ACROSS INDIVIDUALS

CONSISTENCY OF MEASUREMENT

INDIVIDUALS WITHIN
ACROSS AND
ACROSS MEASUREMENT

CONSISTENCY

VERSUS INTERPRETATION APPROACHES TO MEASUREMENT ALTHOUGH PSYCHOLOGISTS
ONE OF THE PRESENT CONVENTIONS IN PSYCHOLOGY INVOLES NOMOTHETIC

STANDARDIZED MEASURATIONS
PSYCHOLOGICAL INFORMATION
EVIDENCE FOR MODE INCONSISTENCY
HISTORY OF BEHAVIOR AFFECT AND COGNITION
CONSISTENCY WITHIN HUMAN RESPONSE MODES: CONGRUENCY
ERROR SIMILARITY
EXAMPLES OF CONTEMPORARY BRIEFNESS STANDARDS
EXPERIMENTAL INHIBITION
HISTORICAL INFLUENCES
EMOTIONAL AND AFFECTIVE INFLUENCES
COGNITIVE INFLUENCES
HUMAN SUBJECTS
STANDARDIZED MEASUREMENT ERRORS: SOURCES OF INCONSISTENCY
A BRIEF HISTORY OF THE
CONSISTENCY OF MEASUREMENT WITHIN INDIVIDUALS
HISTORICAL
HISTORICAL
CONSISTENCY OF MEASUREMENT ACROSS INDIVIDUALS

2
typically discuss this issue in the context of personality psychology, the nomothetic–idiographic debate has relevance across psychological domains.

Nomothetic measurements observe attributes of populations, whereas idiographic measures focus on individuals. The objects of nomothetic measurement are assumed to be present in every person. A nomothetic theorist would maintain that every person could be described as possessing some amount, for example, of the traits of neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness (Goldberg, 1993; McCrae & Costa, 1985; Wiggins, 1982). Idiographic theorists believe that individuals possess unique characteristics that may be shared with all, some, or no other people. An idiographic account is concerned with how the characteristics of a person combine into a unified, unique personality (Allport, 1937). From an idiographic perspective, a particular person might be described as very neurotic and somewhat extraverted, but the dimensions of agreeable, open and conscientious simply would not be meaningful to the study of this individual. From an idiographic perspective, error occurs when a score is assigned to an individual for a category that has no relevance for that individual.

Idiographic researchers study one or a few individuals, often over a long period. Nomothetic researchers study large groups, often on one occasion. Nomothetic researchers search for and believe that their research results apply to all persons, although such goals are also common to some idiographic researchers; for example, C. McArthur (1968) maintained that “we need to know many human natures before we can hope that Human Nature will be revealed to us” (p. 175). Both groups have tended to disparage the other. Allport (1937) quoted Meyer (1926): “A description of one individual without reference to others may be a piece of literature, a biography or novel. But science? No” (p. 271). Allport replied: “The person who is a unique and never-repeated phenomenon evades the traditional scientific approach at every step” (1937, p. 5). Although nomothetic approaches dominate many areas of contemporary psychological measurement, it is not surprising that idiographic measurement has its strongest foothold in such areas as clinical assessment, where psychologists tend to work with single persons.

**Large Samples and Individual Differences Equals Nomothetic Measurement**

How did nomothetic approaches come to dominate measurement? As noted in Chapter 1, psychometricians developed statistical models to describe the relations among the psychological characteristics they studied. To reach sufficient power to detect such relations, statistical methods require large samples of individuals. The larger the aggregate of individuals, the more likely that random errors of measurement would balance each other, thus increasing the chance of detecting the characteristic. If measurement errors balanced or canceled, it did not matter who any particular subject was, as long as you had many subjects. If you required many subjects, however, you also needed to assume that everyone in the sample possessed the characteristic. However, if all persons do not possess the characteristic, you must identify an individual who does and study that individual. As Danziger (1990) wrote, “If the subject is an individual consciousness, we get a very different kind of psychology than if the subject is a population of organisms” (p. 88).

Psychologists such as Allport (1937) took the nomothetic approaches to task because of their emphasis on groups of individuals instead of the individuals themselves. Idiographic psychologists were interested in developing laws that generalized across persons instead of groups of persons (Lamiell, 1990). For Allport, there were no psychological laws to be found outside the study of individuals. Lamiell (1990) provided an example of such a strategy in a series of studies that investigated how individuals rate other persons along psychological dimensions. Lamiell found that subjects typically rate other people not by comparing them to others (i.e., looking for differences between individuals), but by comparing the information provided about the person to an imagined polar opposite of that information. Thus, this comparison is not a retrieval of information from memory, but “mental creations of the person formulating the judgment” (Lamiell, 1991, p. 8).

In practice, the nomothetic approach seemed to work—to a point. With large samples, one could produce bell-shaped distributions of psychological characteristics, thus mirroring the distributions found in other sciences. But psychologists often found only weak correlations between psychological characteristics and the behaviors they were supposed to predict. Although correlations of .30 aided selection decisions in large groups, they still surprised psychologists. Why did x and y only correlate at .30? Were internal psychological variables and behavior really correlated in nature at such a low level? Or had psychology reached the limit of its statistical-measurement capabilities? Idiographic proponents have cited the selection of a nomothetic approach to measurement as the major cause of this and other problems reflecting a lack of scientific progress in psychology. Progress in the accumulation of knowledge, they maintained, cannot be achieved with nomothetic approaches. Similarly, more valid prediction of individual behavior might also be possible if measurement were idio graphically based (cf. Magnusson & Endler, 1977; Walsh & Betz, 1985).

From this discussion it is evident that the purposes of nomothetic and idiographic measurement can be considered complementarity, not anti-
CONSISTENCY OF MEASUREMENT WITHIN INDIVIDUALS

A Brief History of Traits

Social statistics were developed during the eighteenth century for the purpose of linking social and economic variables to social reform (Danziger, 1990). Crime rates, for example, appeared related to geographic locale, with the attendant environmental influences (e.g., poverty) readily recognized. To explain these statistical regularities, Quetelet conceived of the idea that individuals might possess "propensities" to commit acts such as homicide or suicide. Buckle argued that "group attributes were to be understood as nothing but summations of individual attributes" (Danziger, 1990, p. 77). Propensities and attributes became traits, and the application of social statistics to individuals seemed a natural progression.

Psychological measurement and assessment have long been guided by the assumption that psychological phenomena are traits (Maloney & Ward, 1976; R. P. Martin, 1988). This assumption provided psychologists with a set to find consistency in such phenomena. Most definitions of attitudes, for example, have assumed some consistency or persistence. Krech and Crutchfield (1948, cited in Scott, 1968) defined an attitude as "an enduring organization of motivational, emotional, perceptual, and cognitive processes with respect to some aspect of the individual's world" (p. 152). Similarly, D. T. Campbell (1950) wrote that "a social attitude is . . . evidenced by consistency in response to social objects" (p. 31). Many contemporary psychologists continue to assume they are measuring traits, as evidenced by the fact that psychologists typically observe individuals and administer tests in their offices and assume that the resulting behavior generalizes outside of that particular situation (R. P. Martin, 1988).

K. R. Murphy and Davidshofer (1988) suggested that the concept of a trait has three meanings. First, psychological traits are causes. Thus, persons who are introverted avoid extensive social interaction, that is, their introversion motivates them to avoid others. Historically, this is the meaning of traits employed explicitly or implicitly by most psychologists. Second, traits function as convenient organizational schemes for perceiving and remembering similar information. Thus, we might tend to term certain behaviors (e.g., turning in a found wallet, paying all of the taxes you owe) as "honest" although their relatedness may only be illusory. Or the relation may be real: individuals form concepts about how to act across situations, which others perceive as traits (e.g., Stagner, 1984). Third, traits can be considered descriptive summaries of behavioral consistencies or act frequencies (Buss & Craik, 1983). Anastasi (1985) suggested that this conception of traits is being increasingly accepted. The personality traits identified by factor-analytic studies, for example, can be seen as "summarizing behavioral consistencies, rather than as underlying, fixed, causal entities" (Anastasi, 1985, p. 121).

The problem with trait-consistency approaches to measurement is that human behavior is also variable. Behavior changes, yet measurement approaches predominantly depend upon trait ideas. It should be no surprise, then, that measuring change is one of the most difficult tasks for psychological measurement. Regression toward the mean (RTM), for example, is a frequently cited problem in psychological research (e.g., Cook & Campbell, 1979). When a measure is administered two times, it is often observed that scores falling at the extremes of a scale at the first occasion often move toward the mean at the second measurement. This can be a fatal alternative explanation when trying to interpret the results of research that contrasts to...
psychological treatment and control groups that are not equal before an intervention.

Suppose, for example, that you design a study to test the effectiveness of an intervention to decrease classroom behavior problems. For the dependent variable, you choose a checklist of classroom behavior problems. Before the intervention, teachers of students in the treatment class and the control class complete the checklist daily for 1 wk. Figure 8 displays the range of daily behavior problems for the treatment and control classes. As shown in Figure 8, the mean score of students in the treatment group is higher at pretest than the mean score of control students. If the treatment students' problems decline from pretest to posttest, whereas the control group scores remain unchanged, two alternative explanations appear: (a) the treatment worked, thus decreasing behavior problems, or (b) RTM occurred. If the true mean of the problem checklist in this example is 10, then we would expect the treatment group scores to decline upon retesting even without an intervention. RTM is a strong possibility when the treatment and control groups are not randomly assigned, as often occurs in quasi-experimental designs.

Interestingly, tests with more measurement error display more regression toward the mean (Kazdin, 1980). In other words, the more error in a measurement (i.e., the more factors influencing the test score that we do not understand), the more likely that its scores will change.

**SYSTEMATIC MEASUREMENT ERRORS: SOURCES OF INCONSISTENCY**

Error refers to factors that influence measurement in ways we do not recognize or understand. Random errors are those that occur unpredictably. Systematic errors occur in some regular manner and may accumulate, as does a trait, with aggregation. Historically, test theorists and developers have assumed errors in measurement to be random rather than systematic. That is, at the scale of the individual, errors might be systematic—one person may distort responses because of fatigue, another because of poor comprehension—but in large groups such a conglomeration of errors will behave as if they were random (K. R. Murphy & Davidshofer, 1988).

Psychologists have long sought to identify systematic errors in the measurement process. Table 3 displays errors, examined in the traditional psychological measurement and behavioral assessment literatures, that influence the consistency of individuals' test responses and observers' ratings. For example, individuals change their responses to test items when the items are rephrased from positive ("Do you feel the world is a good place to live in?") to negative ("Do you wish that you had never been born?") (Ong, 1965). Individuals behave differently when they are observed and unobserved (Webb, Campbell, Schwartz, Sechrest, & Grove, 1981). Changing the schedule of self-monitoring can influence the resulting data (R. O. Nelson, 1977b); retesting in as short a period as 1 day can reveal changes in test scores (Dahlström, 1969; L. A. Hough, Eaton, Dunnette, Kamp, & McGloy, 1990). Given the emphasis on traits, these inconsistencies are surprising.

**Human Judgment Ratings**

Human Judgment Ratings (HJR) are defined here as qualitative or quantitative assessments made by individuals, about themselves or others, along a psychological dimension. HJR consists of two types: self-reports and

<table>
<thead>
<tr>
<th>TABLE 3</th>
<th>A Partial List of Measurement and Assessment Error Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>R. L. Thorndike (1949)</td>
<td>Paul et al. (1986a)</td>
</tr>
<tr>
<td>(in K. R. Murphy &amp; Davidshofer, 1988)</td>
<td>(1986a)</td>
</tr>
<tr>
<td>Test-taking skills</td>
<td>Carelessness</td>
</tr>
<tr>
<td>Ability to comprehend instructions and items</td>
<td>Fatigue</td>
</tr>
<tr>
<td>Response sets</td>
<td>Boredom</td>
</tr>
<tr>
<td>Health</td>
<td>Information overload</td>
</tr>
<tr>
<td>Fatigue</td>
<td>Emotional strain</td>
</tr>
<tr>
<td>Motivation</td>
<td>Attention shifts</td>
</tr>
<tr>
<td>Stress</td>
<td>Equipment failure</td>
</tr>
<tr>
<td>Set for a particular test</td>
<td>Variations in lighting and temperature</td>
</tr>
<tr>
<td>Examiner characteristics</td>
<td>External distractions</td>
</tr>
</tbody>
</table>
other ratings. Self-reports have also been referred to in the literature as S data and observers’ ratings as R data (Cattell, 1957, 1973). Self-reports are judgments made by individuals about some personal psychological attribute (e.g., rate your current job satisfaction). Other ratings occur when one person rates another on a psychological dimension, as when a manager rates an employee or a teacher rates a child on some attribute (e.g., rate how persistent a student has been over the past semester). Other ratings will be discussed in detail in Chapter 3.

As previously noted, the use of large samples in psychological research meant that one no longer needed to assume, as was the case in early experimental work, that the person who made the report was a psychological expert. As long as the person provided data of some validity, errors would cancel in the aggregation. With both self- and other ratings, Danziger (1990) observed that “during the early period of personality psychology, and to a considerable extent thereafter, it was simply assumed that personality ratings were an unproblematic product of attributes of the task, not attributes of the rater” (p. 160). Test items were assumed to be face valid across individuals and individuals were assumed to be able to respond in a valid manner to those items.

