Cohen, R. J., Montague, P., Nathanson, L. S., & Swerdlik, M. E. (1988). *Psychological Testing: An Introduction to Tests and Measurements*. Mountain View, CA: Mayfield Publishing Co.

Chapter 11
Personality Assessment:
An Overview

1. In a 1950s' vintage oldie-but-goodie rock 'n' roll tune called "Personality," singer Lloyd Price described the subject of his song in terms of walk, talk, smile, and charm. In so doing, Price's use of the term "personality" was quite consistent with the way that most people tend to use the term. For lay people, "personality" refers to components of an individual's make-up that can elicit positive or negative reactions from others. The individual who consistently tends to elicit positive reactions from others is thought to have a "good" personality. The individual who consistently tends to elicit not-so-good reactions from others is thought to have a "bad" personality or, perhaps worse yet, "no personality." Other descriptive terms such as "aggressive personality," "cold personality," and "warm personality" also enjoy widespread usage.
2. When behavioral scientists seek to define and describe personality, the terms they use are more rigorous than those describing simple social skills and are more precise than all-encompassing adjectives. The search has led to the serious study of constructs such as personality traits, personality types, and personality states. In this chapter we survey various approaches to assessing personality and constructing personality tests. Our survey continues in Chapter 12, where we focus exclusively on projective tests. In Chapter 13 we look at other tools that have been used in the process of personality assessment. We begin by defining some of the terms that we use throughout Part 4. As you will see, defining some of these terms is not at all easy. However, logically [begin page 286] speaking, it is important to arrive at working definitions of these terms before proceeding to a discussion of how to measure them.

Defining and Measuring "Personality"

3. Dozens of distinctly different definitions of "personality" exist in the psychology literature (Allport, 1937). Some definitions appear to be all-inclusive in nature. For example, McClelland (1951, p. 69) defined personality as "the most adequate conceptualization of a person's behavior in all its detail." Menninger (1953, p. 23) defined it as "the individual as a whole, his height and weight and love and hates and blood pressure and reflexes; his smiles and hopes and bowed legs and enlarged tonsils. It means all that anyone is and that he is trying to become." Some definitions rely heavily on a particular aspect of the person such as the individual's phenomenal field (Goldstein, 1963) or the individual as a social being (Sullivan, 1953). At an extreme end of the spectrum of definitions are those proposed by theorists who have scrupulously avoided definition. For example, Byrne (1974, p. 26) characterized the entire area of personality psychology as "psychology's garbage bin in that any research which doesn't fit other existing categories can be labelled 'personality.'" Deploring personality theorists who avoid defining their subject matter, Dahlstrom (1970) observed that

   > Some sidestep the issue, apparently to satisfy a demand for ostensive definitions. Thus, Sarason states, "We shall consider personality as an area of investigation rather than as an entity, real or hypothetical" (1966, p. 15). While such a definition makes it easy to point to the definienda ("I am studying what the personologist over there is doing"), it obviously leaves the central definition itself unformulated. (p. 2)

4. In their widely read and authoritative textbook, *Theories of Personality,* Hall and Lindzey (1970, p. 9) wrote that "it is our conviction that *no substantive definition of personality can be applied with any generality"* and that *"personality is defined by the particular empirical concepts which are a part of the theory of personality employed by the observer."* They went on, "If this seems an unsatisfactory definition to the reader, let him take consolation in the thought that in the pages to follow he will encounter a number of specific definitions any one of which will become his if he chooses to adopt that particular theory" (p. 9)[1]

5. At this point you might well ask, "If venerable authorities like Hall and Lindzey aren't going to define personality, who are Cohen, Montague, Nathanson, and Swerdlik to think that they can do it?" Our response is to formulate a middle-of-the-road definition: one that represents a middle ground between the all-inclusive "whole person" types of definitions and the nondefinition types of definitions. We find the following definition useful for our purposes (that is, the teaching of psychological testing): *"Personality* may be defined as *an individual's unique constellation of psychological* [begin page 287] *traits and states.* Accordingly, *personality assessment* entails the measurement of traits and states." Before proceeding to a discussion of strategies used to accomplish such measurement, we should define "traits" and "states." We also define another widely used personality-related term, "types."

Personality Types

6. The vocabulary of personality assessment relies heavily on trait terms (such as "warm," "reserved," "trusting," and "imaginative"). If you have taken a course in personality theory you are probably aware that just as there is no consensus about the definition of "personality," no consensus exists regarding the word "trait." Theorists such as Gordon Allport (1937) have tended to view personality traits as real physical entities that are "bona fide mental structures in each personality" (p. 289). For Allport, a trait is a "generalized and focalized neuropsychic system (peculiar to the individual) with the capacity to render many stimuli functionally equivalent, and to initiate and guide consistent (equivalent) forms of adaptive and expressive behavior" (p. 295). Robert Holt (1971) noted that there "are real structures inside people that determine their behavior in lawful ways" (p. 6), and he went on to conceptualize these structures in terms of changes in brain chemistry that might occur as a result of learning: "learning causes submicroscopic structural changes in the brain, probably in the organization of its biochemical substance" (p. 7). Raymond Cattell (1950) also conceptualized traits as "mental structures," but for him "structure" did not necessarily imply actual physical status.
7. Our own preference is to shy away from definitions that elevate *trait* to the status of physical existence; rather than physical entities, we tend to view psychological traits as attributions made in an effort to identify threads of consistency in behavioral patterns. A definition of *trait* offered by Guilford (1959, p. 6) has great appeal to us. He defined *trait* as, "any distinguishable, relatively enduring way in which one individual varies from another."
8. Inherent in this relatively simple definition are commonalities with the writings of other personality theorists such as Allport (1937), Cattell (1950, 1965), and Eysenck (1961). The word "distinguishable" conveys the idea that behavior labeled with one trait term can be differentiated from behavior that is labeled with another trait term. Thus, for example, behavior within a certain context that might be viewed as "religious" should ideally be distinguishable from behavior within the same or another context that might be viewed as "deviant." Note here that it is important to be aware of the *context* or situation in which a particular behavior is displayed when distinguishing between trait terms that may be applicable; a person who is kneeling and talking to God inside of a church may be described as "religious," while another person engaged in the exact same behavior in a public restroom might more readily be viewed as "deviant." The trait term that an observer applies, as well as the strength or magnitude of the trait presumed to be present, is based on an observation of a sample of behavior. The observed sample of behavior may be obtained in a number of ways, ranging from direct observation of the assessee (such as by actually watching the individual going to church regularly and praying) to the analysis of the assessee's statements on a self-report, pencil-and-paper personality test (on which, for example, the individual may have provided an indication of great frequency in church attendance). [begin page 288]
9. In his definition of "trait," Guilford did not assert that traits represent enduring ways in which individuals vary

---

1. Hall and LIndzey (1970) did point out that important theoretical differences underlie the various different types of definitions of "personality" that exist. After Allport (1937), Hall and Lindzey (1970, p. 8) point out, for example, that a distinction can be made between *biosocial* types of definitions (that is, definitions that equate personality with the social stimulus value of the individual), and *biophysical* types of definitions (that is, definitions that do not take account of the social stimulus value of the individual but are solely rooted within the individual).

from one another; rather, the term *relatively enduring way* was used. The modifier "relatively" serves to emphasize that exactly how a particular trait manifests itself is, at least to some extent, situation-dependent. For example, a "violent" parolee may generally be prone to behave in a rather subdued way with her parole officer and much more violently in the presence of her family and friends. John may be viewed as "dull" and "cheap" by his wife but as "charming" and "extravagant" by his secretary, business associates, and others he is keenly interested in impressing. Allport (1937) addressed the issue of cross-situational consistency-or lack of it-as follows:

> Perfect consistency will never be found and must not be expected. . . . People may be ascendant and submissive, perhaps submissive only towards those individuals bearing traditional symbols of authority and prestige; and towards everyone else aggressive and domineering. . . . The ever changing environment raises now one trait and now another to a state of active tension. (p. 330)

10. Returning to our elaboration of Guilford's definition, note that "trait" is described as a *way in which one individual varies from another.* Here it is important to emphasize that the attribution of a trait term is always a *relative* phenomena. For instance, some behavior described as "patriotic" may differ greatly from other behavior also described as "patriotic." No absolute standards prevail here; in saying that one person is "patriotic," we are in essence making an unstated comparison to the degree of patriotic behavior that could reasonably be expected to be emitted by the average person.
11. Research demonstrating a lack of cross—situational consistency in traits such as honesty (Hartshorne & May, 1928), punctuality (Dudycha, 1936), conformity (Hollander & Willis, 1967), attitude toward authority (Burwen & Campbell, 1957), and introversion/extraversion (Newcomb, 1929) are the types of studies typically cited by Mischel (1968, 1973, 1977, 1979) and others who have been critical of the predominant role of the concept of traits in personality theory. Such critics may also allude to the fact that some undetermined portion of behavior exhibited in public may be governed more by societal expectations and cultural role restrictions than by an individual's personality traits (see Goffman, 1963; Barker, 1963). Research designed to shed light on the primacy of individual differences versus situational factors in behavior is methodologically complex (see Golding, 1975), and the verdict as to the primacy of the trait or the situation is far from being in (see Moskowitz & Schwartz, 1982).

