points earned by a skater may depend to some degree on the quality of the preceding skater's performance.

various angles of the ongoing action, to get close-ups, and so forth. The economic practicality of the situation (let alone other factors, such as the number of hours required to watch footage from multiple views) is that it is seldom feasible to have more than one camera in a fixed position recording the action. The camera is in a sense biased in that one fixed position because in many instances it is recording information that may be quite different from the information that would have been obtained had it been placed in another position—or if multiple recordings were being made.

As we have already noted in the context of self-monitoring, reactivity is another possible issue with regard to behavioral assessment. This means that people react differently in experimental than in natural situations. Microphones, cameras, and one-way mirrors may in themselves alter the behavior of persons being observed. For example, some patients under videotaped observation may attempt to minimize the amount of psychopathology they are willing to record for posterity; others under the same conditions may attempt to exaggerate it. One possible solution to the problem of reactivity is the use of hidden observers or clandestine recording techniques, although such methods raise serious ethical issues. Many times, all that is required to solve the problem of reactivity is an adaptation period. People being observed may adjust to the idea and begin to behave in their typical ways. Most clinicians are aware from personal experience that a tape recorder in the therapy room might put off some patients at first, but in only a matter of minutes the chances are good that it will be ignored.

Some of the other possible limitations of behavioral approaches include the equipment costs (some of the electronics can be expensive) and the cost of training behavioral assessors (Kenny et al., 2008). If training is not sufficient, another "cost"—one that few behavioral assessors are willing to pay—may be unwanted variables in their reports such as observer error or bias.

A Perspective

More than a half-century ago, Theodor Reik's influential book *Listening with the Third Ear* intrigued clinicians with the possibilities of evaluation and intervention by means of skilled interviewing, active listening, and artful, depth-oriented interpretation. In one vignette, a female therapy patient recounted a visit to the dentist that involved an injection and a tooth extraction. While speaking, she remarked on a book in Reik's bookcase that was "standing on its head"—to which Reik responded, "But why did you not tell me that you had had an abortion?" (Reik, 1948, p. 263). Reflecting on this dazzling exhibition of clinical intuition, Masling (1997) wrote, "We would all have liked to have had Reik's magic touch, the ability to discern what is hidden and secret, to serve as oracle" (p. 259).

Historically, society has called upon mental health professionals to make diagnostic judgments and intervention recommendations, and often on the basis of relatively little information. Early on, psychological tests, particularly in the area of personality assessment, promised to empower clinicians—mere mortals—to play the oracular role society imposed and expected. Soon, two very different philosophies of test design and use emerged. The clinical approach relied heavily on the clinician's judgment and intuition and was characterized by a lack of preset and uniformly applied rules for drawing clinical conclusions and making predictions. By contrast, the statistical or actuarial approach relied heavily on standardization, norms, and preset, uniformly applied rules and procedures. Duels between various members of these two camps were common for many years and have been reviewed in detail elsewhere (Marchese, 1992).

It seems fair to say that in those situations where data are insufficient to formulate rules for decision making and prediction, the clinical approach wins out over the actuarial. For the most part, however, it is the actuarial approach that has been most enthusiastically embraced by contemporary practitioners. This is so for a number of reasons, chief among them a passionate desire to make assessment more a science than an art. And that desire may simply reflect that most of us are not oracles. Without good tools, it is difficult if not impossible to spontaneously and consistently see through to what Reik (1952) characterized as the "secret self." Even with good tools, it's a challenge.

The actuarial approach permits hypotheses and predictions that have been found useful to retain; conversely, it enables practitioners to quickly discover and discard untenable hypotheses and predictions (Masling, 1997). Of course, in many instances, skill in clinical assessment can be conceptualized as an internalized, less formal, and more creative version of the actuarial approach.

The actuarial approach to personality assessment is increasingly common. Even projective instruments, once the bastion of the "old school" clinical approach, are increasingly published with norms and subsequently researched using rigorous statistical methods. There have even been efforts—very respectable efforts—to apply sophisticated IRT models to, of all things, TAT data (Tuerlinckx et al., 2002). But academicians have by and large remained unimpressed: "In academic psychology the climate of opinion about projective tests continues as though nothing has changed and clinicians were still reading tea leaves" (Masling, 1997, p. 263).

If the oracle-like, clinical orientation is characterized as the *third ear approach,* we might characterize the contemporary orientation as a *van Gogh approach;* in a sense, an ear has been dispatched. The day of the all-knowing oracle has passed. Today, it is incumbent upon the responsible clinician to rely on norms, inferential statistics, and related essentials of the actuarial approach. Sound clinical judgment is still desirable, if not mandatory. However, it is required less for the purpose of making off-the-cuff interpretations and predictions and more for the purpose of organizing and interpreting information from different tools of assessment. We'll have more to say on this point as we move to the next chapter, Clinical and Counseling Assessment.

Self-Assessment

Test your understanding of elements of this chapter by seeing if you can explain each of the following terms, expressions, and abbreviations:

analogue behavioral observation
analogue study
apperceive
behavioral assessment
behavioral observation
biofeedback
composite judgment
contrast effect
comprehensive system (Exner)
figure drawing test
free association
functional analysis of study
genogram
implicit motive
inquiry (on the Rorschach)
leaderless group technique
need (Murray)
objective methods of personality assessment
penile plethysmograph
percept (on the Rorschach)
phallometric data
plethysmograph
polygraph
press (Murray)
projective hypothesis
projective method
psychophysiological (assessment methods)
reactivity
role play
Rorschach scoring system
Rorschach test
self-monitoring
sentence completion stem
situational performance measure
TAT
testing the limits (on the Rorschach)
thema (Murray)
timeline followback (TLFB) methodology
unobtrusive measure
word association