Given this history, it should be no surprise that self-reports and ratings by others constitute the method most frequently employed throughout the history of psychological measurement. With the exception of a few areas, such as behavioral assessment, most contemporary psychologists continue to rely on self-reports. Kagan (1988), for example, cited research indicating that most personality research during the past 10 years was based on self-report questionnaires. Noting that self-reports have been employed in alcohol research since the beginning of the twentieth century, Babor et al. (1987) observed that verbal reports remain “the procedure of choice for obtaining research data about patient characteristics and the effectiveness of alcoholism treatment” (p. 412). If self-reports are assumed to be credible and valid in some sense, their widespread use is inevitable given their ease of development, administration, and scoring. Indeed, investigators are continually tempted to create new instruments in their research area, as evidenced by a recent APA (1992) advertisement in the Monitor that estimated that 20,000 psychological, behavioral, and cognitive measures are created annually. Hartman (1984) noted that “if a behavior has been studied by two investigators, the chances are very good that at least two different self-report questionnaires are available for assessing the behavior” (p. 133).

Despite the widespread use of self-report, psychologists often seem to adopt one of the following dichotomous beliefs: (1) because individuals can self-report, self-reports must be valid, or (2) because self-reports can be distorted by test-takers, self-reports are useless. The first position represents that taken by most early measurement psychologists. In contrast, self-report critics espousing the second position have pointed to studies comparing self-reports to what the critics see as a more objective criteria, that is, overt behavior. Psychologists have consistently found discrepancies between self-reports of psychological phenomena and overt behavior indicative of or related to the phenomena (e.g., Doleys, Meredith, Poire, Campbell, & Cook, 1977; H. E. Schroeder & Rakos, 1978). Kagan (1988) summarized a variation on this perspective:

A serious limitation of self-report information is that each person has only a limited awareness of his or her moods, motives, and bases for behavior, and it is not obvious that only conscious intentions and moods make up the main basis for variation in behavior. . . . Less conscious indexes, including facial expressions, voice quality, and physiological reactions, occasionally have a stronger relation to a set of theoretically predicted behaviors than do self reports. The reader will remember that when a large number of investigations were derived from psychoanalytic theory, the concept of unconscious ideas was prominent and self-report was treated as replete with error. (p. 617)

Although the discordance between behavior and self-report has led many psychologists to conclude that self-reports are untrustworthy sources of information, other explanations are available for these inconsistencies. Many psychologists have wondered if the prediction of behavior could be improved if systematic errors in the self-reports were identified and corrected in some manner. What are considered some of the most important of these errors are described below.

Response Strategies

Response strategies refer to the processes individuals employ to complete psychological items, problems, and tasks. Response strategies may be classified along an unbiased retrieval–generative dimension. At one end of this continuum are retrieval strategies used in the direct recall and reconstruction of extensive information about self or others from long- and short-term memory storage. At the other end are generative strategies, employed when individuals cannot or will not produce the relevant information from memory.

Test-takers probably employ strategies throughout the continuum when completing psychological tests, but many who use psychological tests often assume a predominance of unbiased retrieval strategies. That is, respondents are assumed to be retrieving valid information from memory as compared to creating distorted data. Considerable evidence exists, however, to suggest the frequent use of generative strategies. For example, when respondents answer questions that request information about events or beliefs that they do not possess, they create the information, as opposed to retrieve it from memory (A. F. Smith, Jobe, & Mingay, 1991). This may occur regardless of whether the items request specific data (as in the question, “How many visits to a physician have you made in the past
6 months?" or beliefs and attitudes ("How often should a person visit a physician per year to maintain good health?"). Whatever the cause, human judgment fallacies have been well documented (Dahlstrom, 1993; Nisbett & Ross, 1980).

**FACTORS RELATED TO STRATEGY SELECTION**

The degree to which respondents employ generative as compared to unbiased retrieval strategies appears to be related to a variety of cognitive, affective/motivational, and behavioral factors. These factors are discussed below.

**Cognitive Influences**

Cognitive factors refer to how individuals think about and process information related to tests. Cognitive psychology’s relatively recent rise as an important paradigm has led many measurement psychologists to consider the role of cognitive factors in the testing process.

*Level of Knowledge and Uncertainty.* To the extent that respondents find ambiguous the elements of the testing situation—such as test items, instructions and format—they may be expected to employ more generative strategies. What do people do when asked a question for which they are uncertain about the answer? MacGregor, Lichtenstein and Slovic (1988) suggest that the simplest strategy to pursue, when estimating an uncertain quantity, is to intuitively produce a number based on whatever information comes to mind. Although inexpensive, simple, and often approximately correct, such intuitive judgments have been shown to lead to systematic biases (MacGregor et al., 1988). For example, MacGregor et al. note that such judgments tend to remain close to an initial value (Tversky & Kahneman, 1974) and may be influenced by recency, salience, and vividness effects.

Some evidence indicates that persons with less experience in a behavioral domain are more likely to exhibit greater attitude–behavior inconsistency in that domain (Fazio & Zanna, 1978; Regan & Fazio, 1977). Regan and Fazio (1977), for example, found that college students who personally experienced a dormitory housing crisis showed greater consistency between their attitudes and behavioral attempts to cope with the crisis than did students with similar attitudes but no direct experience.

*Memory Processes.* Many contemporary descriptions of the item-response process rely heavily on cognitive models (e.g., Babor et al., 1990; Biemer, Groves, Lyberg, Mathiowetz, & Sudman, 1991). Theorists usually include the following stages in the item-response process:

1. Individuals, over time, notice their behavior or the behavior of others.
2. They store that information in memory, where it is subject to some degree of decay.
3. Individuals are presented with an item or task that they attempt to comprehend and then relate to information in memory.
4. Individuals retrieve the information and employ it in response to the item or task.

Distortion may arise from a number of complications during this process. For example, individuals usually do not know what information they will be required to retrieve for a subsequent psychological test, and consequently they cannot plan to systematically store relevant information. The longer the time period between information storage and recall on the test, the more likely the information will decay or be interfered with and retrieval altered in some manner. The most extensive production strategies must be employed when rating the past or the future because the information to make those ratings must be retrieved from memory or generated on some basis. The least generation should occur with ratings of the present, as when a person self-reports about current mood or a teacher rates a student’s current classroom behavior. As the length of time increases between the rated event and the rating itself, so does the inaccuracy (cf. Paul, Mariotto, & Redfield, 1986a). Similarly, the number and complexity of the behaviors that observers must process about themselves or others appear to influence reliability. Nay (1979) reviewed research that found high negative correlations (in the -.50 to -.75 range) between the complexity of the categories recorded by an observer (defined as the number of categories employed divided by the total number of entries) and the reliability of those observations.

*Schemas.* Schemas are cognitive structures or networks that organize information. Usually functioning outside of awareness because they have been overlearned, these existing stores of knowledge and action sequences allow individuals to expect certain events, interpret new information, and solve problems (Gentile, 1990).

Test items and tasks reflect the language and schemas inherent in the culture of the test developer. To the extent that test respondents hold discrepant organizational schemas about test information, distortion from the reference frame of the test developer will occur. Such discrepancies are likely, for example, when the requirements of the test and the experiences of the respondent differ on such dimensions as language and culture. Some evidence suggests that a percentage of individuals store relevant information in a form incompatible with the measurement procedure. Tellegen
(1985) reported a study in which 23 subjects completed self-ratings of mood over a 90-day period. Subjects' responses were factor analyzed separately; for 21 of the subjects, a two-factor solution emerged. However, for 2 of the subjects, no interpretable result was found. To search for suspected differences in perceived item meaning, 15 subjects (including the 2 discrepant subjects) were recontacted and asked to sort the mood terms into subsets with similar meanings. Analysis of the discrepant subjects' sortings indicated that they were processing and understanding the terms in a manner different from the remaining individuals. Tellegen concluded that the two-factor model of emotion was valid "provided respondents are able to report emotional experiences in accordance with consensual semantic rules" (p. 704).

**Summary.** Cognitive factors such as knowledge level, memory processes, and schemas represent potential sources of systematic errors in psychological measurement. However, such internal factors, by themselves, cannot account for the presence of error. Rather, the interaction of these cognitive factors with test design and purpose increases the likelihood of generative responses. Thus, a test that requires the respondent to read at the eighth-grade level is likely to elicit retrieved responses from a group of eighth graders, but guesses or random responses from a group of second graders. I will discuss in more detail the results of such matches and mismatches at this chapter's conclusion.

**Motivational and Affective Influences**

These factors refer to the effects of individuals' affective characteristics and states on the testing process. Interestingly, even these influences tend to be discussed in the literature in terms of cognitive processes.

**Testing Consequences.** Given the widespread use of psychological tests for selection decisions, it would seem apparent that considerable emotion could result from an individual's perceptions of testing consequences. Tests can help decide whether a person obtains a particular job, is admitted to a desired school, or avoids an unwanted intervention. Cronbach (1984) noted that:

> Draftees have been known to report impressive arrays of emotional symptoms, hoping for discharge. In an ordinary clinical test, exaggerating symptoms may be a gambit to enlist sympathy and attention. An unsuccessful student may prefer to have the teacher believe that his troubles are caused by an emotional disturbance rather than to be thought of as stupid or lazy. (p. 471)

In addition, when tests become the vehicle to create a label or diagnosis that becomes known to test-takers and other decision makers, their consequences can have effects that last long beyond any immediate decision. Such labeling can potentially influence individuals' self-concepts and behavior across a range of situations; for example, a student who is placed in a remedial class on the basis of a test may overgeneralize a lack of skill to other content areas (cf. Fairbanks, 1992). This type of effect is one reason psychologists have increased their attention to ethical issues in testing. For example, Messick (1980, 1989a, 1989b) discusses test validity in terms of the function of testing (interpretation and use) as well as the justification for testing (appraisal of evidence and consequence).

Attempts to cope with problems introduced by testing consequences have ranged from complete openness to concealing testing purposes (Cronbach, 1984). Cronbach (1984) suggested that making the testing purpose transparent is most common in situations where respondents are anonymous (as in some types of opinion polling) or when respondents may potentially benefit from valid self-disclosure (as in symptom reports in preparation for a clinical intervention). At the other extreme is a strategy of concealment where test developers attempt to hide the test purpose. For example, developers frequently create innocuous titles for tests (e.g., "Human Services Inventory" instead of the Maslach Burnout Inventory; Maslach & Jackson, 1981) or provide test-takers with a plausible but false rationale for the testing purpose (Cronbach, 1984).

**Test Anxiety.** Gregory (1992) provided a contemporary review of evidence and theory about test anxiety, the emotional experience of individuals who anticipate failure on a test. Although noting that past research has shown that test anxiety negatively affects test performance (e.g., Hembree, 1988), Gregory (1992) also questioned whether poor performance precedes and causes the anxiety. For example, Paulman and Kennelly (1984) found that test-anxious students had ineffective test-taking strategies, whereas Naveh-Benjamin, McKeachie, and Lin (1987) found that many test-anxious students also possessed poor study habits.

Gregory (1992) cited studies that indicate that test-anxious individuals appear to possess a threshold that once crossed, results in severe performance drops. For example, Sarason (1961, 1972) found no difference in performance on a memorization task between high- and low-anxiety individuals when the task was presented as neutral and nonthreatening. When the task was presented as an intelligence test, however, the high-anxious students' performance declined significantly. Similarly, Siegman (1956) found that high-anxious patients performed worse on timed as opposed to untimed Wechsler Adult Intelligence Scale (WAIS) subtests. The results may be explained by the cue-utilization hypothesis (Easterbrook, 1959), which indicates that emotional arousal alters individuals' ability to attend to environmental cues. As arousal increases, attention is narrowed to task-relevant stimuli; however, once arousal crosses a
threshold, individuals lose their capacity to effectively process the cues related to the task.

For individuals who perceive a topic or test situation as anxiety producing, completing the test quickly constitutes escape (negative reinforcement). Gentile (1990; see also Geen, 1987) argues that this is widespread in academic tasks, and it is likely to occur in clinical assessment as well. Similarly, Cronbach (1946) noticed that some students may speed through a test, giving the appearance of random responding.

**Emotional States.** The emotional states that individuals bring to tests or that are induced by tests can affect test response. Brody and Forehand (1986) found that depressed mothers were more likely than mothers with low depression to interpret their children’s noncompliant behavior as indicative of maladjustment. Neufeld (1977) observed that psychologists may avoid testing schizophrenics and some depressed persons because those groups are presumed to be unable to make valid judgments about their psychological attributes. Contrada and Krantz (1987) reported data indicating that illness and accompanying treatment can affect self-reports. Perceptual and cognitive distortions that may interfere with performance on measurement tasks are also apparent in such clinical phenomena as eating disorders (Halmi, Sunday, Puglisi & Marchi, 1989), stress, anxiety, and depression (Meier, 1991a).

Some authors have proposed an association between affective disorders and test response style. Freeman (1955), for example, suggested that (a) obsessive-compulsive persons provide test responses that are too detailed, but also full of uncertainty and doubt; (b) anxious persons have difficulty finding appropriate words or blurt out inappropriate replies; and (c) psychotic individuals demonstrate disorganized thinking and bizarre content in their responses.

**Fatigue and Boredom.** As previously displayed in Table 3, authors who create lists of measurement errors typically include fatigue and boredom. These psychological-state effects are presumed to be an interaction between test-taker and test characteristics. Given that traditional personality and cognitive performance tests can require several hours of effort, it is not surprising that some respondents report fatigue and negative thoughts at the end of tests (cf. Galassi, Frierson, & Sharer, 1981). Fatigue effects have been observed, for example, in surveys of magazine readership, crime reports, and reports of symptoms (Sudman & Bradburn, 1982).

In general, humans attempt to minimize their cognitive-processing load (e.g., Fisher & Lipson, 1985). Sudman and Bradburn (1982) noted that questionnaire respondents who become aware that “yes” answers are followed by lengthy follow-up questions may quickly learn to answer “no.” Similarly, questions may vary in the amount of detail they request respondents to recall (e.g., current salary vs. current interest on savings). As Biemer et al. (1991) noted, when questions become too difficult for respondents, they may prematurely terminate their cognitive processing.