Personality Types

12. Having defined personality as a unique constellation of traits and states we might define a personality *type* as a constellation of traits and states that is similar in pattern to one identified category of personality within a taxonomy of personalities. For assistance in elaborating on this definition of type, we can look to the work of Isabel Briggs Myers and Katherine C. Briggs, authors of the Myers-Briggs Type Indicator (Myers & Briggs, 1943/1962), a test inspired by the theoretical typology of Carl Jung (1923). An assumption guiding the development of this test was that people exhibit definite preferences in the way that they perceive or become aware of, and judge or arrive at conclusions about, people, events, situations, and ideas. According to Myers [begin page 289] (1962, p. 1), these differences in perception and judging result in "corresponding differences in their reactions, in their interests, values, needs and motivations, in what they do best, and in what they like to do."[1] While traits are frequently discussed as if they were something individuals possess, types are more clearly only descriptions of people — not something presumed to be inherent in them.
13. Hypotheses and notions about various *types* of people have appeared in the literature through the ages. Perhaps the most primitive personality typology was the humoral theory of Hippocrates (see Chapter 2). Centuries later, the personality theorist Alfred Adler would differentiate personality types in a way that was somewhat

---

1.In an interesting exploratory study designed to better understand the personality of chess players, the Myers-Briggs Type Indicator was administered to 2,165 chess players (including masters and senior masters). The chess players were found to be significantly more introverted, intuitive, and thinking (as opposed to feeling) than members of the general population. The investigator also found masters to be more judging than would be expected in the general population (Kelly, 1985).

reminiscent of Hippocrates (Table 11-1). Adler's personality types represented different combinations of social interests and varying degrees of vigor with which they attacked life's problems. Adler (1933/1964, p. 127) never developed a formal system to measure these types since he realized that they were generalizations, useful primarily for teaching persons. By contrast, another personality theorist, physician William Sheldon, developed an elaborate typology based on measurements of body mass (see Figure 11-1).

Table 11-1: Two Typologies: Adler and Hippocrates

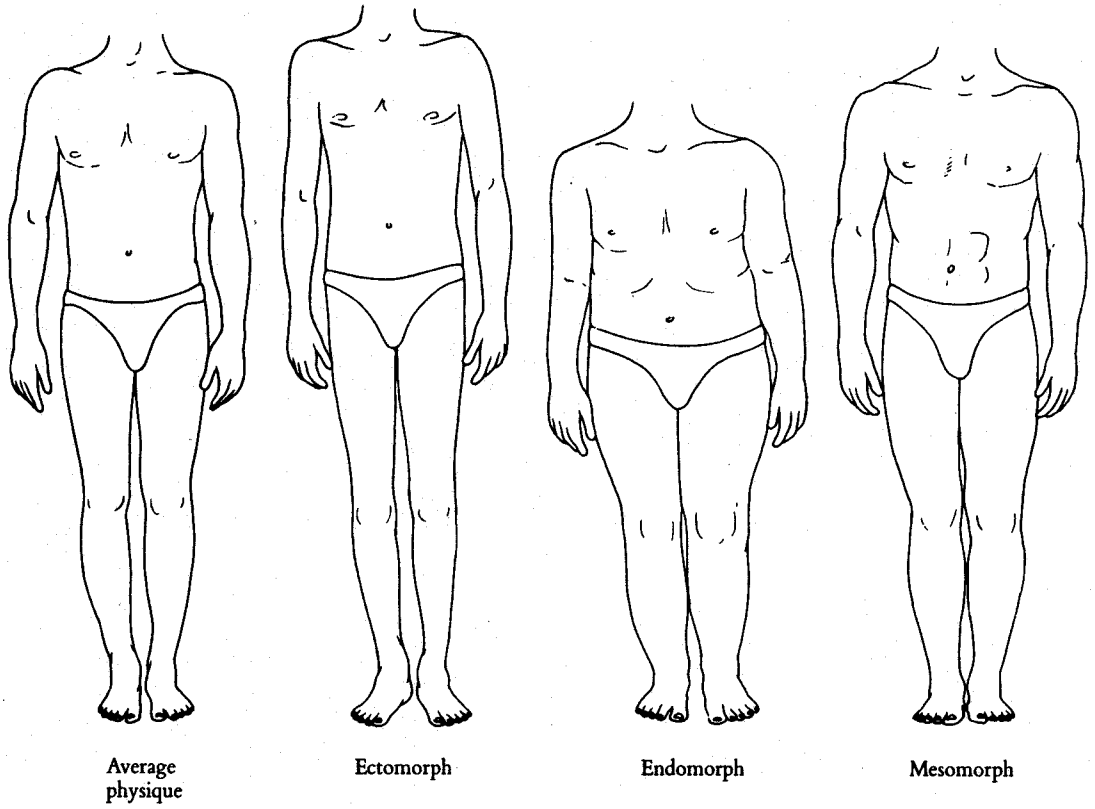
Adlerian Type	Corresponding type of Hippocrates
Ruling type: High activity but in an asocial way; typical of "bossy" people and, in the extreme, homicidal people.	Choleric type
Getting type: This type of person has low social interest and a moderate activity level; typical of people who are constantly depending on others for support.	Phlegnatic or sluggish type
Avoiding type: This type of person has very low social interest combined with a very low activity level; this method of coping relies primarily on avoidance.	Melancholic type
Good Man type: This type of person has high social interest combined with a high activity level; she or he lives life to the fullest and is very much concerned with the well-being of his or her fellow human beings.	Sanguine type

Source: Adler (1927/1965)

Personality States

14. The word *state* has been used in at least two distinctly different ways in the personality assessment literature. In one usage of this term, a personality state is an inferred psychodynamic disposition designed to convey the dynamic quality of id, ego, and superego in perpetual conflict. Assessment of these psychodynamic dispositions may be made through the use of various psychoanalytic techniques such as free association, [begin page 290] See Sheldon's figure.
15. [begin page 291] word association, symbolic analysis of interview material, dream analysis, and analysis of slips of the tongue, accidents, jokes, and forgetting.
16. Presently, a more popular usage of the *state* — and the one that we make reference to in the discussion that follows — refers to the transitory exhibition of some trait. Put another way, the use of the word "trait" presupposes a relatively enduring behavioral disposition, while the term "state" is indicative of a relatively temporary predisosition. Thus, for example, Sally may be described as being "in an anxious state before her midterms, though no one who knows Sally well would describe her as "an anxious person."
17. Measuring personality states amounts, in essence, to a search for and assessment of the strength of traits that are relatively transitory in nature and/or fairly situation — specific. Relatively few existing personality tests seek to distinguish traits from states. Seminal work in this area was done by Charles D. Speilberger and his associates. These researchers developed a number of personality inventories designed to distinguish various states from traits. Included here are the State-Trait Anxiety Inventory (Spielberger, Gorsuch, & Lushene, 1970), the State-Trait Anxiety Inventory for Children (Spielberger, Edwards, Montuori & Lushene, 1973), the State-Trait Anger Scale, (Spielberger et al., 1980a) and the Test Anxiety Inventory, Research Edition (Spielberger et al., 1980b).

Figure 11-1 ***Sheldon's Typology.***



---------- text for sheldon figure-------

William Sheldon and his associates (Sheldon & Stevens, 1942; Sheldon, Dupertuis, & McDermott, 1954) proposed a personality typology based on body build. This complicated typology involved measurements of body mass and ratio that culminated in c assification with respect to three body types: the endomorph, the mesomorph, and ectomorph. Associated with each of these body types are specific predispositions and temperaments. The endomorph, for example, was said to have a "viscerotonic" disposition, which implied, among other things, a love of good food and good company and general even-temperedness. The mesomorph is "somatotonic": action-oriented, adventuresome, and dominating, among other things. The ectomorph is "cerebrotonic": physically and emotionally restrained, future-oriented, and introverted. For Sheldon, the task of assessment was one of classifying persons with respect to three dimensions of physique. Each individual was rated on a scale from 1 to 7 according to the amount of endomorphy, mesomorphy, and ectomorphy that was deemed to be present. An individual who was the epitome of an endomorph would thus be rated as a "7-1-1"; 7 for endomorphy (the highest possible rating), 1 for mesomorphy, and 1 for ectomorphy (the lowest possible rating). An individual who was high on mesomorphy, medium on endomorphy, and low on ectomorphy would be rated 3-7-1; presumably such an individual would also have a termperament that corresponded to this particular "somatotype" (or "body type").

---------- end text for sheldon figure -------

18. In the manual for the State-Trait Anxiety Inventory (STAI), for example, we find that *state anxiety* refers to a transitory experience of tension due to a particular situation. By contrast, *trait anxiety* or "anxiety proneness" refers to a relatively stable or enduring personality characteristic. The STAI test items consist of short descriptive statements, and subjects are instructed to indicate either (1) how they feel "right now" or "at this moment" (and to indicate the intensity of the feeling) or (2) how they "generally feel" (and to record the frequency of the feeling). The test-retest reliability coefficients reported in the manual are consistent with the theoretical premise that trait anxiety is the more enduring characteristic, while state anxiety is transitory; test-retest reli-

ability coefficients for the *state* anxiety measure over a one-hour interval were .33 and .16 for males and females respectively, while the test-retest reliability coefficients for the *trait* anxiety measure for males and females were .84 and .76 respectively. Similar trends were observed in the test-retest reliability coefficients over longer intervals.

19. Take a moment at this juncture to think about how you might go about developng and validating a paper-and-pencil test of personality. Jot down those ideas before continuing to read.
    - What is the purpose of the personality test you've developed? What is it designed to do?
    - Is it to be used to measure traits, types, states, or some combination thereof?
    - Is it to be used to gauge the relative strength of various traits? If so, which traits are to be measured?
    - Is it to be used to distinguish people on the basis of the healthiness of their personality? Is it to be used to distinguish people on the basis of the suitability of their personalities for a particular kind of work? Is it to be used in general research on personality?
    - What kinds of items would your test contain? How would you decide on the content and wording of these items? Would you, for example, rely on a [begin page 292] particular theory of personality in devising these items? Or would you rely on no particular theory, but rather on your own life experiences?
    - In writing your test items, did you use a true/false format or some other format? Will the items of your test be grouped in any particular order?
    - How might you convincingly demonstrate that your test measures what it purports to measure?
20. Like yourself, would-be authors of personality tests have had to struggle with answering questions like these. Some test authors have relied on theories of personality in constructing their test items, while others have steered clear of personality theory and have used more empirical methods. Some test authors have devised forms designed to take a general "inventory" of personality, while others have devised forms to measure specific aspects of it such as the strength of a particular trait. Paper-and-pencil measures of personality differ with respect to the rationale of the measurement model that underlies the test construction. The different models or strategies of test construction have been classified in a number of different ways, and there is even disagreement as to the number of distinctly different models or strategies that exist (Gynther & Gynther, 1976). In the following discussion, we have distinguished four approaches to personality-test construction and have supplemented the discussion with an illustration of at least one test that was developed using each approach.[1] The four approaches are (1) logical or content test construction, (2) factor-analytic test construction, (3) test construction by empirical criterion keying, and (4) the theoretical approach to test construction.