**Summary.** Testing consequences, test anxiety, emotional states, and fatigue and boredom are potential sources of systematic errors. As with cognitive factors, individuals’ affect and motivations interact with test design and purpose. Thus, an ability test administered to select new employees may evoke different motivational states and response styles than the same test administered in a research study investigating learning style.

**Behavioral Influences**

The testing environment can also influence test-takers' responses. Potent factors include the presence or absence of observers, test administrators' gender and race, the physical characteristics of the testing room (e.g., temperature and lighting), and the use of testing apparatus such as pencils or computer keyboards (which may pose difficulties, for example, for persons with physical disabilities). Probably the most studied problem involves the use of behavioral observers.

**Reactivity.** Although the term has been employed in the literature with slightly different meanings, reactivity is defined here as the possible distortion that may result from individuals’ awareness that they are being observed by other persons for the purpose of measurement (Kazdin, 1980). The assumption is that as a result of learning (a) that testing or research is occurring, or (b) the intent of testing or research, individuals may respond differently than they would in unobserved situations. Hartman (1984) reviewed research that found that children’s reactivity is influenced by such observer attributes as gender, age, and activity level, whereas adults are influenced by observers' tact and appearance.

Reactivity has also been described in terms of the transparency of testing or research; the purpose of test items, for example, can be described as more or less obvious to test-takers. The potential importance of reactivity can be illustrated by results reported in Smith and Glass's (1977) meta-analysis of psychotherapy outcome research. Smith and Glass calculated correlations between the amount of psychotherapy gain and such variables as client intelligence, therapist experience, and the reactivity of outcome measures. To gauge reactivity, Smith and Glass rated the transparency of each measure employed in the 375 psychotherapy studies they examined. Of all factors examined, reactivity correlated the highest at .30. This means that studies that employed the most transparent measures demonstrated the greatest therapeutic gain, leaving open an important alternative methodological explanation for study results.
Several strategies have been employed to decrease experimenter influence on subjects' behaviors (Kazdin, 1980; Rosenthal, 1976). Mentors' expectancies and subjects' desire to receive mentors' approval may be motivated to conceal the measures' purpose. Although researchers who employ unobtrusive measures may be motivated to conceal the measures' purpose, practitioners who wish to collect information of a unidimensional construct. Finally, unobtrusive measures may pose ethical problems. The logical direction to move with such an approach is to make measurement unobtrusive, that is, to collect data from individuals without their knowledge. Unobtrusive measures have been proposed as a viable alternative and supplement to traditional assessment strategies (cf. Webb et al., 1981).

**Unobtrusive Measurement.** Examples of unobtrusive measurement include simple observation in naturalistic settings, observation in contrived situations, examination of archival records, or obtaining physical traces (Kazdin, 1980; Webb et al., 1981). Abler and Sedlacek's review (1986) provided several applied examples of unobtrusive measurement. In one study, researchers attempting to determine the effectiveness of an assertiveness training program posed as magazine salespersons and telephoned former participants to determine the program's effects (McFall & Marston, 1970). Another group of researchers found that prospective college students who made more errors filling out orientation applications were more likely to drop out (Sedlacek, Bailey & Stovall, 1984). Epstein (1979) reported a study in which students' self-reports of tension were significantly correlated with the number of erasures on exam answer sheets, number of absences, and number of class papers that were not turned in.

Several problems are inherent, however, with unobtrusive measurements (Kazdin, 1980; Meier & Wick, 1991; Webb et al., 1981). First, considerable effort may be necessary to obtain an unobtrusive measurement. It is much easier, for example, to administer a self-report scale to alcohol treatment participants than to create a simulated bar or observe subjects drink on weekend nights. Second, collecting unobtrusive measurements without arousing subjects' suspicions may be difficult. Third, construct validity is seldom addressed with unobtrusive measures. The behavior of individuals in naturalistic or contrived situations, for example, may not be direct reflections of a unidimensional construct. Finally, unobtrusive measures may pose ethical problems. Although researchers who employ unobtrusive measures may reveal this fact at debriefing, practitioners who wish to collect multiple unobtrusive measures (e.g., at the beginning, middle, and conclusion of multiple treatments) may be motivated to conceal the measures' true intent.

In experimental situations, researchers have documented that experimenters' expectancies and subjects' desire to receive experimenters' approval influence subjects' behaviors (Kazdin, 1980; Rosenthal, 1976). Several strategies have been employed to decrease experimenter expectancy and experimenter approval effects. One might attempt to keep the person who actually runs the study—the experimenter—as well as the subjects blind to the study's hypotheses. One might also include a control group whose expectations have been set similar or counter to the experimental group's analyses would contrast the changes in such a control with those of the intervention. With such control groups, expectancies and desirability factors become objects of investigation rather than error.

**Summary.** The environment and context of testing provide a third category for describing sources of systematic errors. The most central of these problems is reactivity, changes in behavior that occur in individuals who become aware of being observed or measured on some psychological dimension. Unobtrusive measures and concealing knowledge about testing purpose are among the strategies employed to circumvent these problems.

**Examples of Generative Response Strategies**

Generative response strategies such as socially desirable responding and random responding have been extensively studied by psychologists. As will be seen below, however, little consensus exists about the importance of such strategies, and with a few exceptions, about methods to minimize them.

**Random Responding.** Systematic errors at the level of the individual may result in what appear to be random test responses. For example, respondents' lack of motivation to cooperate with testing may be manifested by responding to items randomly.

Test developers attempt to identify random responding by including items likely to be true or false for all persons (e.g., "I eat every day"). Clinicians may become experienced in recognizing random response profiles. Random response profiles, however, may also resemble those produced by persons with psychiatric diagnoses such as psychosis (K. R. Murphy & Davidshofer, 1988). Consequently, psychologists who administer tests will also conduct an interview to separate random responders from individuals with other problems.

Berry et al. (1992) investigated random responding in a series of studies with the MMPI-2. In a study of college students, they found that 60% gave one or more random responses to the 557 items. Seven percent reported random responding to many or most of the items; subjects who acknowledged some random responding averaged 36 such responses. Berry et al. found few correlations between self-estimates of random responding and subjects' demographic characteristics. In a second study, Berry et al. found that most subjects who admitted to random responding reported having done so at the end of the test, although another sizable group scattered responses throughout. A third study with subjects from the general
population found that the number of self-admitted random responders dropped to 32%. Finally, a study of 32 applicants to a police training program found that 53% indicated that they had randomly responded to some items.

**Dissimulation and Malingering.** Dissimulation refers to faking good or bad on test items (K. R. Murphy & Davidshofer, 1988), whereas malingering occurs when individuals simulate or exaggerate psychological conditions (G. P. Smith & Burger, 1999). Given that many test items and tasks are transparent in their intent to detect such phenomenon as psychopathology or dishonesty, test-takers may be motivated and able to generate answers that suit their purposes rather than reflect valid or retrieved information. For example, prejudiced individuals may very well tell a pollster that they would vote for an African-American presidential candidate when in fact they would not. Similarly, individuals who wish to receive disability payments may exaggerate their complaints and symptoms. Dahlström (1985) noted that as early as the 1930s investigators were able to demonstrate the ease of faking answers on psychological tests. Terman and Miles (1936), for example, found that the most discriminating items on a scale designed to show personality differences between men and women were also the most susceptible to change under explicit instructions to fake test answers in a masculine or feminine direction.

Test developers attempt to identify and reject items that may be easily faked during the test construction process. Developers have created scales to detect malingering (e.g., Beaber, Marston, Michelli, & Mills, 1985; Rogers, Bagby, & Dickens, 1992; Schretlen, 1986; G. P. Smith & Burger, 1993) as well as tests that include special items designed to detect dissimulation. Psychiatric patients, for example, appear less able to provide normal responses when item subtlety increases (R. P. Martin, 1988). R. P. Martin (1988) reviewed the best-known MMPI scales designed to identify distorted responding, including the (1) Lie scale, items in which the respondent may claim great virtue, (2) F scale, infrequently answered responses that may indicate a tendency to fake illness, and (3) K scale, subtle items intended to assess defensiveness and willingness to discuss oneself. A weighted derivation of the K scale is added to other MMPI clinical scales to correct for the sensitivity to dissimulation of such items may also be sensitive to other factors. The F and Fb scales of the MMPI-2, made up of items reflecting clinically aberrant and statistically rare responses, are also affected by symptom exaggeration, psychosis, and random responding (Berry et al., 1992). The Variable Response Inconsistency (VRIN) scale (Tellegen, 1988) is composed of statistically and semantically rare item pairs and appears to be able to separate random responders from other groups (Wetter, Baer, Berry, Smith & Larsen, 1992).

Social Desirability. Social desirability (SD) is a type of response set, that is, a tendency to respond with answers that the respondent believes are most socially acceptable or makes the respondent look good (Edwards, 1953; Nunnally, 1967). Paulhus (1991) noted that psychometricians have been aware of SD effects since at least the 1930s (e.g., Bernreuter, 1933; Vernon, 1934). SD researchers maintain that it is a separate trait that varies among individuals (i.e., individuals have different needs for approval) and that it is SD that most personality tests actually measure. Edwards (1970), for example, summarized research demonstrating strong correlations between the probability of personality item endorsement (contained in tests such as the MMPI) and their SD value.

Although SD has been researched primarily with personality tests, the phenomenon has also been noted with other measurement methods, such as the clinical interview. Barlow (1977) describes a patient who came to treatment with problems of anxiety and depression, which the patient indicated were associated with social situations. Over a 1-yr period the patient made progress in a treatment emphasizing the learning of social skills, but still complained of anxiety and depression. Finally, the patient blurted out that the real cause of the discomfort were strong feelings of homosexual attraction he experienced in some social situations. Asked why he did not report this previously, Barlow (1977) wrote that “he simply said that he had wanted to report these attractions all year but was unable to bring himself to do so” (p. 287). Although homosexuality may not be the taboo subject it was for many people in the 1970s, issues surrounding such sensitive topics as sexuality and substance abuse remain subject to SD errors. Hser, Anglin, and Chou (1992), for instance, found that self-reports of male addicts showed greater inconsistency between two interviews for more socially undesirable behaviors, such as narcotics use, than for socially desirable behaviors, such as employment.

Given the evidence that SD affects test responses, psychologists have attempted to eliminate its effects during scale construction and during test-taking (Paulhus, 1991). Although no consensus about best methods has been reached, strategies have included the following:

1. Instructing test-takers to respond honestly (e.g., Benjamin, 1988). Little research is available to document this instruction’s effectiveness (R. P. Martin, 1988).

2. Developing instruments such as the Social Desirability Scale (Crowne & Marlowe, 1964) to identify and eliminate test items (during scale development) or test-takers (during concurrent administration of other tests), which correlate too highly with SD scores. Similarly, judges may rate new test items on a scale from extremely desirable to extremely undesirable in an effort to detect relatively neutral items. Research results suggest considerable agreement among groups of judges, including preschool children,
different psychiatric populations, and even judges from different cultures, on the desirability of specific items (Edwards, 1970; Jackson & Messick, 1969).


5. Warning respondents that methods to detect distortion exist. L. A. Hough et al. (1990) cited four studies that found support for the efficacy of these approaches (Haymaker & Erwin, 1980; Lautenschlager & Atwater, 1986; Schrader & Osburn, 1977; Trent, Atwater & Abrahams, 1986).

Despite an acknowledgment of the potential effects of SD, no consensus has been reached about the size of its effects. Dahlström (1969) suggested that SD may simply be another component of, instead of substitute for, factors such as neuroticism that are measured by scales such as the MMPI. SD, then, becomes not so much an error that must be eliminated or controlled but another component or type of psychological trait. Similarly, R. P. Martin (1988) suggested that socially desirable responses may not be invalid because most people typically do behave in a socially desirable manner. That is, individuals do attempt to manage the impressions they present to other people (Messick, 1991).

Problems such as SD bias may have persisted partially because of the dominating assumptions of selection testing. For example, McCrae and Costa (1983) wrote:

As an item or scale characteristic, therefore, SD is a potentially serious problem in situations in which information is required about the absolute level of a response. For most psychological applications, however, absolute levels of scale scores are meaningless or difficult to interpret. Instead, normative information is used to compare the score of one individual to those of other persons from populations of interest. If the scores of all individuals are uniformly inflated or decreased by SD, it will make no difference in interpreting scores, since rank order and position in a distribution are unaffected. (p. 883)

But for many testing purposes, including selection, the absolute level of response is important.

In psychotherapy outcome research, it is quite plausible that SD effects would influence the mean of individuals’ scores on such negative constructs as stress, anxiety, depression, and drug abuse. As shown in Figure 9, pretest scores on a socially undesirable construct such as anxiety might demonstrate a range restriction or floor effect (A). Many psychological interventions, however, teach clients to recognize and accept the experience of some amounts of anxiety as normal. If this education was the primary intervention effect—thereby reducing the socially undesirable perception of anxiety—posttest scores in the intervention condition might demonstrate a greater range as well as an increase in mean anxiety level from pretest to
posttest. Use of a retrospective pretest (e.g., G. S. Howard, Millham, Slaten, & O'Donnell, 1981) might demonstrate the expected pretest-posttest decrease in anxiety (C), but the strength of placebo effects—the intervention’s expectation that change has occurred, regardless of the intervention’s actual efficacy—makes acceptance of retrospective reports controversial.

To the extent that SD moves test scores toward the floor or ceiling of scale values and thereby restricts the range, interpretation of theoretical research and selection relations becomes problematic. As discussed in Chapter 1, theory development requires precise measurement that demonstrates the smallest distinctions (and therefore, greatest range) possible; to the extent that social desirability reduces such distinctions, measurements cannot reflect the full characteristics of the actual phenomena. The usefulness of selection testing depends on concurrent and predictive validity coefficients, correlations that will be attenuated when range restriction occurs. SD thus has the potential to affect many types of psychological measurement and assessment.

**Response Styles.** Lanyon and Goodstein (1982) differentiated between response styles and response sets. They described a response set as a tendency to distort responses so as to generate a specific impression; thus, social desirability is a response set. Response style is a distortion in a particular direction regardless of item content; agreeing with all items regardless of what the item states is an example of a response style. Response style and set have often been employed interchangeably in the measurement literature, causing some confusion. In the discussion below, what researchers have termed both styles and sets refer to Lanyon and Goodstein’s description of response styles.

The two most recognized response styles have been the tendency to agree with an item, called acquiescence, and the tendency to disagree with an item, called criticalness (K. R. Murphy & Davidshofer, 1988). Other proposed dimensions (see Messick, 1991) include uncritical agreement, and the tendency to disagree with items that are clear and relevant to the test-taker (Jackson, 1967). Cronbach (1946) suggested that acquiescent responding was itself one component of authoritarianism, a claim that never obtained much empirical support (Messick, 1967).

In the 1960s, acquiescence became a concern with the MMPI because 85% of its items are positively phrased (Messick, 1991). Several studies found that factor analysis of MMPI items revealed two factors that could be interpreted in terms of social desirability and acquiescence (e.g., Jackson & Messick, 1961, 1962, 1967). In other words, these researchers suggested that the MMPI did not measure its purported content as much as it reflected individuals’ response sets and styles. Block (1965) responded to this criticism by revising MMPI items to reduce the effects of response styles and sets and then subjecting responses to the resulting items to factor analysis. Results indicated that the revised MMPI had a factor structure similar to the unrevised scale. This raised doubts about whether response styles actually confounded content measurement, suggesting instead that sets and styles were actually reflections of the very traits measured by the MMPI (see also Dahlstrom, 1969). The issue continues to be raised in present times, as in Messick’s (1991) contention that problems caused by response styles have been documented in the measurement of mood states and androgyny.

Some response styles occur in the presence of certain item and test-taker characteristics. For example, acquiescence appears most pronounced when test-takers are presented with ambiguous items (Jackson & Messick, 1958). Berg (1955) suggested that acquiescence is a logical response of individuals in our culture who are presented with questions about matters they deem to be unimportant. Individuals with low verbal ability or high impulsivity are also more likely to employ response styles (Jackson, 1967).

To reduce acquiescence and criticalness, many test developers maintain a balance of items that can be scored true or false as indicative of the measured attribute. Suggestions have also been made to increase the content saturation of tests (Messick, 1991) and to write items that are clear and relevant to the test-taker (Jackson, 1967). Cronbach (1946) suggested that response sets could be reduced by any procedure that increased the structuredness of a test, such as better defining response alternatives or changing test instructions.

Interestingly, efforts to decrease response styles have been found to increase response sets and vice versa. R. P. Martin (1988) suggested that
response sets partially result from the clarity of item content: the clearer the item, the more likely that a response set such as social desirability (SD) will occur. If the item is ambiguous, however, then the probability of a response style such as acquiescence increases. R. P. Martin (1988) noted that projective tests were partially constructed on the assumption that more ambiguous stimuli would lead to less faking and socially desirable responding. This assumption, however, has not received much empirical support (Lanyon & Goodstein, 1982). Similarly, test experts have debated the usefulness of more subtle but ambiguous items, the intent of which may be less transparent to test-takers, but which may also invite acquiescence or criticalness because individuals have little basis on which to respond. For example, K. R. Murphy and Davidshofer (1988) suggest that a question like “I hate my mother” is very clear and invites a response based on its content. Test-takers may also suspect, however, that such a question may be intended to measure neuroticism or psychopathology. A question like “I think Lincoln was greater than Washington” is less transparent, but a respondent who must generate a response may simply agree because of the positive item wording. Such a respondent might also agree with the statement that “I think Washington was greater than Lincoln.” As noted previously, studies tend to favor the validity of obvious items over subtle ones. Consequently, the use of subtle items to diminish response sets may increase the likelihood of a response style and thereby diminish test validity.

ERROR SIMULATIONS

As noted above, most of the attention paid to systematic errors occurs during the item selection process. Test developers, for example, might examine correlations between test items and scores on an SD scale to eliminate items with a high SD relation. Few psychological scales contain items designed to detect systematic errors, and subscales designed to detect such errors often cannot differentiate between error types or their causes. Yet systematic errors are likely to be a function of persons and situations as well as items, so we should expect such errors even with items designed to minimize them. If errors such as random and socially desirable responding are present in test data, how could they be detected? Would such errors, for example, be apparent in the statistical procedures typically applied to describe and evaluate psychological scales?

One approach to this problem is to create a series of data sets containing ideal and error-laden values for comparison. Figure 10 displays responses to ten 5-point Likert items by 100 hypothetical individuals that form a unidimensional Guttman scale: all the persons in a higher level possess the characteristics of those at the next lower level, plus one more (Reckase, 1990). The data in Figure 10 have a mean of 30.00, a standard deviation of 14.00, and a coefficient alpha of 1.00.

\[
\begin{array}{cccccccccc}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
& \end{array}
\]

FIGURE 10 A Guttman scale.
Suppose I simulate random responding by creating a computer program that takes these data and substitutes random numbers for 50% of the values. Table 4 displays means, standard deviations, and coefficient *alphas* for 10 such simulations. The means of these data are near the original’s value; the standard deviations are considerably lower, but are relatively consistent within themselves. Surprisingly, the *alphas* are moderately high, near the median reported for actual scales sampled by Meier and Davis (1990). Given these results, I think it unlikely that most researchers and practitioners would be able to identify, in typical psychological test data, the type of systematic errors discussed above.

**CONSISTENCY WITHIN HUMAN RESPONSE MODES: DESYNCHRONY OF BEHAVIOR, AFFECT, AND COGNITION**

Binet and his contemporaries believed that physical and mental processes were closely linked. But Binet’s cognitive tasks outpredicted Cattell’s physical measures. Why should that be so? Should not individuals’ performance on sensory and perceptual tasks be the first link in the chain of psychological processes? The lack of correlation between psychological and physiological measures surprised early psychologists and opened the way to fuller inquiry into the relations between different human systems.

Awareness of desynchrony—the lack of correlation between human systems that seemingly should be interrelated—has expanded over time. Social psychologists initially assumed that individuals’ attitudes caused subsequent behavior, only to find that many attitude-behavior correlations are low (Liska, 1975). Researchers who wished to decrease alcohol and drug use in adolescents by improving attitudes and knowledge about alcohol often find little behavioral change (Tobler, 1986). And psychotherapists who assumed that changes in clients’ cognitions, affect, or behaviors would result in immediate changes across other systems have been wrong (Rachman & Hodgson, 1974). This is a measurement problem in the sense that psychological theories have often failed to describe what system should be measured when.

Psychological theorists have found it useful to divide human functioning and personality into distinct systems or modes. A large number of such modes have been suggested. A partial, overlapping list includes behavior (e.g., motor and interpersonal), affect, cognition, sensation, imagery, physiology, and communication (e.g., verbal and nonverbal) (Lang, 1968; Lazarus, 1981). Such divisions allow theorists to propose causal mechanisms to explain functioning. Other factors external to individuals, such as social, cultural, and physical environments, have also been discussed as causal factors, but they will not be included here.

Psychologists commonly discuss three modes as distinct and basic: behavior, affect, and cognition (B-A-C). B-A-C represents a simplified system that most psychologists understand. Behavior refers to the overt actions individuals perform; we may measure, for example, how frequently a manager speaks with an employee and the duration of those conversations. Affect is the emotion or mood a person experiences; an employee might experience such feelings as anxiety or satisfaction on the job. Cognition refers to the covert thought processes of an individual; a manager might think “Giving Jones a pay raise might increase her job satisfaction and keep her at this company.”

Desynchrony refers to inconsistencies within an individual’s B-A-C modes. For example, an adolescent might express, during a health education...
higher psychological processes. In 1896 psychologists. Fechner and Cattell, one of the first crises in the history of psychological measurement. As noted in Chapter 1, the success of physiology as a science strongly influenced early psychological laboratories, but viewed physiology as the foundation for the higher psychological processes. In 1896 Cattell asked, "To what extent are the several traits of body, of the senses, and of mind interdependent? How far can we predict one thing from our knowledge of another? What can we learn from the tests of elementary traits regarding the higher intellectual and emotional life?" (Sokal, 1987). Psychologists were interested in discovering general laws that would mark psychology as a legitimate science, and a genuine body–mind link would certainly qualify as an important instance of such a law.

It follows that psychologists would first employ physiological tasks such as grip strength in an attempt to predict such psychological phenomena as intelligence. But Binet's tasks, which were distinctly nonphysiological, better predicted school performance, leaving open the question of why physiological states did not better correspond with psychological ones. These issues remain important 100 years later. Psychologists still search for physiological markers and causes of psychological states, but this has proven to be a difficult task (cf. Cacioppo & Tassinary, 1990). For example, Goldstein and Hersen (1990) maintained that efforts to identify biological markers of most forms of psychopathology have been unsuccessful. Babor et al. (1990) similarly noted the lack of success in identifying biochemical markers of alcoholism. Matarazzo (1992) predicted that intelligence testing would become increasingly linked with physiological measures, particularly those assessing brain activity. He reviewed studies (Hendrickson, 1982; Reed & Jensen, 1991) that found moderate to high correlations between brain activity and intelligence scores. However, rather than demonstrating that biology causes intelligence, as hereditarians since Galton have believed, these studies illustrate the concurrence of brain activity and cognitive processes.

Surprisingly few gains in knowledge have occurred about consistency across human response modes. Psychologists have typically assumed that these modes work as we experience them, that is, as a unified whole, or that only one of the modes, typically behavior, is worthy of study. Regarding the latter, Loewinger (1957) maintained that "the common error of classical psychometrics and naively operational experimental-theoretical psychology has been to assume that only behavior is worth predicting" (p. 688).

Evidence for Mode Inconsistency
Evidence of desynchrony is plentiful. Behavioral, cognitive, and affective measures of fear and anxiety, for example, often demonstrate low to moderate intercorrelations (e.g., Abelson & Curtis, 1989; Craske & Barlow, 1988; King, Ollendick & Gullone, 1990; Leitenberg, Agras, Buiz, & Wincze, 1971; Mineka, 1979). Measures of subjective sexual arousal do not always correspond to physiological measures of arousal (Hall, Binik, & diTomasso, 1985; Henson, Rubin, & Henson, 1979). Some psychotherapeutic and psychopharmacological treatments appear to produce desynchrony (Kinsey & Benjamin, 1984), whereas others do not (McLeod, Hoehn-Saric, Zimmerli,
de Souza, & Oliver, 1990). Desynchrony between cardiac responding and skeletal action has also been observed in animals (Gantt, 1953). In alcohol prevention programs, it is common to find a positive, but low correlation between measures of attitudes toward alcohol consumtion and alcohol consumption itself (Tobler, 1986).

Mavissakalian and Michelson (1982; see also Barlow, Mavissakalian & Schofield, 1980; Michelson et al., 1990) studied patterns of change with 26 agoraphobics who were assigned to different 12-wk treatment programs. They measured clinical, behavioral, and physiological variables at pre-test, during treatment, and at a 1-month follow-up. Mavissakalian and Michelson found that the appearance of synchrony and desynchrony was at least partially caused by the examined interval between measurement periods. In general, behavioral and clinical measures changed most quickly, followed by physiological measures. Hodgson and Rachman (1974) reported similar findings for the order of mode change. The most common form of individual desynchrony during treatment was for self-reports of anxiety to decline while heart rate increased.

Hall et al. (1985) employed 20 heterosexual male college students in a study designed to assess physiological and subjective sexual arousal. While listening to audiotapes describing heterosexual intercourse, subjects moved a dial signifying low to high arousal; simultaneously, penile tumescence was measured by a strain gauge. Subjects demonstrated considerable variability on correlations calculated between physiological and subjective arousal measures. The highest correlations were present for individuals who were both most physiologically aroused and slower to report maximum levels of subjective arousal.

**Causes**

Why does desynchrony occur? Rachman and Hodgson (1974) reviewed three possibilities that remain viable. First, different modes could be linked to different types of reinforcement and reinforcement schedules (Gray, 1971). Thus, some agoraphobics maintain their avoidance behavior by attending to their home as a signal of safety; cognitive and affective states might not be reinforced by those same cues. This explanation has found support in studies that demonstrate that when highly motivated, phobic subjects can perform threatening behavior despite the accompaniment of fear (Hodgson & Rachman, 1974). Second, Schachter's (1966) research indicated that affect could be defined in terms of cognitive appraisal of physiological states. From this perspective, individuals could misinterpret their physiological status, thus resulting in desynchronous measures. Only at high arousal would individuals be unlikely to misinterpret their response modes, a hypothesis that has also received empirical support (Craske & Craig, 1984; Kazdin, 1980; Marks, Marset, Boulougouris & Huson, 1971; Rachman, 1978; Watson, 1988) as well as disconfirmation (Kalupekok & Levis, 1985).