Logical of Content Test Construction

21. One strategy of personality-test construction has been variously referred to as the "logical," "content," "intuitive," or "rational" approach. Here the personality inventory comprises items that logically, intuitively, or rationally seem to belong in the test. Inherent in the logical approach to personality-test construction is the assumption that the test constructor has indeed been logical in the selection of test items. As an adjunct to his or her own logic or intuition, the test developer frequently employs aids such as textbooks, clinical records, experimental data, and conversations with colleagues and others. Suppose you were going to follow the logical or content approach in the construction of a test designed to measure "attitudes toward school." Intuition might tell you that items such as the following should be included:

1.It is important to note that these approaches to test development are not necessarily mutually exclusive; different aspects of a test's development may contain features of each. For example, prospective items for a test could be selected on a rational/logical basis and/or on a theoretical basis. The selected items could then be arranged into scales on the basis of factor analysis. The utility of each item might then be empirically demonstrated.

(Answer TRUE or FALSE)
I enjoy getting up in the morning for school.
I like my teacher(s).
I enjoy seeing my friends at school.
I enjoy the subjects I learn about at school.

22. [begin page 293] Logically, items like those listed would appear to belong in any test that purported to measure attitudes toward school. The first formal efforts to measure personality employed the logical approach to test construction. The Personal Data Sheet (Woodworth, 1917), later known as the Woodworth Psychoneurotic Inventory, was an early test of personality designed to screen World War I recruits for personality and adjustment problems. The test items tapped self-report of fears, sleep disorders, and other problems deemed to be symptomatic of a trait called psychoneuroticism; the greater the number of such problems, the more psychoneurotic the respondent was presumed to be.
23. A content-constructed instrument still in use today is the Mooney Problem Checklist (Mooney & Gordon, 1950). Items on this checklist were developed after evaluating statements of problems obtained from approximately 4,000 high school students, as well as on the basis of counseling interviews and a review of clinical records. The Checklist items relate to emotional functioning in areas such as home and family; boy/girl relations; courtship and marriage; morals and religion; school/occupation; economic security; social skills and recreation; and health and physical development. Respondents are instructed to underline all problems that are of conscern to them and to circle those items that "are of most concern."
24. There are four forms of the instrument, each appropriate for administration to a different age group from junior high school through adult. The test may be administered individually or in groups. Test-retest reliability coefficients for the various forms of the Mooney Checklist have been found to be relatively high, suggesting consistency in the way that test takers perceive their problems over time. The test results have been found to be especially useful in counseling situations where they may be used as a kind of catalyst to treatment and as a pre- and post-measure of the effectiveness of treatment.
25. In gneral the logically constructed test has a certain appeal to test takers since the content is so straightforward and so directly related to the objective of the test. The respondent typically feels more in control of the information he or she is revealing in a content-constructed device than, for example, on an indirect measure of personality such as the Rorshach Inkblots Test. A drawback inherent in the logically constructed test is the case with which the respondent may withhold or distort important information by failing to respond to items honestly. For this reason, a test developer may initially approach a test's development by selecting logically appealing items, but then depart from logic in order to structurally modify the test to detect deceptive responses (see the discussion of "deviant" responses in the last section of this chapter). Another drawback of the logical approach to test construction pertains to the fact that test takers might not necessarily have the insight and perspective on their problems and their assets to accuractely assess themselves.

Factor-Analytic Test Construction

26. Recall from our previous discussion (see Chapter 6) that factor analysis is a data reduction method. Here, we focus on the use of this statistical technique to identify the minimum number of variables or "factors" that account for the intercorrelations in a number of observed phenomena. To illustrate, let's use an example where the "number of observed phenomena" are a multitude of colors. Let us suppose that you [begin page 294] want to paint your apartment but have no idea as to the color that would go best with your "early undergraduate" decor. You go to the local paint stores in your area and obtain free card samples of every shade of color paint known to humanity—thousands of color samples. Let's further suppose you undertook a "factor analysis" of these thousands of color samples — that is, you attempted to identify the minimum number of variables or factors that account for the intercorrelations between all of these colors. You would discover that, accounting for

the intercorrelations, there existed three factors (which might be labeled "primary factors") and four more factors (which might, be labeled "secondary" or "second-order" factors), the latter set of factors being combinations of the first set of factors. Since all colors can be reduced to three primary colors and their combinations, the three primary factors would correspond to the three primary colors, red, yellow, and blue (which you might christen factor R, factor Y and factor B), and the four secondary or second-order factors would correspond to all of the possible combinations that could be made from the primary factors (factors RY, RB, YB and RYB).

27. The color illustration may be helpful to keep in mind as we review how factor analysis can be used in the construction of personality tests. Popular tests such as the Eysenck Personality Inventory, the Guilford-Zimmerman Temperament Survey, and the Sixteen Personality Factor (16 PF) Questionnaire all were derived through the use of factor-analytic strategies. We have chosen the 16 PF to describe in greater detail.

The 16 PF

28. Just as you might have an idea that you wish to analyze all colors into their primary factors, so the notion Raymond Bernard Cattell had when he set out to construct a personality test was the analysis of all personality traits into what might be called primary or "source" traits. Construction of the test items began with a look at the previous research by Allport and Odbert (1936), which suggested that there were more than 18,000 personality trait names and terms in the English language. Of these, however, only about a quarter were "real traits of personality" or words and terms that designated "generalized and personalized determining tendencies — consistent and stable modes of an individual's adjustment to his environment . . . not . . . merely temporary and specific behavior" (Allport, 1937, p. 306). Cattell added to this list some trait names and terms employed in the professional psychology and psychiatric literature and then had judges rate "just distinguishable" differences between all of the words (Cattell, 1957). The result was a reduction in the size of the list to 171 trait names and terms. College students were asked to rate their friends with respect to these trait names and terms, and the factor-analyzed results of that rating further reduced the number of names and terms to 36, which were referred to by Cattell as "surface traits." Still more research indicated that 16 basic dimensions or "source traits" could be distilled. The Sixteen Personality Factor Questionnaire is a test that contains items tapping each of the 16 source traits listed in Table 11-2.

29. The 16 PF was designed for use with junior and senior high school students as well as college and general adult populations. The test was normed on more than 15,000 people. Short-term test-retest reliability estimates have been relatively high, though estimates of long-term test-retest reliability have been considerably lower. The poor long-term test-retest reliability coefficients raise questions concerning the stability of the traits the test purports to measure. Indeed, academicians are by no [begin page 295] means in unanimous agreement that Cattell has discovered *the* "source traits" of personality or that the data from the test yields 16 factors (see Cattell & Krug, 1986)

30.

Table 11-2: Factors of the Sixteen Personality Factor Questionnaire (16 PF)

	Low Score	High Score
Sociable	Reserved	Warm, cooperative
Intelligent	Dull	Bright
Mature	Affected by feelings, undemonstrative	Emotionally stable, calm
Dominant	Obedient, submissive	Assertive
Cheerful	Sober, serious	Enthusiastic
Persistent	Disregards rules, undependable	Conscientious

Table 11-2: Factors of the Sixteen Personality Factor Questionnaire (16 PF)

	Low Score	High Score
Adventurous	Shy	Venturesome
Effeminate	Toughminded, realistic, vigorous	Tenderminded, sensitive
Suspicious	Trusting	Suspicious
Imaginative	Practical, conventional	Imaginative
Shrewd	Forthright, naive	Sophisticated, shrewd
Insecure	Self-assured	Guilt prone, timid
Radical	Conservative, traditional	Experimenting
Self-sufficient	Group-dependent	Self-sufficient, resourceful
Controlled	Uncontrolled	Controlled
Tense	Relaxed	Tense

31. Numerous other forms of this test have subsequently been developed, including an abbreviated version of the test, a "low literate" form for people with third- to sixth-grade reading levels, a taped version for the visually handicapped, and translations into various languages. The philosophy of the 16 PF was extended downward in the construction of various other personality tests, including the Early School Personality Questionnaire (for ages 6 to 8), the Children's Personality Questionnaire (for ages 8 to 12), and the High School Personality Questionnaire (for ages 12 to 18). The use of this series of tests from childhood through adulthood could provide a relatively consistent yardstick by which to gauge personality functioning at various developmental stages.
32. One of the limitations inherent in the factor-analytic technique is the problem of naming factors that have been identified through the statistical analysis. Suppose, for example, you obtained high intercorrelations between the following traits on a test of personality:
    - Depression
    - Anger
    - Fatigue
    - Conservative
    - Bright
33. How would you name the factor that all of these traits seemed to "load on?" Of course there is no rule to naming factors, and the name that you choose might be meaningful for you but not necessarily a name that others would readily accept. [begin page 296] Another limitation inherent in factor-analytic approaches to test construction concerns the controversy that may arise concerning the selection of a particular factor-analytic technique. As has been pointed out by Comrey, Backer, and Glaser (1973, p. 11), "There are many different methods of carrying out a factor analysis. Several different factor analysts can take the same data and come up with as many different solutions . . . all of these different solutions for the same data by different analysts represent interpretations of the original correlation matrix that may be equally correct from the mathematical point of view."