Watson (1988) reported that scales measuring positive and negative affect, often observed to be independent factors, exhibit higher (negative) intercorrelations during periods of greater emotion. Craske and Craig (1984) divided a group of pianists into high- and low-anxious groups and recorded self-report, behavioral, and physiological measures during a performance before an audience. The high-anxious group displayed increased anxiety across measures, whereas the low-anxious group was desynchronous. Such results seem to support a flight-or-fight stress response in which organisms oriented to a threat focus all their resources on coping with the threat (Selye, 1956). Finally, Lang (1971) suggested that verbal reports of affect may be more precise than data produced by measurement of autonomic systems. Intercorrelations between self-reports and physiological measures would be reduced because of range restriction in the latter. On the other hand, Kagan (1988) suggested it may be difficult to translate certain physiological phenomena into natural language items, and this may partially account for difficulties that respondents encounter when answering test items related to affect and physiology.

Psychologists often validate latent traits by correlating behaviors with self-reports of the construct representing that behavior. But Evans (1986) believes that "psychometric principles maintain the fallacy that behaviors are 'measures' of more fundamental underlying entities" (p. 149). Self-reports and behaviors may be organized by a third mechanism: reinforcement (Evans, 1986). Responses can form a cluster because they are under the control of a reinforcement contingency. Cognitions, psychophysiological responses, and overt behaviors "all interact in mutually dependent sub-modes of individual repertoires, not . . . alternative measures of a construct from different 'modes' of behavior" (Evans, 1986, p. 152). Evans suggested that reinforcement can create stable response repertoires across modes (i.e., synchrony) and that similar groupings may be shared by different individuals. In other words, stability that has been attributed to personality traits more accurately reflects, in Evans' view, the stability of environments inhabited by individuals.

Evans (1986) noted the common finding in the psychophysiological literature that individual differences exist in the ability to detect physiological processes. For instance, some individuals may be good judges of their heart rate, whereas others are not. Evans indicated that this variation may be partially explained by different strategies employed to monitor physiology. In a study of the correspondence between self-report and objective measurement of penile circumference during sexual arousal, some subjects appeared to observe their tumescence, whereas others made judgments on the basis of their appraisal of the erotic materials (Evans, 1986). Evans
show greater decreases following a relaxation intervention than persons whose initial rate is lower. For example, persons with high pulse rates are more likely to cause environmental information may be differentially processed and acted upon, independence of modes expands the degrees of freedom afforded individuals. A child moderately frightened by a harmless pet cat can talk himself into touching the cat. An adult mildly obsessed with a thought can often stop the obsession by doing a behavior. A behavior may be delayed or modified while an individual considers its consequences.

Psychological Interventions

Desynchrony poses problems for theorists who devise psychological interventions and for practitioners who implement them. In the past, psychologists examining intervention effects have often assumed that treatments cause uniform effects across modes. Subsequent measurement has often focused only on one mode; misleading results occur, however, if intervention effects differ by mode. Even when multiple modes have been measured, they often did not correlate, thus raising questions about the construct validity of the scales and their constructs.

In psychotherapy outcome research, investigators attempt to change behavior, affect, and cognition as part of a therapeutic intervention (Bergin & Lambert, 1978). The effects of the intervention are assessed by measures of these three modes. The common desynchrony problem encountered in such research is that individuals often change on one, but not all of the measures, and sometimes in unexpected directions. Because the type of change produced by different forms of psychotherapy is difficult to predict, researchers typically include scales that assess all relevant domains (Bergin & Lambert, 1978). Although this enables researchers to describe the varying changes produced by therapy, it does not explain them.

In a study of desynchrony in 21 female agoraphobics, Craske, Sanderson, and Barlow (1987) found that high heart rate was strongly related to positive treatment outcome. Heart rate was associated in these patients with a willingness to approach a feared situation. Craske et al interpreted these results as indicating the importance of patients’ willingness to tolerate intense physical sensations in treatment success. It is also possible, however, that the patients employed awareness of their heart rate as concrete evidence of a phobia that they wanted to overcome. Rather than simply tolerating the physiological signs, these patients may have used those signs as motivation to change.

Assumptions of synchrony by psychological interveners have also been found to influence clinical judgment. Evans (1986) reported a study by Stark (1981) that examined staff ratings of the behavior of an autistic adolescent girl. Staff at the facility treating the girl reported that she frequently had off days in which they believed all aspects of her performance deteriorated and during which they believed it would be unsuitable to provide her with the usual educational programming. Stark found that staff judgments of off days were correlated only with the frequency of echolalic speech
(i.e., involuntary repetition of speech said by others), but not with other measures of task performance. The professional staff, then, made the mistake of overgeneralizing from the speech problems to other modes: they assumed a consistency that did not exist. Chapter’s 3 review will show that this is common problem with interviewers and raters.

**SUMMARY AND IMPLICATIONS**

Consistency assumptions have led psychologists to pursue such questions as, Are psychological dimensions common to all individuals? A better question might be: For what purposes are nomothetic tests best? One answer is to suggest that nomothetic tests seem to make the most sense for selection purposes where few resources are available for training or intervention. Idiographic assessment seems to fit better in intervention contexts with more resources to spend on individual data gathering.

In general, few resources are available in most testing situations, and the result has been a reliance on economical, nomothetic self-reports. The major question surrounding self-reports has been: What would cause individuals to respond inconsistently? A categorization of systematic response errors indicates that individuals employ item-response strategies influenced by such variables as cognitive, affective or motivational, and behavioral or environmental factors.

As shown in the decision tree in Figure 13, the possibility of such variables creating generative response strategies depends upon the degree of match between the test and the test-takers’ cognitive, affective, and behavioral states and traits. If test-takers’ reading levels preclude them from fully understanding items, for example, it is reasonable to assume that they may guess on such items. To the extent that test-takers’ cognitive traits differ from test requirements and purposes—because of language or cultural differences, lack of education and cognitive skills, or insufficient knowledge—even objective tests may function like projective ones. That is, when the mismatch between test items and respondents’ cognitive characteristics is sufficiently large, objective test items become ambiguous stimuli to these respondents, with the result being idiosyncratic associative responses. Such responses are desired in projective tests and error in objective tests.

Even when a cognitive match exists, however, test-takers’ affective characteristics may not. Unmotivated individuals may respond randomly or employ a response style such as acquiescence; highly motivated test-takers may fake good or bad or employ a response set such as socially desirable responding. Finally, factors in the test environment, such as the presence of observers and test administrators as well as their characteristics, may influence responding. Even respondents who have appropriate motivation (e.g., few concerns about doing well or behaving properly) may restrict or alter their responses or behaviors until they become accustomed to the unusual aspects of the environment.

In general, the larger the number and degree of mismatches, the greater will be the use of generative response strategies. Correspondingly, more generative responses means lower validity for the test’s designed purpose. If mismatches occur, it may be possible to intervene cognitively (e.g., explaining difficult words to a less educated test-taker), affectively (e.g., increasing motivation for valid responding by informing test-takers’ that their responses will be checked for accuracy), or behaviorally (e.g., observing unobtrusively or allowing time for adaptation to the observer or administrator). Alternative methods, such as the interview, may be useful for determining mismatch causes or implementing an intervention. For example, listening skills are designed to increase a client’s trust in the counselor, with one intent being that the client self-discloses more accurate information as the relationship develops (cf. Egan, 1990). Finally, it is also possible that type of generative response style (e.g., socially desirable responding) may be indicative of a specific type of mismatch (e.g., affective and motivational errors).

**FIGURE 13** Decision tree of systematic errors for person variables. The use of generative response strategies depends on the degree of mismatch between the test and the test-takers’ cognitive, affective, and behavioral states and traits.
Table 5: Desynchrony as a Result of Systematic Errors

<table>
<thead>
<tr>
<th>Answer</th>
<th>Cognitive</th>
<th>Affective</th>
<th>Behavioral</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>Match</td>
<td>—</td>
<td>—</td>
<td>Synchrony possible</td>
</tr>
<tr>
<td>A2</td>
<td>Mismatch</td>
<td>—</td>
<td>—</td>
<td>Desynchrony</td>
</tr>
<tr>
<td>A3</td>
<td>Match</td>
<td>Match</td>
<td>—</td>
<td>Synchrony possible</td>
</tr>
<tr>
<td>A4</td>
<td>Match</td>
<td>Mismatch</td>
<td>—</td>
<td>Desynchrony</td>
</tr>
<tr>
<td>A5</td>
<td>Match</td>
<td>Match</td>
<td>Match</td>
<td>Synchrony</td>
</tr>
<tr>
<td>A6</td>
<td>Match</td>
<td>Match</td>
<td>Mismatch</td>
<td>Desynchrony</td>
</tr>
</tbody>
</table>

The answer column refers to the particular answer (1–6) to the questions posed in the flow chart in Figure 13.

The sequence of answers to the questions in Figure 13 also has implications for desynchrony. As shown in Table 5, matches and mismatches of tests and respondents can be examined in the cognitive, affective, and behavioral domains. Mismatches within any one mode mean that desynchrony is likely to be evident between that and any other mode. Synchrony can occur only when matches occur first within and then between modes. One solution to desynchrony has been to aggregate scores across modes, but this strategy may produce misleading data if synchrony and desynchrony are the result of systematic and not random errors. If theorists can specify when synchrony and desynchrony should occur in psychological and psychophysiological phenomena, more precise and valid measurement should result.

Desynchrony and systematic errors may be recurring problems partially because test developers and researchers have been guided by assumptions of consistency. Theorists working without consistency assumptions may be more likely to consider the roles and interrelations of cognitive, affective, and behavioral variables when they specify how constructs should be measured. From an idiographic perspective, systematic errors occur when the test developer assumes consistency across individuals, that is, when the test attempts to measure a cognitive, affective, or behavioral characteristic that either is not present or cannot be accessed by an individual. Nomothetically inspired items and tasks will thus lead to mismatches in many individuals and the use of generative response strategies.

If one does not entirely trust psychological self-reports, what then? A logical direction is to employ raters who do not share the biases of subjects and who have some experience or training in gathering information from individuals. In the history of measurement, the most commonly employed method of gathering psychological information has been the interview.

Interviews' greatest advantage and disadvantage is their face validity. That is, the data produced in an interview appear credible to the interviewer and the interviewee. Fremer (1992, p. 4) quoted R. L. Thorndike (1969):

The teacher who is most critical of standardized testing is often endowed with unlimited faith in the accuracy of his own judgments. He knows! It is vitally important that we do not, in identifying the shortcomings of test data, manage at the same time to build up the teacher's view that his own judgment is infallible.
Traditional psychological measurement and behavioral assessment represent one of many schisms found in psychology (Staats, 1983). Proponents of these two approaches tend to be interested in traits or behavior (D. Fiske, 1979). In one world spins traditional measurement, with its emphasis on traits, “real, relatively stable differences between individuals in behavior, interests, preferences, perceptions, and beliefs” (K. R. Murphy & Davidsheimer, 1988, p. 17). In another is behavioral assessment, with its emphasis on psychological states as influenced by the environment. Rarely do the literatures of these worlds mix.
The tradition started by Watson focused on behavior. Behaviorists maintained that a psychological phenomenon was not real unless it could be directly observed. Given such assumptions, there was little need for measurement concepts; whatever could be operationally defined (in the laboratory) or observed (in the clinic), existed. On the other hand, the psychological phenomena of interest to the mental testers were almost always assumed to be unobservable: traits were latent. No single behavior could constitute intelligence, for example, but intelligence formed an abstract, useful construct that could explain clusters of behavior. Intelligence itself was not directly observable, and to this day the most advanced theories of mental testing discuss intelligence in terms of latent traits.

Some initial integration of the two positions has occurred (e.g., Silva, 1993), but the process remains incomplete. Behavioral assessors, for example, have more recently observed that fear is a construct. That is, fear itself cannot be directly observed, but the construct may be useful for certain purposes. In the clinic, many effective behavioral techniques have been developed to alleviate patients' fears. In the course of applying those techniques, however, assessors have observed a desynchrony between different indices of fear. At the conclusion of treatment, it is common for a snake phobic to be able to approach a snake and verbally report no fear, but still possess an accelerated heart rate. Unless we define each phenomenon by its measurement mode, we must resort to the use of a construct to make sense of this situation. As described in Chapter 2, Rachman and Hodgson (1974) cited Lang's (1968) definition of fear as a construct composed of "loosely coupled components" (Rachman & Hodgson, 1974, p. 311) that may covary or vary independently. Because of instances like these, behavioral assessors have become interested in applying traditional measurement concepts in their domain (e.g., Harun, 1984).

As previously defined, traits are enduring, stable psychological characteristics of individuals. In the context of this chapter, traits are assumed to be stable across situations. Thus, persons described as honest are expected to display honest behavior regardless of the situations they find themselves in. Figure 17 displays examples of cross-situational consistency and inconsistency. For example, individuals who score low on a test of honesty may behave dishonestly in classrooms and stores, whereas more honest individuals behave honestly in those settings. In religious situations, however, both high and low honesty individuals may behave honestly. Honest behavior in this case is situation-specific.

Use of the term trait implies that enough cross-situational stability occurs so that "useful statements about individual behavior can be made without having to specify the eliciting situations" (Epstein, 1979, p. 1122). Similarly, D. T. Campbell and Fiske (1959) stated that "any conceptual formulation of trait will usually include implicitly the proposition that this trait is a response tendency which can be observed under more than one experimental condition" (p. 100).

The assumption that psychological phenomena are traits has long guided psychological measurement, and the continued faith in trait measurement has been maintained at least partially by its success. For example, most measurement in vocational psychology is guided by trait-and-factor theory. Typically the tests employed by vocational psychologists measure interests, also assumed to be stable. Test-takers' interest scores are compared with the scores of successful workers in all occupations to determine the best fit. The Theory of Work Adjustment (Dawis & Lofquist, 1984; Lofquist & Dawis, 1969) proposes that to maximize worker satisfaction and production, individuals' abilities and needs must be matched with job requirements and reinforcers. Abilities, needs, job requirements and reinforcers are assumed to be relatively stable. Yet vocational counselors have also been influenced by a developmental concept, that of vocational self-concept crystallization (Barrett & Tinsley, 1977). Developmental theorists have suggested that vocational identities—beliefs about one's abilities and needs—tend to grow and shift until they crystallize, typically around age 18. That is, one's vocational identities become a trait, presumably stable for the remainder of one's life. A similar idea, the differentiation hypothesis ( Anastasi, 1985), has been proposed to account for the emergence in individuals of group factors of intelligence.