Test Construction by Empirical Criterion Keying

34. Personality-test construction by the strategy of empirical criterion keying may be summed up in the following

simplified way:

1. Create a number of test items that presume to measure one or more traits.
2. Administer the test items to at least two groups of people:
   a. a "criterion group" composed of people you know to possess the trait being measured, and
   b. a control group of people who are presumed not to possess the trait in question.
3. Items that discriminate in a statistically significant way with respect to the criterion and control groups are retained, while those items that do not discriminate between the two groups are discarded.

35. This method of test construction is referred to as "empirical" because only those items that demonstrate an actual (empirical) relationship between the test item and the trait in question are retained. It is called "criterion keying" since each item of the test is keyed to a criterion, the criterion being related to the particular trait in question. Since test construction by means of empirical criterion keying always involves the comparison of at least two groups of people (one group possessing the trait, the other not), this approach to test construction has also been referred to as the method of "contrasted groups." Two well-known personality tests developed by this method are the Minnesota Multiphasic Personality Inventory (MMPI) and the California Psychological Inventory (CPI).

The MMPI

36. Conceived in the 1930s by psychologist Starke R. Hathaway and psychiatrist/neurologist John C. McKinley as an aid in assessing the mental health of patients seen in medical practice, a test first called the "Medical & Psychiatric Inventory" was renamed when published by the University of Minnesota Press in 1941 as the "Minnesota Multiphasic Personality Inventory" (MMPI). Hathaway (Figure 11-3) reminisced that "It was difficult to persuade a publisher to accept the MMPI" (Dahlstrom & Welsh, 1960, p. vii), though the test quickly gained popularity among psychologists and has become the single most widely used objective personality test (Lubin, Larsen, & Mattarazzo, 1984).

37. The MMPI consists of 550 statements to which the examinee responds "true" or [begin page 302] "false. In one form of the test, statements are printed on cards, and a third category, "cannot say," is included (Dahlstrom, Welsh, & Dahistrom, 1972). For the group-administered version ofthe test, all unanswered items in the test booklet are scored in the "cannot say" category. The MMPI may he used with persons 16 or older who have at least a sixth-grade education (or an IQ of 80). Tape-recorded and foreign-language versions of the inventory have also been constructed.

38. As reported by the test authors (Hathaway & McMinlcy, 1940, 1951), research preceding the final selection of items involved the study of psychiatric textbooks, psychiatric reports, and previously published personality-test items. The test items that were ultimately selected reflected 26 content categories, including general health, family issues, religious attitudes, sexual identification, and psychiatric symptomatology (Hathaway & McKinley, 1951). These items were then presented to both criterion groups and a control group. Lanyon and Goodstein (1971, p. 76) described the normal control group as follows: " . . . 1500 control subjects were drawn from hospital visitors, normal clients at the University of Minnesota Testing Bureau, local WPA workers, and general medical patients." The criterion group was eight clinical groups of psychiatric in-patients from the University of Minnesota hospital. Those items reflecting statistically significant differences between the responses of the clinical criterion group and the control subjects were retained. Analysis of the clinical groups' responses in contrast to the control group made it possible to develop "scales" that corresponded to each disorder. The MMPI consists of eight clinical scales that were developed in this fashion (as well as two additional scales, Masculinity-Femininity and Social Introversion-Extraversion, that employed nonpsychiatric Criterion groups in their development). A brief description of each criterion group used in the development of the ten clinical scales appears in Table 11-3. More detailed information concerning the construction and validation of the MMPI can be found in Welsh and Dahlstrom (1956).

39. In addition to ten clinical scales, the MMPI contains three "validity scales" that were designed to serve as indicators of factors such as the operation of response Sets, attitudinal factors, or misunderstanding of directions that may influence test results. These include the L scale (sometimes referred to as the "Lie" scale), the F scale (sometimes referred to as the "Infrequency" scale), and the K (correction) scale, The L scale contains 15 items that arc somewhat negative but that apply to most people, such as "I do not always tell the truth," or "I gossip a little at times" Dahlstrom et al., 1972, p. 109). The preparedness of the examinee to reveal *anything* negative about himself or herself will he called into question if the score on the L scale does not fall within certain limits. The 64 items on the F scale (1) arc infrequently endorsed by members of nonpsychiatric populations (that is, normal people) and (2) do not fit into any known pattern of deviance. A response of "True" to an item such as the following would be scored on the F scale: "It would be better if almost all laws were thrown away" Dahlstrom et al., 1972, p. 115). An elevated F score may mean that the respondent did not take the test seriously and was just responding to items randomly. Alternatively, the individual with a high F score may be a very eccentric individual or someone who was attempting to "fake bad." Malingerers in the armed services, people intent on committing fraud with respect to health insurance, and criminals attempting to "cop a psychiatric plea" are some of the groups of people who might he expected to have elevated F scores on their profiles.
40. Like the L score and the F score, the K score is a reflection of the frankness of the [begin page 303]

Table 11-3: The Clinical Criterion Groups for MMPI Scales

	Scale	Criterion Group
1.	Hupochondriasis (Hs)	The criterion groupfor this scale was patients who showed exaggerated concerns about their physical health.
2.	Depression (D)	The criterion group for this scale was clinically depressed patients; unhappy and pessimistic about their future.
3.	Hysteria (Hy)	The criterion group for this scale included patients with conversion reactions.
4.	Psychopathic deviate (Pd)	The criterion group for this scale was patients who had had histories of delinquency and other antisocial behavior.
5.	Masculinity-femininity (Mf)	The criterion group for this scale was Minnesota draftees, airline stewardesses, and male homosexual college students from the University of Minnesota campus community.
6.	Paranoia (Pa)	The criterion group for this scale was patients who exhibited paranoid symptomatology such as ideas of reference suspiciousness, delusions of persecution, and delusions of grandeur.
7.	Psychasthenia (Pt)	The criterion group for this scale was anxious, obsessive-compulsive, guilt-riddent, and self-doubting patients.
8.	Schizophrenia (Sc)	The criterion group for this scale was patients who were diagnosed as schizophrenic (various subtypes)
9.	Hypomania (Ma)	The criterion group for this was patients, most diagnosed as manic-depressive, who exhibited manic symptomatology such as elevated mood, excessive activity, and easy distractibility.
10.	Social introversion (Si)	The criterion group for this scale was college students who had scored at the extremes on a test of introversion-extraversion.

test taker's self-report. An elevated K score is associated with defensiveness and the desire to present a favorable impression. A low K score is associated with excessive self-criticism, desire to detail deviance, and/or desire to fake bad. A "True"

response to the item "I certainly feel useless at times" and a "False" response to "At times I am all full of energy" Dahlstrom et al., 1972, p. 125) would be scored on the K scale. The K scale is sometimes used to "correct" scores on five of the clinical scales; the scores are statistically corrected for an individual's overwillingness or unwillingness to admit deviancy.

41. The MMPI may be computer-scored, even computer-interpreted; computerized reports range in detail from simply a numerical score for each scale to long and detailed narrative reports. Whether computer-scored or hand-scored, the raw test scores are converted to standard scores that have a mean of 50 and a standard deviation of 10. Standard scores of 70 or greater on the clinical scales arc considered to indicate a problem that must be investigated. For example, a score of 88 on the Depression scale would suggest an extremely depressed and pessimistic individual, while an 85 on the Hypochondriasis scale would be reflective of an individual who has frequent physical [begin page 304] complaints and excessive concern with bodily functioning. Interpretations on the MMPI are generally made, however, on the basis of the entire test pattern or profile, not on the basis of a score on any one scale.

42. In contemporary usage, MMPI scales are referred to by number rather than their original name. This is so because literal interpretation of the names of the scales would be inaccurate. A high score on the Schizophrenia (Sc) scale does not necessarily mean that the test taker would be diagnosed as schizophrenic; the test taker might well be diagnosed as suffering from some other form of psychosis. It might even be possible for an individual with an elevated Sc scale to be diagnosed as normal. In practical usage, the scales are viewed as continuums with respect to particular personality traits associated with the criterion group the scale was based on. For example, a person scoring high on the Paranoia scale would be regarded as high in suspiciousness, feelings of persecution, and distrust. Note that this use is inconsistent with the purpose of the test as conceived by the test authors (to be an instrument used for classification and differential diagnosis).

43. Since its inception in the early 1940s, the MMPI has been used in clinical and research settings with a variety of individuals. The consequence of decades of use and research is a proliferation of new MMPI scales based on the test patterns of various populations. Over 400 new MMPI scales have been devised since the test's publication, and there may well be another 400 new scales by the time this textbook goes into its second edition. Researchers have examined and compared not only the MMPI responses of normals and persons with various psychiatric diagnoses, but also the test protocols of members of more "offbeat" populations as well. Included in the latter category is research with members of groups as diverse as a serpent-handling religious cult (Tellegen et al., 1969), castrated males (Yamamoto & Seernan, 1960), submarine school dropouts (King, 1959), and civilians selected for isolated northern stations (Wright, Sister, & Chylinski, 1963). Several encyclopedias of MMPI profiles—referred to in the profession as "cookbooks"—are available for use by clinicians (for example, Hathaway & Meehl, 1951; Dahlstrom, Welsh, & Dahlstrom, 1972; see also, Swenson, Pearson, & Osborne, 1973; Butcher, 1979; Dahlstrom, Lachar, & Dahlstrom, 1986).

44. Critics of the MMPI have cited limitations relating to its construction or use. In light of the widespread use of this instrument, the original normative sample has been criticized as being deficient in terms of size and the represcntativcness of the general population. Other criticism has been leveled at the sheer age of the norms; as Dahlstrom et al. pointed out (1972, p. 8), "Each subject taking the MMPI, therefore, is being compared to the way a typical man or woman endorsed those items. In 1940, such a Minnesota normal adult was about thirty-five years old, was married, lived in a small town or rural area, had had eight years of general schooling, and worked at a skilled or semiskilled trade (or was married to a man with such an occupational level)." Dahistrom and colleagues are currently involved in a large-scale project designed to update the entire MMPI (Greene, 1985).