As noted in Chapter 2, traits were first employed with the concept of intelligence, a phenomenon assumed to be transmitted through heredity and immune to environmental influences. Although measures of intelligence predicted school performance better than physiological measures, these mental tasks still fell considerably short of the mark of total prediction. If intelligence tasks fail, where is the next logical place to look for variables to assist in prediction? Personality and temperament were psychologists' answer to that question.
Developers of personality and temperament tests closely copied the assumptions and procedures of intelligence tests (Danziger, 1990). With both intelligence and personality tests, scaling involved aggregation, that is, a total score was obtained by adding the total number of correctly performed tasks or endorsed items. Most importantly, personality, like intelligence, was assumed to be consistent across persons and independent of environments.

**THE CONTROVERSY OF MISCHEL AND PETERSON**

But trait-based tests are not entirely consistent. Loewinger (1957) wrote that “circumstances contrive to keep behavior largely unpredictable, however constant its propensities” (pp. 688–689). The importance of environment was emphasized in 1968 with the publication of Mischel’s *Personality and Assessment* and Peterson’s *Clinical Study of Social Behavior*. These books ignited a controversy among measurement and personality theorists that continues to smolder.

Mischel (1968) contended that personality constructs were unstable, that is, the influence of traits was relatively small compared to the influence of situations or environments. He reviewed findings of measurement studies and proceeded to criticize personality psychologists for failing to account for environmental factors when measuring traits: “What people do in all situations and on all tests can be affected, often quite readily, by many stimulus conditions and can be modified substantially by numerous environmental manipulations” (p. 10). Mischel favored measuring behavior in specific situations as opposed to measuring signs of underlying mental processes that could presumably predict future behavior. Mischel believed that new theories of measurement needed to be developed that could account for human adaptability, perception, cognition, self-regulation, and self-modification.

Peterson (1968, p. 23) sounded a similar theme when he stated that research had “suggested very strongly that traditional conceptions of personality as internal behavior dispositions were inadequate and insufficient” because of the influence of situations. Peterson reviewed studies supporting this position in a number of areas, including research documenting the effects of examiners on the behavior of individuals taking projective devices. Whether Rorschach examiners were friendly or distant, Peterson concluded that because “strong positive evidence for validity and utility is nowhere to be seen . . . it looks as if entirely new approaches to the clinical study of behavior will have to be developed” (p. 3).

Mischel’s and Peterson’s publications prompted many psychologists to reexamine trait-based measurement approaches. The violation of the expectation of personality consistency produced three major responses. Traditional theorists sought out more evidence for the consistency of traits. In contrast, some psychologists came to reject intrapsychic traits entirely. Most mainstream psychologists did not follow this extreme direction, however, preferring to search for explanations in the interaction between psychological traits and environments.

**Reinforcing the Trait Argument**

Some contemporary psychologists consider the attack to be repulsed and the battle won (e.g., Block, 1977; Epstein, 1990; Goldberg, 1993). For example, Anastasi (1985) stated that “the long-standing controversy between situational specificity and personality traits has been largely resolved” (p. 134). Anastasi’s solution was to redefine traits as repositories of behavioral consistencies. Traits so defined are not causes, but simply convenient descriptions of psychological regularities that occur and may be influenced by environmental contexts.

Based on studies employing a variety of research methodologies and samples, personality researchers have become increasingly confident that long-term stability of personality traits exists. West and Graziano (1989) concluded that research studies have demonstrated substantial long-term stability of personality in children and adults. They also noted, however, that stability declines across longer measurement intervals, is lower in children, and depends on the particular traits measured. Moreover, predictions of personality from one time point to another typically account for only about 25% of the variance, leaving considerable room for environmental and person–environment influences. Examining the stability of vocational interests, Swanson and Hansen (1988; see also D. P. Campbell, 1971) found similar results: although individual variability and environmental influences existed, trait stability could be demonstrated over time. Similarly, Staw and Ross (1985) studied 5000 middle-aged men and found that job satisfaction remained stable even when employees changed jobs and occupation. In a laboratory study with 140 undergraduates, Funder and Colvin (1991) found behavioral consistency across laboratory and real-life settings, although consistency varied by type of behavior.

Epstein (1979, 1980) proposed that trait inconsistency results from insufficient aggregation of measurement observations. For example, one can aggregate 30 test items into a total score, and this total score is likely to predict a set of criteria better than any one of the individual items. Similarly, one can aggregate scores across different measurement occasions. Although acknowledging evidence that behavior changes as a result of situational variables, Epstein (1979) reviewed research that found that aggregating psychological measurements results in a substantial increase in validity coefficients. In terms of classical test theory, aggregation works because
behavioral consistencies accumulate over multiple measurements whereas random errors do not (Rushton, Jackson, & Paunonen, 1981). Epstein (1979) also conducted a series of studies that found that through aggregation, intercorrelations of measures of behavior, self-reports, and ratings by others could exceed the typical .30 ceiling.

Most contemporary psychologists view the scores produced on intelligence tests as stable and as indicative of latent traits that operate across environments. For example, Schmidt and Hunter’s (1977) work on validity generalization indicated that for general classes of occupational groups, tests of cognitive abilities may have validity across a wide variety of situations. Their research demonstrated that a significant portion of the variability among validity coefficients reported in the literature results from methodological problems such as small sample size, criterion unreliability, and scale range restrictions. When these sources of error are removed, cognitive tests have relatively stable validity within occupational groups. This work complements the position of Mischel (1968) who found that cross-situational consistency existed for behavioral correlates of cognitive abilities. However, validity generalization proponents’ claim of negligible variation over sites has not been universally accepted (Tenopyr, 1992; also see Cronbach’s [1991b] analysis of Hedges’ [1987] data), and the question of situational interactions with cognitive abilities is probably not as closed as many psychologists consider it to be.

**The Rejection of Traits: Behavioral Assessment**

Behaviorists emphasized the dominance of environmental reinforcement in shaping individual’s behavior, be it motor or verbal. In contrast with traditional psychological measurement, behavioral assessors are interested in measuring individuals’ past learning histories and current environmental influences (R. O. Nelson & Hayes, 1986). Behavioral assessors observe behavior in natural or contrived settings and attend to stimuli, behavioral responses, and the consequences of those responses.

The processes, assumptions, and procedures of behavioral assessment differ from traditional measurement. Hartman (1984) emphasized that behavioral assessment is direct, repeated, and idiographic. Assessment is direct in that the psychologist measures observable behavior. Any observed behavior is considered to be a sample of potential behavior, as opposed to a sign of an underlying, unobservable trait (cf. Goodenough, 1950, cited in Cronbach & Meehl, 1955). Behavior is measured repeatedly for the purpose of demonstrating relative stability before intervention and change after intervention, thus demonstrating that the intervention is the cause of the behavioral change. Assessment may consist of continuous recording of behavior (when only a few behaviors occur) or some type of time sampling.

With the exception of areas driven by accountability concerns (e.g., psychiatric inpatients), nonbehavioral psychologists typically do little or no formal measurement during the intervention process.

Behavioral psychologists assess idiographic variables, that is, those unique to the individual in question, such as a behavior, affect, and cognition (B-A-C) mode sequence. Cone (1988) argued that nomothetic, trait-based measurements produce data remote from single cases. He suggested that idiographic instruments will be more sensitive to individual behavior change. In this context, idiographic measures are criterion-referenced (i.e., scores are compared to some absolute measure of behavior), whereas nonnomothetic are norm-referenced (i.e., scores are compared among individuals). Norm-referenced tests are constructed to maximize variability among individuals (Swezey, 1981). However, items that measure behaviors infrequently performed by the population are unlikely to be included in norm-referenced tests. Jackson (1970), for example, suggested that items checked by less than 20% of the test development sample be dropped because they will not contribute to total score variability. Yet those infrequent, idiographically relevant items may be the very ones of interest to change agents and to theorists.

Even the label behaviorists employ differs from traditional measurement. Behavioral assessment, rather than behavioral measurement, is the term employed to describe these measurement approaches because (a) psychologists (and to some extent, clients) typically perform the measurement in a clinical setting, in conjunction with behavioral interventions, and (b) psychologists gather different types of measurement data and integrate them in an assessment.

In contrast with traditional psychological measurement, where anyone can be a self- or other-observer if enough measurements are gathered to decrease measurement error, behavioral assessment involves training observers. Training consists of learning an observation manual (containing definitions of relevant behavior and scoring procedures), conducting analogue observations, on-site practice, retraining, and debriefing (Hartman, 1984; Nay, 1979; Paul, 1986). Hartman (1984) noted that research has indicated that better observers tend to be older, female, and to possess higher levels of social skills, intelligence, motivation, and attention to detail.

Behavioral assessors often express ambivalence about the utility of traditional psychometric analyses. For example, behavioral assessors have begun to attend to validity estimates of what are called higher order variables, such as anxiety and fear, that are likely to have more than one factor influencing their scores. For example, whether measures of eye contact, voice volume, and facial expression all relate to a client’s complaint about shyness (Kazdin, 1985) can be framed as a question of construct validity. Yet Cone (1988) stated that:
Construct validity will be of no concern to behavioral assessors, in one sense since constructs are not the subject of interest, behavior is; in another sense, behavior can be seen as a construct itself, in which case the instrument will have construct validity to the extent that it "makes sense" in terms of the behavior as the client and the assessor understand it. (p. 59).

Cone (1988) also questioned the importance of discriminant validity, saying that it "is not relevant to an assessment enterprise that is built on the accuracy of its instruments. By definition, an accurate instrument taps the behavior of interest and not something else." (p. 61).

Cone (1988) and Pervin (1984) indicated that additional theoretical and psychometric criteria need to be established for behavioral assessment. For example, Cone (1988) proposed that a behavioral measure, to be considered accurate, must be able to (a) detect the occurrence of a behavior; (b) detect a behavior's repeated occurrence; (c) detect its occurrence in more than one setting; and (d) have parallel forms that allow detection of covariation to demonstrate that the behavior can be detected independent of any particular method. Cone also observed that no guidelines currently exist for selecting dimensions relevant to particular clients or for developing instruments to assess these dimensions; interestingly, both of these criteria are strengths of nomothetic approaches (cf. Buss & Craik, 1985). Cone proposed that such guidelines include (a) determining the environmental context of the problem; (b) determining how other people (or models) cope with the problem in that environment; and (c) constructing a template of those effective behaviors to match against the clients' current repertoire. Such a template could be used to guide therapy and as a gauge of therapy's effectiveness.

**Person–Environment Interactions**

Although the importance of person–environment interactions has been recognized for some time (Kantor, 1924; Lewin, 1935; Murray, 1938, cited in McFall & McDonel, 1986), interest has surged in the past two decades. Here behavior and environment are viewed as a feedback loop in which both factors influence the other (Magnusson & Endler, 1977). Instead of focusing on persons or situations, behavior must be measured in context, as a process that occurs in a steady stream. Bowers (1973) noted that from an interactionist perspective, individuals influence their environments as much as their environments influence them. To a significant extent, people create their own environments to inhabit (Bandura, 1986; Wachtel, 1973).

Bowers (1973) approached interaction from the perspective of Piaget’s concepts of assimilation and accommodation. Individuals assimilate observations from the environment into preexisting cognitive schemas. At the same time, those schemas are modified to accommodate new information in the environment. Bowers (1973) stated that “the situation is a function of the observer in the sense that the observer’s cognitive schemas filter and organize the environment in a fashion that makes it impossible ever to completely separate the environment from the person observing it” (p. 328).

Most interactionists assume that cognition mediates the perception of the environment. This is important because it means that behavior that appears inconsistent may actually be indicative of a single construct. For example, Magnusson and Endler (1977) observed that high anxiety may motivate a person to speak excessively in one situation and withdraw in another. The behaviors differ, but the causal construct (anxiety) is the same across situations. As shown in Figure 18, the relation between behavior and construct may be nonlinear. Thus, anxiety may motivate an individual to increase the amount of talking until it reaches a threshold where the individual begins to decrease speech and finally withdraws.

Magnusson and Endler (1977) discussed this type of consistency using the term *coherence*. They suggest that coherent behavior can be understood in terms of the interaction between an individual's perception of a situation and the individual's disposition to react in a consistent manner to such perceived situations. The factors that influence this interaction, such as intelligence, skills, learning history, interests, attitudes, needs and values, may be quite stable within individuals. As shown in Figure 19, individuals C and D, who score highly on a test of honesty, may show more honest behavior across two situations than individuals A and B who obtain low scores. However, C and D may also display differences between themselves in honest behavior across situations—perhaps because of slight differences in their perceptions of those situations—even though their mean behavior score is the same across situations. From the perspective of the individual, the behavior appears coherent. From the perspective of the observer who looks only at group differences, the behavior appears consistent. From the perspective of the observer who looks at individuals across situations, the behavior appears inconsistent.

Appropriate techniques for measuring and analyzing the processes suggested by interactionist theory remain in dispute (Golding, 1975; McFall & McDonel, 1986; Walsh & Beitz, 1985). For example, McFall and McDonel...
Situations

FIGURE 19. Coherent, consistent, and inconsistent behavior patterns. Two individuals (C and D) who score highly on a test of honesty may show more honest behavior across two situations than two individuals who obtain low scores (A and B). However, C and D may also display differences between themselves in honest behavior across situations even though their mean behavior score is the same across situations. (Reprinted by permission from Magnusson and Endler, 1977.)

(1986) stated that (a) analysis of variance (ANOVA) procedures that examine statistical interactions fail to investigate the theoretically central question of how person-situation variables interact over time; (b) investigators can easily manipulate experiments to show the relative importance of person, situation, or interaction factors; and (c) problems of scale remain, that is, no framework exists for how to determine the meaning of different units or chunks of the person-situation process. Bowers (1973) maintained that a rigid adherence to research methodologies has obscured the interactionist perspective. Experimental methods help investigators primarily understand the influence of situations, and correlational methods assist in the understanding of person differences.

Aptitude-by-Treatment Interactions

Treatments can be conceptualized as types of situations (Cronbach, 1975b; Cronbach & Snow, 1977). In a study where an experimental group is contrasted with a control group, both groups are experiencing different types of situations. Persons can also be conceptualized as having aptitudes, that is, individual characteristics that affect response to treatments (Cronbach, 1975a). As shown in Figures 20 and 21, in an aptitude-by-treatment interaction (ATI) study researchers attempt to identify important individual differences that would facilitate or hinder the usefulness of various treatments (Snow, 1991). From a common sense perspective, ATIs should be plentiful in the real world. From the perspective of selection, intervention, and theoretical research, finding ATIs would seem to be of the utmost importance.

ATIs were Cronbach’s answer to the problem of unifying correlational and experimental psychology (Snow & Wiley, 1991). Cronbach (1957, 1975b) noted that the battle over the relative dominance of traits versus environment was maintained by ignoring the possibilities of interactions. For example, Cronbach (1975b) reported a study by Domino (1971), which investigated the effects of an interaction between learning environment and student learning style on course performance. Domino (1971) hypothesized that students who learn best by setting their own assignments and

A PT I T U D E

FIGURE 20 Aptitude-by-treatment interaction—1. In an ATI, the aptitude of subjects interacts with treatments to produce different outcomes. Here individuals high on an aptitude enjoy a much better outcome when completing treatment A than treatment B. (Reprinted from Cronbach, 1957.)

FIGURE 21 Aptitude-by-treatment interaction—2. Here A refers to a group of individuals with high anxiety and B to a group with low anxiety. Situation 2 causes an increase in anxiety for both groups, but the change is disproportionately larger for group B. Note that treatments and situations are interchangeable.
tasks (independent learners) might show the best outcomes in a class when paired with teachers who provided considerable independence. Similarly, students who learn best when provided with assignments by the teacher (achievement through conformity) might perform better when paired with instructors who pressed for conformity (e.g., teachers stressed their own requirements). Domino did find empirical support for this interaction.

But Cronbach (1975b) and others (cf. D. Fiske, 1979; McFall & McDonel, 1986; Scriven, 1969) have largely abandoned the search for general laws via ATIs in favor of local, descriptive observations. Cronbach (1975) noted that results supporting ATIs are inconsistent, often disappearing when attempts to replicate occur. He saw time and history as the major culprits: many psychological phenomena change over time, frustrating attempts to fix them in terms of general laws. Cronbach (1975) indicated that trait conceptions do not hold to the extent necessary to demonstrate consistent ATIs. This may also be interpreted as support for the position that situations have stronger effects than traits or trait-situation interactions.

I conducted unpublished research that demonstrates the difficulty of finding ATIs. I investigated whether college students' comfort with computer use interacted with one of two interventions designed for alcohol education. The interventions were a computer-assisted instruction (CAI) program for alcohol education and a set of written materials on which the CAI program was based. The interaction hypothesis suggested that students most comfortable with computers and who completed the CAI program would demonstrate the greatest improvement. Students were pretested and posttested on measures of alcohol attitudes, knowledge about alcohol, and alcohol consumption. Given the resources necessary to run the interventions, data were collected over a period of several years, with preliminary analyses occasionally conducted. Analyses at various stages of data collection found the expected interactions, but when the n per cell reached about 15, only an alcohol attitude scale showed a significant interaction. Given these results and a change in my access to a major source of research subjects, the project was suspended.

Environmental Assessment

Instead of searching for stability in individuals, another group of theorists and researchers sought to find consistency in environments and situations. Attempts to categorize and measure environments form the essence of this measurement approach (Conyne & Clack, 1981; Walsh, 1973).

One of the major tasks of environmental assessment is an analysis of environment types, and many classifications systems have been proposed to accomplish this task (cf. Goodstein, 1978; Huebner, 1979; Steele, 1973). Conyne and Clack (1981) proposed that an environment consists of physical, social, institutional, and ecological-climate components that shape and are shaped by people. Moos (1973) classified human environments into ecology, behavior setting, organizational structure, inhabitants' behavior and characteristics, psychosocial climate, and functional reinforcements (i.e., environmental stimuli).

Many vocational psychologists hold a similar assumption about the relative stability of work environments. An occupational setting may attract certain types of individuals on the basis of the setting's fit with the needs and abilities of the worker. One of the best-known and well-researched occupational classifications has been proposed by Holland (1959, 1985). He suggested that work environments may be classified as involving one or more of the following dimensions:

1. Realistic environments, where work entails mechanical skill, physical strength, motor coordination, and concrete problems
2. Investigative environments, with an emphasis on research activities, scientific accomplishments, mathematics ability, and abstract problems
3. Artistic environments, involving artistic activities and competencies, and an emphasis on expressive, original, and independent behavior
4. Social environments, where work involves social interactions, liking others, cooperation, and helping
5. Enterprising environments, involving selling and leading activities, self-confidence, aggressiveness, and status
6. Conventional environments, involving recording and organizing records and data, conformity, and dependability

Holland (1985) believed, however, that individuals' characteristics may change the climate of the work setting. The most important variable in this regard is the extent to which an individuals' needs and abilities are congruent with the work environments in which individuals find themselves. Very incongruent individuals leave environments, whereas moderately incongruent individuals will change, moving toward the dominant persons in the environment.

As shown in Table 7, most person–work environment fit theories suggest that the degree to which individuals fit their work environment determines their level of productivity and job satisfaction. Thus, realistic individuals working in realistic occupations will be most productive and satisfied, Investigative individuals in investigative occupations, and so on. Holland's theory provides for similarity of occupational types (e.g., investigative and artistic occupations are more similar than investigative and conventional) so that different fits may be ranked ordered in terms of their degree of expected productivity and satisfaction.

A crucial question in environmental assessment is whether to classify environments or perceptions of environments. The person many consider the founder of person–environment interaction, Kurt Lewin, continued a Gestalt perspective in which behavior was believed to occur in the context...
of an individual's total perceptual field of an environment (Lewin, 1951). That is, people are surrounded by a self-generated psychological environment and a nonpsychological environment. As in other person-environment theories, cognition has been proposed as a significant mediator of how environmental events are perceived, understood, and transformed by individuals (Bandura, 1986; Conyne & Clack, 1981). From this perspective, an understanding of how an individual thinks about a situation is necessary for a person-environment analysis. For example, R. B. McCall (1991) observed that after completing a Marine Corps' confidence course, recruits might view the course as a confidence builder or as intimidating. Some theorists hoped that cognition might prove to be stable across situations, but research results have not been supportive. For example, attributional style—characteristic ways individuals explain and interpret life events (S. Fiske & Taylor, 1984)—appears to possess little consistency across situations (Bagby, Atkinson, Dickens, & Gavin, 1990; Cutrona, Russell, & Jones, 1984).

What kinds of measurements are undertaken in environmental assessment? Conyne and Clack (1981) provided several examples. Cognitive maps are spatial representations of individuals' psychological environments. Conyne and Clack described a researcher who instructed students to plot where in their neighborhoods they felt high and low stress. The resulting map helped to explain truancy by showing that a city school bus route stopped at many high-stress areas where students were afraid of being physically attacked. Geographic maps can be used to locate individuals with psychological characteristics and events (e.g., academic achievement, depression) to examine potential relationships between environment and person. However, many environmental assessment procedures consist only of self-report questionnaires that ask respondents to rate environments along different theoretical dimensions. In Moos' (1979a, 1979b) social climate scales, for example, individuals rate such environments as college residence hall, classroom, family, and work along such dimensions as relationships, personal growth, and system maintenance and change. Similarly, vocational psychologists typically measure environments by assessing (via self-reports) the interests and abilities of persons successful in specific occupations.

**Moderators of Cross-Situational Consistency**

Moderator variables are those that change the nature of the relation between two other variables. For example, one may propose that investigative individuals would be most productive and satisfied in an academic or scientific environment as compared to other occupational situations. However, one might find that other variables, such as ethnicity or gender, affect that relation. Female and male academics might produce equal number of publications, but females might also experience less job satisfaction.

A variety of potential moderating variables have been proposed and investigated. Research has suggested differences in consistency by response mode and by levels of aggregated measures (Diener & Larsen, 1984; Epstein, 1983; Mischel & Peake, 1982; Rushton, Brainerd, & Pressley, 1983). For example,Violato and Travis (1988) found that male elementary school students demonstrated more cross-situational consistency on the variable of behavioral persistence. Similarly, Connell and Thompson (1986) found infants' emotional reactions were more consistent across time than their social behavior. Variables such as age (Stattin, 1984), gender (Forzi, 1984), socioeconomic status, and cognitive abilities (Violato & Travis, 1988) have also been found to moderate cross-situational consistency.

Bem and Allen (1974; also see Diener & Larsen, 1984; Lanning, 1988; Zuckerman et al., 1988) suggested that individuals themselves are moderators of cross-situational consistency. That is, some persons may act consistent across situations, and others may not; cross-situational consistency could be considered an individual difference variable. Thus, in person A, the trait of honesty is manifested across all situations; with person B, honesty occurs only at church; and person C exhibits little honesty in any situation. Investigators who conduct studies averaging these individuals would find no support for cross-situational positions, but a disaggregation of the data might demonstrate such consistency for pairings of similar individuals such as A and C. Bem and Allen (1974) found that students' ratings of their cross-situational consistencies often did match their behaviors in different situations. For example, students who said they were friendly across situations did show more consistency. However, more recent research has provided mixed support for this position. Chaplin and Goldberg (1984) failed to replicate Bem and Allen's results, whereas P. A. Burke, Kraut and

---

**TABLE 7 Rank Order of Expected Productivity and Satisfaction Indices by Person-Environment Fit**

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>I</th>
<th>A</th>
<th>S</th>
<th>E</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

*R refers to realistic occupations, I to investigative, A to artistic, S to social, E to enterprising, and C to conventional. The hexagon at right shows the degree of similarity among these different occupational types; adjacent occupations are more similar. Thus, R occupations are more similar to I and C than to S. The matrix at left displays the resulting rank order of expected productivity and job satisfaction indices based on the similarity of person-work environment fit.*
Dworkin (1984) found that subjects' ratings of personal traits, cross-situational consistency, and the traits' importance to their self-schemas were highly correlated. Greener and Pennebaker (1982) found a low reliability estimate for the 1-item consistency measure previously employed by Bem and Allen.

**States and Traits**

Psychologists have also suggested that psychological phenomena may be manifested through traits and states. States are transitory psychological phenomena that change because of psychological, developmental or situational causes. Spielberger (1991) credited Cattell and Scheier (1961) with introducing the state–trait distinction. Even theorists interested in measuring traits acknowledge the presence of state effects in psychological testing. F. M. Lord and Novick (1968), for example, observe that in mental testing "we can perhaps repeat a measurement once or twice, but if we attempt further repetitions, the examinee's response changes substantially because of fatigue or practice effects." (p. 13).

State constructs have occasionally been invoked to explain the inconsistencies found in psychological measures. For example, Matarazzo suggested that the MMPI measures states, not traits, and that it "reflects how you're feeling today, or how you want to present yourself that day" (Bales, 1990, p. 7). Similarly, Dahlstrom (1969) indicated that MMPI scales' reliability should not be assessed through test-retest methods because "there is scarcely any scale on the MMPI for which this general assumption [of temporal stability] is tenable for any period of time longer than a day or two" (p. 27). Even intelligence may possess state properties: IQ test scores have been shown to be affected by amount of schooling (Ceci, 1991; see also Bandura, 1991, and Frederiksen, 1986). Retest of intelligence scores at 1-yr intervals show high stability but decrease substantially when the retest interval lengths beyond that period (Humphreys, 1992).

Although trait conceptions have historically dominated measurement, psychologists have always recognized states and struggled to integrate them into measurement theory and practice. Loevinger (1957), for example, reviewed D. W. Fiske and Rice's (1955) theory for explaining intraindividual response variation. Fiske and Rice suggested that individuals will change their response to an item because

1. something changes within the individual (e.g., the individual matures),
2. the order of item presentation changes, or
3. the stimulus situation changes.

Fatigue, for example, is a physiological variable that might produce inconsistent responding on psychological tests. D. W. Fiske and Rice (1955) suggested that such variability is lawful, although Loevinger (1957) noted that such a belief is at variance with the classical test theory assumption that these errors are random. Loevinger also reviewed studies that demonstrated, over retests, improvements in personality functioning and intelligence. Such improvements, Loevinger suggested, are a function of practice effects and learning by the organism. In other words, psychological phenomena treated by traditional measurement approaches as traits may also demonstrate state effects.