45. In October 1983 a new set of MMPI norms for normal adults was published. The norms were developed by a group of researchers from the Mayo Clinic of Rochester, Minnesota (Colligan, Osborne, Swenson & Offord, 1983) and included MMPI responses from 1,408 normal subjects (people who were not under the care of any health—care professional), ranging in age from 18 through 99 years and living in [begin page 305] the same general geographic area as the sample used by Hathaway and McKinley (1940). The results indicated that peo-

ple living in the 1980s tended to have elevated MMPI profiles in contrast to a comparable sample of people living in the 1940s (and the increases tended to be greater for men than for women). Colligan, Osborne, Swenson, and Offord (1984) offered two alternative (though not mutually exclusive) explanations for this finding: (1) people in the 1980s may be under more psychological and physical stress than were people in the 1940s, and (2) changes in response patterns may be due to changes in societal mores and perceptions. Colligan et al. (1984) interpreted their findings as being of practical as well as statistical significance, and they cautioned that "clinicians take a somewhat more conservative approach to profile interpretation with more careful consideration of the impact of age and sex on profile configuration."

46. At this writing, published experience with the updated norms has been scarce and the byword with respect to their use seems to be "caution." Miller and Streiner (1986) examined MMPI data for 2,083 people using the original norms and those from Colligan et a!. (1983). These researchers noted sufficient lack of comparability between the two sets of norms to caution that the newer norms not be used independently — but rather in conjunction with the original norms — until the clinical relevance of the differences are determined. In reviewing the work of Colligan et al., Greene (1985) reached a similar conclusion:

> The real issue is whether the use of contemporary MMPI norms results in more accurate predictions. . . . . In the empirical spirit with which the MMPI was developed, it seems that we must wait to see the data. Until then, all we can say is that contemporary adults cam somewhat different scores on the various MMPI scales than the adults of the 1930s. Hopefully, such research will be forthcoming so we can begin to evaluate the issue of interpretive accuracy. (p. 109)

47. Whether the new or original norms are employed, it has always been important for the test user to temper interpretations made from the test data with reference to the limitations of the population used as a normative sample. Thus, for example, Colligan et al. (1983) pointed out that their norms would not be appropriate for use with ethnic minority groups, and they encouraged the development of norms expressly designed for such use. In this vein, it would also be important to learn more about the applicability of the new norms to other geographic areas and groups (Miller & Streiner, 1986).

48. From the standpoint of test construction, the MMPI has been criticized for having some of the same items used in the different scales. The result of this structural redundancy is that some of the scales are highly correlated with one another. If the instrument is to be used as a tool of differential diagnosis, it would be preferable for the scales not to correlate with one another. There also exists some confusion as to the meaning of a low score on the clinical scales; while the meaning of an elevated score on a clinical scale may be clear, Wiggins (1973) has pointed out that given the way the MMPI was constructed, the meaning of a significantly low score is unclear. Other frequently cited limitations of the MMPI have to do with the ready availability of its computerized scoring and the possible misuses inherent in any computer—generated test reports (more on that subject in Chapter 20); the offensiveness of some of the [begin page 306] questions to some test takers (Butcher & Tellegen, 1966; Gallucci, 1986), particularly questions related to sex, religion, bladder and bowel functions; and the length of the test (which has been viewed by some as excessive). One attempted remedy for the latter criticism has been the development of short forms of the test — forms that contain only a sampling of items from each of the scales and a fraction of the original total of items (Stevens & Reilley, 1980). In general, however, the short form of the MMPI seems not to have lived up to its promise in terms of psychometric soundness or clinical utility (Helmes & McLaughlin, 1983; Hart, Lutz, McNeill, & Adkins, 1986).

49. In spite of its limitations, the MMPI remains the most used and researched of all the existing personality inventories. Its use as a tool to describe aspects of one's personality has found application in a variety of clinical, counseling, educational, worksite, and research settings. The large and ever-expanding literature on this test provides a library of reference material to MMPI users. Although the test is seldom used in the way it was designed to be used — as a measure of differential diagnosis — it is no doubt of value to clinicians in their everyday work with psychiatric patients; MMPI results provide insight into the extent and magnitude of patients' problems. The test results are frequently viewed as tentative hypotheses about the examinee's psychopathology that await clarification and validation from other sources of data (see Graham, 1977).

California Personality Inventory

50. Another test constructed by the method of empirical criterion keying is the California Personality Inventory (CPI). This test is a "kissing cousin" of the MMPI in that many of its items were drawn directly or revised from the MMPI. In contrast to the MMPI, which was developed to assess maladjustment, the CPI was designed for use with normal populations aged 13 and older, and its scales emphasize more positive and socially desirable aspects of personality than do the scales of the MMPI.
51. The CPI is available from its publisher in its original form (Gough, 1956) or in a revised edition (Gough, 1987). The original edition of the test contains 18 scales, which may be grouped into four categories depending upon whether they primarily measure interpersonal effectiveness (including measures of poise, self-assurance, and self-acceptance), intrapersonal controls (including measures of self-control and tolerance), academic orientation (including measures of achievement potential), or general attitudes toward life (including measures of conformity and interests). Eleven of the personality scales were empirically developed based on the responses of subjects known to display certain kinds of behaviors. Factors such as course grades, participation in extracurricular activities, and peer ratings were used in selecting the criterion groups (see Gough, 1957, 1975). Four scales, Social Presence, Self-Acceptance, Self-Control, and Flexibility were developed through internal-consistency item-analysis procedures. Also built into the inventory were scales designed to detect response sets for faking favorable and bad impressions.
52. The 1987 revision of the test retained the 18 original scales with only minor changes in content and some rewriting or deletion of items to reduce sexist and/or other bias. Two new scales were added, Independence and Empathy, bringing the total number of scales contained in the 1987 revision of the test to 20. The 20 scales can [begin page 307] be organized with reference to three independent themes derived from factor-analytic studies: (1) interpersonal orientation, (2) normative orientation, and (3) realization. Like its predecessor, this edition of the CPI may be hand- or computer-scored.
53. Normative data for the original version of the CPI was obtained from the testing of 6,000 males and 7,000 females of varying age, socioeconomic status, and place of residence. Test-retest reliability coefficients reported in the CP1 manual range from .55 to .75. Included in the manual is research concerning the feasibility of making various kinds of predictions with the test scores; predictions ranging from the probability of delinquency or dropping out of school to the probability of success among those in training for various occupations (such as dentists, optometrists, accountants, and so on). An abbreviated form of the original edition of the CPI has been found to correlate in the range of .74 to .91 with the original (Armentrout, 1977).
54. Like the MMPI, studies reporting on new scales for the CP1 can be found in the professional literature. For example, Gough (1985) reported on the development of a "Work Orientation" (WO) scale for the CPI. The WO scale is composed of 40 items that were found to be correlated with criterion measures such as a job performance rating. It was reported that high scorers on WO were dependable, moderate, optimistic, and persevering.
55. Also like the MMPI, the widely used CPI has its critics. The test has been criticized for the relatively high intercorrelations between the scales and for relatively low coefficients of reliability (Megargee, 1972). Other criticism is leveled at the methods used to establish some of the criterion groups. Still, the test is a widely used, widely researched instrument that has proven its value as a useful tool with normal subjects. Whether the 1987 edition will prove more psychometrically sound than its predecessor is a question that will be answered as published reviews become available.

The Theoretical Approach to Test Construction

56. Some personality tests are closely tied to a particular theory of personality, and all of the items on such a test

are designed to measure traits or states presumed to exist on the basis of that theory. For example, a personality test constructed within a psychoanalytic framework might have items on it designed to assess id, ego, and superego functioning. Some of the personality inventories that have employed the theoreti cal approach or strategy in their construction include the Myers-Briggs Indicator (based on the personality typology set forth by Carl Jung, see Myers & Briggs, 1943/1962; Myers & McCaulley, 1985; and Briggs, Myers, & Saunders, 1987), the Personality Research Form (based upon Henry Murray's work; see Chapter 1987), the Personality Research Form (Jackson, 1984) which was based on Henry Murray's work; see the Close-up in Chapter 7), and the Edwards Personal Preference Schedule (EPPS), which we describe below.

57. **The Edwards** Personal **Preference Schedule (EPPS)**

58. The EPPS (Edwards, 1953) is a personality inventory based on the theory of personality presented by Henry Murray in *Explorations in Personality* (1938). *Explorations* pre [begin page 309] sented a complex but academically elegant theory of personality that not only introduced new concepts (such as "press," "regnancy," and "serial programs"), but also provided the impetus for renewed study of more traditional concepts.[1] In the latter context, for example, Murray explored the parameters of the word "need," defininig it, writing about its consequences, and detailing how various needs could be inferred. According to Murray, needs could be either primary or secondary, overt or covert, focal or diffuse, proactive (determined from within) or reactive (occurring in response to or as a result of some environmental event), and modal (done for the sheer pleasure of doing) or effect (done to effect some result). The list of needs originally published in *Explorations* appears in Table 11-4.

59. Edwards selected 15 of the needs listed by Murray and constructed items designed to assess each of those needs. He next conducted research designed to assess th social desirability of each of the items he wrote. Items assessing different needs tha were found to be generally equivalent with respect to social desirability were then placed into pairs (Edwards, 1957a, 1957b, 1966). For example, a pair of statcments deemed to be approximately equivalent with respect to social desirability might be

    - I feel depressed when I fail at something.
    - I feel nervous when giving a talk before a group.

Table 11-4: List of Needs Presented in Murray (1938)[a]

	Need	Definition (the need to . . .)
1.	Abasement	submit passively accept blame, injury, criticism, or punishment admit inferiority, error, wrongdoing, or defeat
2.	Achievement	accomplishment and excel rival and surpass others
3.	Affiliation	please, win affection of, and remain loyal to a friend draw near to others
4.	Autonomy	be independent, unattached, and defy convention

1."Press" is a construct Murray used to refer to significant determinants of behavior that lie outside of the person. **it** is a term used in contrast to the construct "need," which refers to the significant determinants of behavior from ***within.*** "Regnancy" is a concept Murray used to link physiological *(brain)* processes to psychological processes (see Murray, 1938, p. 45). "Serial program" is a term used to refer to a set of subgoals that must be reached before some final goal can be attained.

Table 11-4: List of Needs Presented in Murray (1938)[a]

	Need	Definition (the need to . . .)
5.	Counteraction	make up for failure with renewed efforts overcome a weakness of a fear
6.	Defendance	protect or shield from blame, criticism, assault, and humiliation
7.	Dominance	influence or direct others by authority or force
8.	Exhibition	influence others by entertaining, shocking, exciting, or enticing them
9.	Harm avoidance	avoid physical injury, pain, illness, and death
10.	Infavoidance Nurturance	help, support, protect, comfort, nurse,heal, and give sympathy
11.	Order	achieve balance, precision, and organization
12	Play	participate in games, sports, other pleasurable activities act sheerly for "fun"
13.	Rejection	separate or snub a person deemed to be inferior in some way
14.	Sentience	seek and enjoy sensuous activities
15.	Sex	have erotic relationships and sexual outlets
16.	Succorance	be nursed, supported, sustained, protected, advised, forgiven, consoled have a steadfast, sympathetic supporter
17.	Understanding	question, theorize, analyze, speculate, generalize

a.We have abbreviated the definitions of these needs for the puposes of this tabular presentation. Consult Murray (1938, pp. 152-226) for complete definitions.

60. Edwards constructed his test of 210 pairs of statements in a way such that respondents were "forced" to answer "True" or 'False" or "Yes" of "No" to one of two statements that were equivalent in terms of social desirabilit This "forced-choice" technique represented an attempt to control for respondents' attempts to fake good or fake bad. Note also that each of the two statements above, like each of the statements in every pair of EPPS statements, is keyed to a different need in Murray's system. Endorsement of an item keyed to one scale in essence serves to reject an item keyed to an alternative scale. The score that is computed for each ofthe EPPS needs or scales thus represents the intensity of a particular need *in relation to* the intensity of the individual respondent's other needs. EPPS scores are, in psychometric jargon, *ipsative* in nature; the scores do not represent the strength of the need in absolute terms but rather the strength of the need in relation to the individual respondent's other needs. To elaborate, ipsative scoring allows for comparison of personality characteristics exhibited by an individual examinee with respect only to that examinee and does not allow for comparison between examinees. Stated another way, such scoring is useful in intra-individual comparison and not in inter-individual comparison. For example, on the basis of personality inventory data derived by means of ipsative scoring, it might be appropriate to make a statement like "John's need for achievement is higher than his need for succorance." However, it would be inappropriate on the basis of such data to compare any of John's needs to those of another person's as in a statement like, "John's need for achievement is higher than Jane's need for achievement."

61. In addition to the use of the forced-choice format, Edwards built other precautionary measures into the EPPS in an effort to detect and/or minimize the effects of faking, response sets, and other factors that would threaten the validity of the obtained scores. A Consistency scale is designed to check on the consistency of the exam-

ince's responses. This scale consists of 15 identical items that are repeated in various places throughout the inventory.

62. As a further measure of consistency, a "stability" score may be obtained; this score is equal to the correlation coefficient that describes the relationship between two halves of the test (odd and even scores in the 15 scales).

63. Normative data for the EPPS were initially gathered on a sample of 760 male and 749 female college students from 29 campuses throughout the country and approximately 9,000 men and women from the general adult population. Subsequently, data based on the test results for 559 male and 986 female high school students were added. Test-retest reliability coefficients for the 15 scales based on one-week intervals were found to range between .74 and .87. Internal-consistency measures resulted in split-half reliability coefficients ranging from .60 to .87 with a median of.78. Interpretation of these findings is complicated because the test contains repeated items. In general, the test is viewed as being within acceptable standards of test-retest and interitem reliability; the objection many reviewers have raised concerns the lack of compelling validity data (Heilbrun, 1972). Additionally, questions have been raised concerning the extent to which the forced-choice format of the test does indeed eliminate the social desirability response set from affecting scores (Heilbrun & Goodstcin, 1961a, 1961b; Rorer, 1965; Wiggins, 1966). Reviewers have also questioned the appropriateness of converting ipsative scores into normative percentiles. Still, in spite of these limitations, the EPPS remains a widely used and widely researched instrument. [begin page 312]

Some Problems and Issues in Assessing Personality

64. Many personality assessment instruments of the paper-and-pencil variety rely heavily either on the self-report of the assessee or on a rating made by the assessor(s). We conclude this chapter by considering some limitations inherent in the use of such techniques.

65.

Limitations of Self-Report Techniques

66. Were employers to faithfully rely on job applicants' representations concerning their personality and their suitability for a particular job, they might well receive universally glowing references — and still not hire the most suitable personnel. The problem here is that many of the applicants might be expected to try to "fake good." Were local draft boards to faithfully rely on draft resisters' representations concerning their personality and lack of suitability for military service, few resisters would be inducted into military service. The problem here is that many of the resisting registrants might be expected to try to "fake bad." One problem inherent in assessing personality, a problem particularly acute with respect to self-report methods, is the problem of faking or "impression management." We now discuss this problem as well as the related problem of response sets in taking tests.

67. **Impression management**. After Goffman (1959), Braginsky, Braginsky, & Ring used the term "impression management" to refer to the fact that:

> we can and generally do manage our expressive behavior so as to control the impressions that others form of us. Through selective exposure of some information **(it** may be false information) consistent with the character we mean to sustain for the purpose of an interaction, coupled with suppression of information incompatible with that projection of self, we establish a certain definition of ourselves that we attempt to maintain throughout the interaction episode. (p. 51)

68. In essence, we all try (to varying degrees) to "manage impressions" of ourselves to others. According to Goffman (1959), an individual may want his audience "to think highly of him, or to think that he thinks highly of them, or to perceive how in fact he feels towards them, or to obtain no clear-cut impression; he may wish to ensure sufficient harmony so that the interaction can be sustained, or to defraud, get rid, confuse, mislead, antagonize, or insult them" (p. 3). In many personality assessment situations, the examinee may be highly motivated to manage a favorable impression of himself — to "fake good" as it were. For example, if the data

from the assessment will be used to determine if the individual is admitted to college or considered for promotion, the temptation to present oneself in as favorable a light as possible is strong. Conversely, there are other situations in which an individual may be tempted to "fake bad" to achieve some desired result. A chronic mental patient who prefers the environs of a mental hospital to the outside world may attempt to "fake bad" on a personality test if he or she is led to believe that the data from that test may result in discharge from the hospital. Criminals may attempt to "fake bad" on personality tests in order to be declared on the basis of insanity. [begin page 314]

69. Another variation ot impression management concerns **not** the desire to take good or bad, but simply to manage the impression—good, bad, or indifferent—that the actor believes the audience is expecting. This point has been elaborated on by Goffman (1959):

> Doctors who are led into giving placebos, filling station attendants who resignedly check and recheck tire pressures for anxious women motorists, shoe clerks who sell a shoe that fits but tell the customer **it** is the size she wants to hear — these are cynical performers whose audiences will not allow them to be sincere. (p. 18)
>
> If a baseball umpire is to give the impression that he is sure of his judgment he must forego the moment of thought which might make him sure of his judgment: he must give an instantaneous decision so that the audience will be sure that he is sure of his judgment. (p. 30)

70. In the personality assessment situation, some examinees may respond in a way that they believe will confirm or deny the expectations of the examiner.

71. **Response sets.** A *response set* refers to the tendency to respond to a question in some characteristic manner regardless of the content of the question. For example, some individuals are more apt to answer "Yes" or "True" than "No" or "False" on short-answer tests. Psychologists have distinguished several different types of response sets. One type has been referred to as a "socially desirability response set." This refers to examinees' tendency to respond in such a way as to present themselves in the most socially acceptable way in order to manage a favorable impression. Another response set has been referred to as "acquiescence." The acquiescent responder tends to agree rather than disagree on true/false, yes/no, and agree/disagree types of tests. At the other end of the continuum from the acquiescence response set is the nonacquiescence response set characterized by a test taker who exhibits a tendency to disagree.

72. A third type of response set has been referred to as "deviance," the tendency to give unusual or uncommon responses to test items. As we have seen, some personality tests contain items that are part of the test for the express purpose of identifying the respondent who has a tendency to give unusual or uncommon responses. Thus, for example, a "True" response to an item like "I recently vacationed in downtown Beirut" might lead the test scorer/interpreter to raise some questions about the findings: Did the test taker understand the instructions? Did the test taker take the test seriously? Did the test taker respond "True" to all of the items on the test? Did the test taker respond randomly to items on the test? Analysis of the entire protocol might help to provide additional answers.