Perhaps the most well-known state–trait measure is the State–Trait Anxiety Inventory (STAI; Spielberger, Gorsuch, & Lushene, 1970). The STAI consists of two 20-item Likert scales to measure state anxiety (i.e., situation-specific, temporary feelings of worry and tension) and trait anxiety (i.e., a more permanent and generalized feeling). Both scales contain items with similar and overlapping content: state scale items include "I am tense," "I feel upset," and "I feel content," whereas trait scale items include "I feel nervous and restless," "I feel secure," and "I am content." However, the state scale asks test-takers to rate the items according to how they feel "at this moment," whereas the trait scale requests the ratings to reflect how the test-takers "generally" feel. The instructions do seem to produce the desired difference: test–retest reliabilities for the state scale, for example, are considerably lower than for the trait. Spielberger (1991) also developed the State–Trait Anger Expression Inventory. Spielberger described state anger as a condition that varies over time as a result of such factors as perceived injustice and frustration. He distinguished state anger from trait anger, the latter being a disposition to perceive many situations as frustrating or annoying and to respond to those situations with state anger. Thus, the two constructs are related: persons high in trait anger will experience more frequent state anger.

**Process Research**

Psychotherapy researchers often distinguish between outcome research, designed to test the efficacy of various interventions, and process research, designed to detect variables that change during the intervention. Process research typically consists of single-subject, within-subject, or between-subject designs that assess such variables as counselor verbal behavior and aspects of the client–counselor relationship (Heppner et al., 1992).

Process research focuses on changes in the counselor, client, or counselor–client relationship within (e.g., treatment sessions) and across (e.g., different types of treatments) situations. The hope is that these changes will be related to intervention outcome, although it has been difficult to demonstrate a strong process–outcome link (Elliott et al., 1987; Heppner et al., 1992; Hill et al., 1988). The failure to show such a relation may partially result from researchers' emphasis on studying counselor variables to the exclusion of client variables (Heppner et al., 1992; Hill, 1982). Heppner
et al. (1992) observed that process research typically assumes the client to be a passive agent instead of an active information processor.

Measurement of process variables is performed by trained raters who assess segments or all of the counseling sessions. How much to measure has been one of the central questions of process research, and it appears that the purpose of the research can suggest answers (Friedlander et al., 1988; Heppner et al., 1992). For example, if the researcher is interested in process variables that apply across groups of counselors and clients, small segments of sessions (totaling as little as 10% of the session) have been found to be representative of the process in groups. If the researcher is interested in a single case, however, it appears necessary to sample entire sessions.

**SUMMARY AND IMPLICATIONS**

Situational influences on human behavior have been the most noticed sources of trait inconsistency. In response, psychologists have proposed the following:

1. Reinforcing arguments that traits exist, as in work on aggregation
2. Rejecting traits, as in behavioral assessment
3. The existence of person–environment interactions, where traits and environments influence each other in a continuous system
4. ATI, where aptitudes interact with environments, the latter conceptualized in terms of treatments
5. Environmental assessment, involving classification of environmental types
6. Moderators of cross-situational consistency, variables that facilitate consistency of traits across situations
7. Psychological states, constructs defined as variables that change over time
8. Process research, where researchers attempt to isolate and study variables that change during psychological interventions

Many psychological phenomena demonstrate trait and state characteristics. In other words, these phenomena are likely to demonstrate some stability (that is, to be a reflection of psychological traits) and change (that is, to be influenced by environmental and developmental factors). Goldberg's (1993) question, "Do traits exist?" may be rephrased as Cone's (1991), "What levels of aggregation of tests and criteria are needed to demonstrate trait properties?" To demonstrate trait consistency, it appears necessary to aggregate items, persons, and occasions of measurement. In addition, some people and dimensions appear more stable than others (Bem & Allen, 1974; R. P. Martin, 1988).

Idiographic approaches appear to have been more successful than nomothetic procedures for displaying cross-situational consistency. For example, C. G. Lord (1982) found that idiographic measures of conscientious behavior were consistent across situations, whereas nomothetic methods were not. Walsh and Betz (1985) maintained that behavior is "reasonably predictable, given knowledge of an individual's perception of the situation and of the individual's disposition to respond in that situation" (p. 13). The problem may center primarily on handling error; so many unknown factors operate in the interaction between person and situations that only repeated idiographic assessment may provide some sense of how any particular individual will behave across situations. Under these conditions, Magnusson and Endler suggested (1977, p. 11) that "individual behavior across different situations provides a consistent, idio graphically predictable pattern." Clinical assessment, for example, provides a unique perspective on individual and situational factors. However, clinical assessors still cannot, for example, make predictions of person–environment interactions that would result in suicidal or homicidal behavior by a specific individual at a particular time.

Despite their initial promise, approaches such as ATIs and person–environment interaction theories have had relatively little impact on psychological measurement. The problem, Walsh and Betz (1985) believe, "is one of measuring and describing the multidirectional transactions. Currently this is a measurement task that has been very difficult to operationalize and make real" (p. 13). Measurement error is a plausible explanation for the failure to find person–environment interactions (Cronbach, 1991b). Imprecise measurement may obscure interactions, and measurements not guided by theory may lead to inappropriate or insensitive measurement (R. B. McCall, 1991). Partially because of cost and the fact that investigators have yet to settle upon a taxonomy for situations (McFall & McDonel, 1986), person–environment approaches have yet to command the attention given to traditional measurement.

McFall and McDonel (1986) saw the person–situation question as "inherently unresolvable" (p. 238) and suggested that the debate should be dropped. Instead, psychologists should continue elsewhere their search for the "meaningful units with which to describe, predict, and explain behavior" (p. 238). Tryon (1991) reached a similar conclusion. He maintained that situational specificity simply involves different mean levels of behavior in one situation compared to another. As shown in Figure 22, individuals may demonstrate, in general, more anxiety in two testing situations (situations 2 and 4) than in two lectures (situations 1 and 3). Arguments for traits, Tryon suggested, involve persons maintaining their rank on a construct within the distribution. If Persons A, B, C, and D rank first, second, third, and fourth on the amount of anxiety they display in a lecture, and then maintain that ranking in other settings, trait arguments will be upheld.
even if overall levels of anxiety change. Tryon (1991) believed that it is possible to hold both the situational specificity and trait positions because “activity is both very different across situations yet predictable from situation to situation” (p. 14). He concluded:

Situational differences are so large that they stand out immediately. Person consistency is more subtle and requires aggregation to reach substantial effect size. The Spearman-Brown prophecy formulate indicates that either effect size can be made arbitrarily large depending upon the level of aggregation chosen. The implication for research and clinical practice is that one should choose the level of aggregation that provides the necessary effect size to achieve the stated purpose of the empirical inquiry at hand. (p. 14)

HISTORY

According to the classical definition of validity, a valid test is one that measures what it is supposed to measure; Cronbach (1992) reported that such a definition was officially noted in a 1921 publication by a group that became the American Educational Research Association (AERA). From this perspective, a test of honesty that actually measures honesty is valid; an honesty test that measures social desirability and honesty is less valid; an honesty test that measures only social desirability is invalid. Validity has been contrasted with reliability: Reliability is the consistency of scores on the same test, whereas validity is the consistency of scores between related tests (D. T. Campbell & Fiske, 1959). Similarly, Mischel (1968) defined reliability as agreement between tests under maximally similar conditions, whereas validity is agreement under maximally dissimilar conditions. It is
CONCLUSIONS AND SPECULATIONS

Accepting any successful paradigm carries with it benefits and costs. The promoters of a successful paradigm will obviously emphasize its benefits, but eventually other groups will attempt to demonstrate and publicize the accompanying costs. To the extent that a paradigm cannot explain the phenomena of interest, calls for revisions and modifications, and revolutions are likely to be increasingly heeded.

A paradox exists between the sharp criticisms of psychological measurement by academic scientists and others and the widespread use and acceptance of current tests by practitioners and researchers. The paradox may be understood by noting that psychological measurement devices appear adequate for some purposes, but not for others. For example, selection testing is one of psychology’s most important societal functions. In education, business, and the military, psychological measurement, particularly cognitive ability tests, represent the best and fairest approach to selection decisions. That success led psychologists to apply selection assumptions and procedures to tests in other areas, such as explanatory theory building and intervention, where the results have been less successful. Thus, ample reason exists both to respect contemporary tests and to work toward their improvement.

I disagree with those who believe or imply that important psychological constructs are largely immeasurable (cf. L. Goldman, 1990; Mayer, 1966). The evidence does not support Faust and Ziskin’s (1988) assertion that human behavior “resists objective, direct, or reliable observation and measurement” (p. 33). No other set of procedures exceeds the utility of psychological measurement and assessment in their proper applications. To improve measurement’s effectiveness, what is required is a renewed emphasis on theoretical and technological development, application of existing substantive theories to measurement procedures, experimentation with new procedures, and programmatic research. In turn, these goals should be supported by increased grant funding, additional attention to measurement and assessment in graduate coursework, and greater attention to these issues in scholarly publications. I am astounded at how little grant support is available for measurement and assessment research, particularly from nonmilitary sources. For example, a 1970 plan to revise the MMPI—

1Portions of this chapter are reprinted from Meier (1993) by permission of the American Psychological Association. © 1993 by the American Psychological Association.
probably the most important noncognitive psychological test in the world—was delayed until recently partially because of a lack of funds (Graham, 1990).

The issue of sufficient resources to study measurement and to do applied assessment is a “what-comes-first?” problem. Without the additional resources necessary to improve measurement and assessment, little progress in these areas can be expected. Yet better measurement and assessment will be expected by those who will judge our requests for additional resources in science and practice. Economics is likely to be a critical factor in applied measurement innovations: for example, as long as costs are the primary factor in deciding the extent to which assessments should be conducted in practice settings, the most inexpensive methods, self-reports and interviews, will continue to reign. It is ironic that the managed care companies now dominating psychological practice, who publicly profess to value solid evaluations of health care, show so little interest in paying for psychological assessment, research on assessment innovation, br training practitioners in effective assessment.

Dawis (1987) maintained that “researchers should not be reluctant to experiment with different scale construction approaches—and should report their results, so that the rest of us can find out what method is best” (p. 488). D. T. Campbell and Fiske (1959) expressed similar hopes:

Psychologists today should be concerned not with evaluating tests as if the tests were fixed and definitive, but rather with developing better tests. We believe that a careful examination of a multitrait-multimethod matrix will indicate to the experimenter what his next steps should be: it will indicate which methods should be discarded or replaced, which concepts need sharper delineation, and which concepts are poorly measured because of excessive or confounding method variance. (p. 105)

Sidman (1960) observed that much important work in experimental psychology is devoted to “… improvements in measuring instruments, advanced methods of recording data, sophisticated data analysis, the design of specialized apparatus to do a particular job or generalized apparatus to perform many functions, and the extension of old techniques to new areas” (p. 16). Sidman (1960) also reported that Skinner suggested that apparatus (and measurement) failures could be a source of new discoveries.

When measurement and assessment advance, my guess is that testing will move away from its strict reliance on self-report. For example, several researchers have begun to investigate the potential benefits of combining self-reports with the response latencies of those reports (e.g., R. R. Holden et al., 1992; Scheidt & Meier, 1993). Test construction will focus not just on maximizing individual differences, but on other criteria as well. For example, test developers might construct a measurement device containing items and tasks that differentiate between adolescent, minority women who respond and fail to respond to a particular psychological intervention. If they are theory-based, items and tasks retained and dropped during test development will contain information relevant for theory revision. I believe that scale construction will become more specialized and closely linked with substantive and measurement theories. Different types of instruments will be designed for different tasks.

Kuhn (1970) implied that one essentially gambles when choosing between paradigms. My bets are on the theories and procedures, such as cognitive approaches and GT, that pay the most attention to construct validity and construct explication. I suspect that the validity procedures proposed by D. T. Campbell and Fiske (1959) and Cronbach and Meehl (1955) won over other proposals because their methods dealt with measurement at the level of the total score. In contrast, proposals by Loewinger (1957) and Jessor and Hammond (1957) to evaluate construct validity at the item level received less attention because no theory or methodology was available to handle the variability of responses at the item level. Cognitive and cognitive-behavioral theories are now available that have the capacity of guiding investigations at the item level.

Although students still require instruction in such basics as reliability, validity, norms, and item analysis, recent developments argue for a revision and possible expansion of the measurement curriculum. I would argue for greater attention to approaches that emphasize construct validity and construct explication. To diminish the artificial science-practice split in the graduate curriculum, faculty should also increase the measurement emphasis in applied and practice courses. Instructors could construct assignments in which students apply basic and advanced measurement procedures to important problems in their subspecialties. Such practical applications might increase students’ interest in measurement and assessment and intrigue some students sufficiently to investigate the utility of new measurement procedures in their areas. Few students realize that the investigator-developed scales frequently employed in basic research (Meier & Davis, 1990) lack sufficient evaluation in terms of reliability and validity, thus presenting alternative methodological explanations that weaken confidence in research results. In general, measurement instructors should teach students how to (a) evaluate the psychometric properties of measurement devices employed in published studies and (b) routinely evaluate and report the psychometric properties of students’ measurement devices.

Throughout the history of the science and the profession, psychologists have been presented with a pressing need for useful measurement and assessment. At present many psychological measures produce data comparable to a map that contains the names and relative locations of towns and villages, but little information about scale or the surrounding topography. The practice of measurement and assessment will continue, and without substantial improvement, so will the crises. Faust and Ziskin’s (1988) controversy about the adequacy of psychological assessment and diagnosis as presented by expert witnesses in legal testimony is now moving away from
center stage in favor of questions about clinicians and their clients' abilities to reliably distinguish between repressed memories and false ones (Loftus, 1993).

Cronbach (1992) concluded a presentation on the history of validation in cognitive ability testing by observing that "there has been progress. We can look forward to more." I believe that significant progress in measurement and assessment is the single most important prerequisite for unification of psychological science and practice.

APPENDIX
10 DATA SETS USED FOR SIMULATING RANDOM RESPONDING