Problems Attendant on Rating Scales

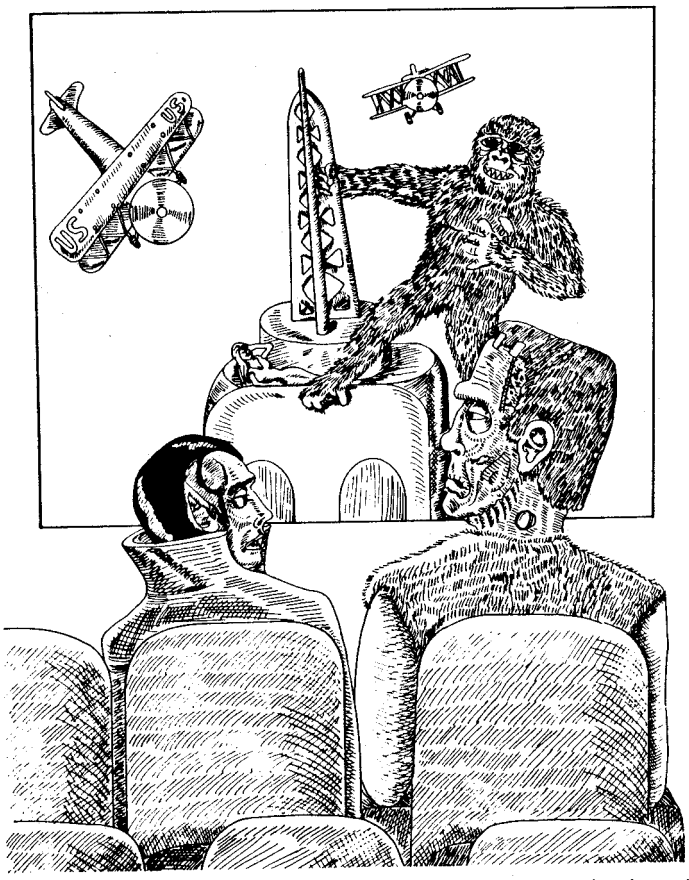
73. Some measures involve procedures where one individual observes and evaluates someone else. The considerations that need to be kept in mind in such a situation have already been touched on in Chapter 6, in the section on bias. Here we review and expand on that discussion with reference to rating scales and raters.

74. **The rater.** Mrs. Jones, a third-grade teacher, had Alvin Farkas's brother Fred in her class five years ago. She remembers Fred to be an excellent, all-around student, and he [begin page 315] was every bit the "teacher's pet." Will this fact enter into Mrs. Jones's judgment when she evaluates Alvin? Maybe it shouldn't, but few people would be surprised if it did. Teachers are human, too, and past experience, attitudes, hopes, and fears are some of the factors that might enter into — and bias — their ratings. In the situation of two brothers, a *halo effict* may be operative with respect to Mrs. Jones's ratings of Alvin; the Farkas name has generated so much

goodwill in the mind of Mrs. Jones that Alvin may be perceived as "capable of doing no wrong." More broadly, a halo effect is a type of error in rating wherein some single attribute or combination of attributes biases judgments or ratings regarding other attributes.

75. Many raters have an investment in the people they rate. Thus the school, industrial, or organizational instructor who has spent six months teaching a particular course has a personal investment in the ratings of the students; it doesn't look well for the instructor if too many of the students fail on some final measure of outcome. Thus, situations might exist where the rater's own self-interests are at odds with — and may interfere with — a fair and unbiased rating (Figure 11-4).

Figure 11-4 ***A halo effect.***



"Monsters and screaming women have always worked for me; I give it 'thumbs up,' Roger."

76. [begin page 316] Numerous other factors may contribute to bias in a rater's ratings. The rater may feel competitive with, physically attracted to, or physically repulsed by, the subject of the ratings The rater may not have the proper background experience and trained eye needed for the particular task The rater's judgments may be limited by his or her general level of conscientiousness and willingness to devote the time and effort required to do the job properly. The rater may harbor biases concerning various stereotypes. The rater may have a tendency to rate highly (a *leniency* or *generosity error*), a tendency to rate harshly (a *severity error*) or a tendency to rate everyone at some point around the midpoint of the rating scale (an *error of central tendency*). Subjectivity based on the rater's own subjective preferences and taste may also enter into judgments; Bo Derek was a perfect "10" for Dudley Moore in the film by the same name, though others may find this woman less than perfect to greater or lesser degrees.

77. One attempt at controlling for raters' biases involves educating raters as to the types of biases that exist and the ways in which they may interfere with the accuracy of ratings. Another attempt at controlling for raters' biases has been to provide training sessions for raters. Such training sessions afford the opportunity for raters to (1) clarify terminology to increase the rehability of their ratings (for example, terms such as "satisfactory" and "unsatisfactory" may be construed differently by different people), (2) to obtain practice in observing and rating others, and (3) to compare their ratings with those of experienced raters. Research has demonstrated the effectiveness of rater-training programs (see Bernardin, 1978).

78. **The instrument.** By now you have already acquired much firsthand experience with a small sample of the various rating systems that have an impact on everyone's academic, business, and social life. Some of these familiar rating systems are as follows:

- "X" is a rating of a motion picture in which there is rather graphic presentation of sexual and/or violent material. When you were younger, such a rating prohibited you from entering the theater.
- "****" is a rating used in many travel guidebooks to denote the highest quality accommodations and dining.
- "/ / /" is something your friend Jane uses in her little black book next to the names of men she has dated to distinguish those who have conformed to her highest specifications in terms of mental, physical, and related attributes.
- "D" is the rating your instructor gave you as a final grade in your economics course. This is why you decided to shun the business world and become a psychology major.

79. Rating scales are used to classify, to determine eligibility, and to predict effectiveness. Ratings are also useful in the process of validating a particular test because they provide a convenient criterion against which test scores can be compared. Thus, for example, scores on a paper-and-pencil "Work Effectiveness Test" might be compared against a supervisor-filled-out "Work Effectiveness Rating Scale." Given that rating scales may play a large part in terms of individuals' academic and business futures, a word about the construction of these types of instruments is in order. Rating scales (like tests) with the same name may be focusing on vastly different things. For [begin page 317] example, one "Worker Ettectiveness Rating Scale" might contain items on it that relate mostly to a worker's creativity and initiative while another "Worker Effectiveness Rating Scale" might contain items that focus more on the worker's ability to cooperate with fellow workers. Thus, a rating scale, like a test, must be judged by its validity for use in a specific context and for a specific purpose, not by its name.

80. Rating scales come in many varieties. There are rating scales to rate the self and there are rating scales to rate others. Some rating scales require the rater to make careful observations (such as "Does the patient make his bed?"), while others require the rater to make evaluations and express opinions (such as "How well does the patient get along with the other patients on the ward?"). Rating scales vary in format; in general, they are either alphabetical, numerical, graphic, or of the forced-choice variety. The alphabetical rating scale uses letters keyed to some type of description as the rating system. The letter-grade rating system of A to F (excluding the letter "E") is an example of an alphabetical rating system as is the movie industry's "G," "GP," "R," and "X" rating system. A numerical format, as its name implies, employs numbers keyed to descriptions (for example, 0 = the least, 100 = the most). With graphic rating scales, the rater's task is to check off or mark some line, number, letter, or point on a figure. One widely used rating scale of the graphic variety is called the "semantic differential." Developed by Osgood, Suci, & Tannenbaum (1957), the semantic differential is a technique that employs bipolar adjectives and a seven-point rating scale (Figure 11-5). The examinee is instructed to respond to the presentation of some idea, concept, or issue by checking off one of the seven spaces between the bipolar

adjectives. Forced-choice rating scales contain two or more descriptions from which the rater must select the most appropriate. Forced-choice ratings are useful in self-rating instruments and in other situations where there might exist a special need to minimize errors in ratings as a function of bias or response sets.

81. One form of rating that requires special discussion is ranking. In essence, ranking entails an ordering of ratings with reference to some bipolar variable (such as highest-lowest, most-least, or strongest-weakest). Like forced-choice procedures, ranking

Myself

Warm	____	____	____	____	____	____	____	Cold
Tense	____	____	____	____	____	____	____	Relaxed
Optimistic	____	____	____	____	____	____	____	Pessimistic
Frugal	____	____	____	____	____	____	____	Extravagant
Weak	____	____	____	____	____	____	____	Strong
Brooks Brothers Suit	____	____	____	____	____	____	____	Hawaiian shirt

Figure 11-5. The Semantic Differential. This is a technique that can be applied to the rating of people, products — most anything. Here the rate is being asked to place checkmarks at the point in the continuum that best describes himself or herself.

82. [begin page 318] procedures may force the rater to make fine distinctions and to identify positive as well as negative choices. The *paired-comparison method* of ranking entails individually comparing every item to be ranked with every other item to be ranked. Another ranking method entails comparing each item or individual to be ranked according to some preestablished standard or criterion. Rankings generally provide little information in and of themselves. For example, what does it mean to be ranked fifth in a class of gifted children? To make such a ranking meaningful, we would have to know more (such as measures of central tendency and variability, the method by which the ranking was derived, and so forth).

83. Inter-rater reliability tends to increase as a function of the clarity and specificity with which terms on a particular rating scale are defined. Thus, all other things being equal, a random group of raters will probably exhibit less agreement on a rating scale that merely has categories such as “Excellent,” “Good,” “Fair,” and “Poor” than on one where clear behavioral referents to these terms are specified.

Figure 11-3 *Starke Rosecrans Hathaway (1903—1984).*

84. “With his consistent emphasis on objectivity and eclecticism, his insistence on data in preference to inference, his commitment to collegiality and scientific openness, and his scholarly respect for both the biological and psychological dimenions of human personality, Starke Hathaway has an assured place as one of the founders of modem clinical psychology”—so read the obituary for the co-developer of the MMPI, a test that in “its many versions and in nearly 50 languages . . . has been employed in hundreds of different research uses and practical applications for nearly five decades” (Dahistrom, Meehl, & Schofield, 1986).

85. Born in Michigan, Hathaway spent much of his youth in Marysville, Ohio. He earned his bachelor’s and master’s degree at Ohio University in Athens and his Ph.D. at the University of Minnesota. Through the efforts of a psychiatrist at the University Medical School, J. Charnley McKinley, Hathaway was granted a position in the neuropsychiatry division. The two men would subsequently collaborate in the development of the Minnesota Multiphasic Personality Inventory MMPI (Hathaway & McKinley, 1940).

86. Dahlstrom, Meehi, & Schofield (1986, p. 835) remind us that "Hathaway's identification with the MMPI overshadowed his equally important contributions as a teacher and therapist. He was a master clinician to whom medical colleagues frequently referred puzzling or difficult patients for diagnosis or treatment. The more difficult and challenging the case was, the more intense, persistent, and innovative were Hathaway's efforts. He rarely failed to achieve a significant result. . . . Many of Hathaway's treatment methods anticipated the behavioral interventions of today, including such methods as mild aversive shock, suggestion and hypnosis, modeling, and habit retraining."

87. Hathaway's long list of lifetime achievements includes being recipient of the American Psychological Association's award for Distinguished Contributions for Applications in Psychology. Hathaway retired from the University of Minnesota in 1971 and he died in his home in Minneapolis on July 4, 1984.

> "When I came to the University hospitals in about 1937 and began to work with patients, I started to change from a physiological psychologist toward becoming a clinical psychologist. As we went on grand rounds, I, with my white coat and newly developing sense of role, expected that the medical staff would want the data and insights of a psychologist. I still remember one day when I was thinking this and suddenly asked myself, suppose they *did* turn to me for aid in understanding the patients' psychology; what substantive information did I have that wasn't obvious on the face of the case or that represented psychology rather than what the psychiatrist had already said. I could, perhaps, say that the patient was neurotic or an introvert or other such items suggested from my available tests. I had intelligence tests, and a few other inventories. I didn't have any objective personality data that would go deeper or be more analytically complex than what would suggest general statements, such as that the patient was maladjusted. . . . [As] I then perceived [personality inventories, the] variables and interpretation were not in current jargon nor did they develop suggestions that would be of value to a staff required to make routine diagnostic, prognostic, and treatment decisions.
>
> The real impetus for the MMPI came from reports of results with insulin shock treatment of schizophrenia. The early statistics on treatment outcomes, as is characteristic of new treatment ideas, promised everything from 100% cure to no effect and no value. It occurred to me that the enormous variance in effectiveness as reported from hospital to hospital depended partly upon the unreliability of the validity criterion — the diagnostic statements. If there were some way in which we could pick experimental groups of patients using objective methods, then outcome tests for treatment efficacy should be more uniform and meaningful. I did not have any objective personality instrument that was adaptable to such a design; and, thinking about the needs, I got the idea of an empirically developed inventory that could be extended indefinitely by the development of new scales." (S. R. Hathaway, quoted in Mednick, Higgins, & Kirschenbaum, 1975, pp. 350-351).

88. Clinical Versus Actuarial Prediction

89. There are two different general approaches to interpreting data derived from personality (as well as more clinically oriented) tests and related sources. Referred to as the *clinical* and the *actuarial* approaches, these approaches represent two distincdy different ways in which data are combined to yield forecasts of future performance. Underlying the clinical approach is a reliance on clinical experience and judgment. Underlying the actuarial approach is a reliance on normative data and statistical formulas.

90. Data derived from tests, interviews, case-history material, and other sources will ultimately be used to formulate a description of, predict something about, or make a decision pertinent to an assessee. Questions concerning the optimal method for integrating all of the data and formulating such descriptions, predictions, and/or decisions have been a matter of longstanding controversy within the profession of psychology. One method, referred to as the *actuarial* approach (Meehi, 1954), is distinguished by *its* exclusive reliance on statistical procedures, empirical methods, and formal rules as opposed to reliance on the interpreter's own judgment in evaluating the data. By contrast, the *clinical* approach is characterized by less formal rules and reliance on the clinician's own intuition, judgment, and experience.

91. To illustrate some of the differences inherent in these two approaches, suppose that two psychologists, one who subscribes to the actuarial approach, "Dr. Actu," and one who subscribes to the clinical approach, "Dr.

Clin," were called upon to make a recommendation concerning whether a "Mr. T. Taker" should be hired as an executive with a large corporation. Both clinicians are given identical files on Mr. Taker, containing scores on various standardized tests, case-history data, projective-test data, and interview material. Both clinicians are aware that the corporation wants to hire executives with superior abilities in the areas of leadership, decision making, organizing and planning, interpersonal skills, and creativity.

92. Dr. Actu might approach his task by going through all of the available data on Mr. Taker and then applying certain preset rules (for example, some equation to combine the data for each variable) to come up with a score on each of the five variables to be judged. If the scores on, say, three out of five of these variables exceed a certain preset cutoff score, Dr. Actu would recommend that Mr. Taker be hired. Dr. Clin may or may not arrive at the same recommendation on the basis of his analysis of the same data. The process employed by Dr. Clin is more free-wheeling and less replicable than that employed by Dr. Actu. Something — virtually anything — in the data on Mr. Taker is capable of influencing Dr. Chin's judgment as to whether this applicant has executive potential. For example, Dr. Clin may have noticed that the written physical description of Taker included the fact that he wore one gold earring to the interview. On the basis of this fact alone, Dr. Clin might recommend that Taker not be hired; having interviewed hundreds of executives and prospective executives for this firm, Dr. Clin has mentally formulated an image of what the successful male executive looks like — and there is no provision for one gold earring in that picture.

93. The sample situation we describe is exaggerated for the purposes of illustration, for the clinical approach is characterized by careful scrutiny of all available data; and conclusions are typically drawn on the basis of a constellation of factors, not just one (such as preference for wearing earrings). Still, our summary is useful in highlighting the nature of clinical as opposed to actuarial judgments. Dr. Clin may have rejected Taker solely on the basis of an element of his attire. Taker might also have "lost points" with Dr. Actu for this manner of dress as well, but only if "manner of dress" were one of the preset criteria to be rated in the assessment equation; exactly what importance, weight, or relevance the earring would be given in the hiring equation would have to have been placed into the selection equation before the selection procedure had begun. The actuarial approach, in contrast to the clinical one, is strictly empirical in nature. If a large body of existing data indicates that males who wear one earring to employment interviews (or, stated more broadly, persons who dress in a manner inconsistent with the "image" of a particular corporation) turn out to be poor executives, such persons will lose points in their evaluation. With respect to the clinical approach, the body of data being used as a reference is the information, knowledge, and experience of the clinician making the judgment.

94. A difference between the two approaches that must be emphasized concerns the *meaning* assigned to certain data. Because the actuarial approach is so empirical in nature, meaning of responses and behaviors is deemphasized in favor of how such responses and behaviors correlate with a certain criterion. If successful male executives for the company in question do not tend to sport earrings, that will be sufficient for Dr. Actu to reject the applicant. Alternatively, Dr. Chin might overlook and "see beyond" the earring, noting that other data suggest Taker to be a highly creative, artistic, and independent individual who would do well in a particular executive slot that the corporation needs to fill. Clin's report to the corporation might recommend Taker be offered the executive position, conditional upon his removal of the earring. If Taker was hired, consented to removing the earring, and did very well in the position, the corporation might then seek to recruit other applicants who fit a similar profile.

95. Since there is a finite set of data available to the clinician, it would be nice if there was one best way to interpret that data. An architect of the actuarial approach, Mcehl (1984) likened the clinical approach to leaving a supermarket and saying, "Well, it looks like I spent about 17 bucks worth" instead of consulting the cash register receipt to know what was actually spent. Citing reasons why the actuarial approach has failed to achieve widespread adoption, Mechi's list included the following factors: (1) the ubiquity of irrationality in the conduct of human affairs, (2) sheer ignorance, (3) the threat of technological unemployment, (4) strong theoretical identifications on the part of some clinicians, (5) claims that actuarial techniques are "dehumanizing," (6) mistaken concepts of ethics, and (7) computer phobia.

96. Einhorn (1984) has asked how we can presume to make predictions about the course of human life if we can't even do it for interest or mortgage rates. Einhorn argued that clinicians must accept the reality that there will always be error in prediction. Since clinicians have more limited information—processing than computers, there would appear to be more room for error in the clinical approach.

97. Others have added that the process of making predictions clinically may be tedious while computers may make the same or better decisions within seconds. And others have argued that computers compute and can at best

show low levels of relations; in essence, they yield regression equations with neither understanding, compassion, nor the ability to anticipate unforeseen and unanticipated (that is, nonprogrammed) events. With respect to the latter point, no computer ever predicted that there would be a national oil shortage in this country in the early 1970s. The shortage arose as a result of an Arab fuel boycott, which arose in part as a consequence of the support of the United States for Israel in the Yom Kippur war. Thus while there was no shortage of computer printouts indicating rates of fossil fuel consumption and production in this country and throughout the world, no computer could have forecasted the unlikely chain of events that resulted in not only the oil shortage but also a number of related consequences (such as gas-station lines, federal energy usage restrictions and incentives, and the imposition of a national speed limit of 55 miles per hour).

98. Clearly, both the clinical and the actuarial approach have much to be said for them. The actuarial approach rends to he much more efficient than the elinical one in terms of making predictions in a variety of situations, especially those in which many predictions must be made and a large data base for making those predictions exists (Meehi, 1954, 1959, 1965). Owing to its rigor, the actuarial approach lends itself well to research; volumes have been written, for example, concerning descriptions of persons with particular MMPI patterns. Being less subject to empiricism and to rules, the clinical approach has as its chief advantage flexibility and the potential for using the novel cOmbination of data (“programmed” as well as “unprogrammed”) to arrive at decisions, descriptions, predictions, and hypotheses.

99. In summary, the difference between the clinical and the actuarial approach to assessment is in some ways similar to the difference between a courtroom trial that will result in a ruling by either a judge or a computer. Both the computer and the judge will take in all of the evidence and weigh it. Each will arrive at a verdict on the basis of the weight of the evidence and the applicable standard (“guilty beyond a reasonable doubt” in a criminal proceeding and “preponderance of the evidence” in a civil proceeding). The computer will weigh the evidence according to preprogrammed rules and arrive at a verdict. The judge will also weigh it according to ("preprogrammed") rules but with more openness to nuances of information that might not be in the “rulebook.” While the computer’s decision can be expected to conform to the letter of the law, the judge’s decision can be expected to conform with not only the letter of the law but its spirit as well.