GLOSSARY OF MAJOR TERMS

Administration

Definition. Process by which a test taker completes a test.

Description. In most testing situations, the administrator's primary job is to ensure *standardization* (i.e., the establishment of similar test procedures) of the testing environment. Worthen et al. (1993) suggested several guidelines for administered tests, including (a) checking the physical setting for appropriateness (e.g., adequate lighting, temperature); (b) ensuring that participants know what they are supposed to do; (c) monitoring the test administration; and (d) following any standardized instructions carefully (e.g., as provided with a published test). Test takers, however, bring their unique individual differences with them to the testing situation—some of which complicate the standardization effort. For example, Kahn and Meier (2001) found that how individuals defined the construct they were asked to report (i.e., power in a family) influenced the scores they actually reported. Relatedly, completing a test on a computer may present challenges to individuals who lack computer experience.

Recall that I employ *test* generically, that is, to mean any type of measurement and assessment device. How a test is administered at least partially distinguishes between measurement and assessment types. In *self-reports* the participants themselves read and respond to items. In *interviews* the assessor reads items/questions to participants. Fewer resources is the advantage for self-reports (i.e., you do not need an interviewer), whereas greater depth of understanding (i.e., you can ask respondents to elaborate and they can ask you to clarify) is an advantage of interviewing.

For example, Blais, Norman, Quintar, and Herzog (1995) compared two methods of administering the Rorschach projective test (i.e., the Rapaport and Exner systems). The Rorschach consists of administration of 10 inkblots designed to provide ambiguous stimuli. Rapaport and Exner administrations, which differ mainly in the examiner – examinee seating arrangements and questioning instructions, were randomly assigned first to 20 women with bulimia. Significant differences were found between the two administration systems, with Exner producing more color and shading responses. Interestingly, system differences were most prominent on the first presentation of the two administrations. Other research has also shown that Rorschach scores can be changed because of administrators' differing instructions (Exner, 1986).

Administrator – respondent relationship

Definition. The degree of rapport and trust established between the test administrator/interviewer and the person taking the test.

Description. Traditionally, the relationship between the administrator and test taker has been placed in the background. However, test developers and publishers urge administrators to establish rapport with test takers, but seldom is the presence of this rapport assessed or monitored (cf. Worthen et al., 1993). Research has been conducted to examine the effects of administrator characteristics on respondents. Little attention has been paid to the relationship, however, because test theorists and developers usually do not consider the relationship an important factor.

In qualitative assessment, the relationship is assumed to influence the honesty and accuracy of information shared by the test taker (Strauss & Corbin, 1997). That is, to the extent that the test taker trusts the administrator, the test taker is more likely to make an effort to produce reliable and valid information.

One way of approaching the issue of administrator and interviewer effects is to compare traditional testing administration to situations where little or no administrator – test taker interaction occurs. For example, are tests administered or introduced by a person equivalent to computer-administered tests and interviews? In other words, does the automation of test procedures affect the method's reliability and validity? Some researchers have found no differences between traditional and computer-administered versions of tests (e.g., Calvert & Waterfall, 1982). However, some who take computer – administered tests alter their rate of omitting items (Mazzeo & Harvey, 1988) or increase their faking good responses (Davis & Cowles, 1989). Students who have recently taken the computer-administered version of the GRE or similar tests should compare their experiences to other testing situations. Given the equivocal research findings, the equivalence issue currently must be considered on a test-by-test, sample-by-sample basis.

Aggregation

Definition. Summing or averaging of measurements.

Description. Aggregation often improves the reliability and validity of measurements because random measurement errors cancel or balance each other (Rushton, et al., 1983, 1981). Even if systematic errors are present, if they are of a sufficiently different type, they may offset each other. In most instances, then, an aggregated score should reflect the construct of interest better than any one item.

One problem with aggregation is that you may sum incompatible sources. For example, you may be interested in studying parents' ratings of their children's behavior. It may be that mothers, as compared with fathers, have more experience with their children and thus can provide more valid data. Adding the fathers' data to mothers' may be introducing a source of error.

Epstein (1979; see also Martin, 1988) provided examples of the benefits of aggregation. Epstein asked 45 undergraduates to keep daily records, for 14 consecutive days, of such behaviors as number of social phone calls made, social contacts, headaches, hours of sleep, and similar constructs. Epstein found that the average correlation of these constructs for 1 day with data provided for the 13 other days was quite low (e.g., .09 for hours slept). That is, little relationship existed between behavior on any 1 day and behavior exhibited on the other 13 days. To demonstrate the effects of aggregation, Epstein summed scores for the even and odd days and correlated these groups. For every behavior measured, the aggregated correlations exceeded the 1-day correlations. For example, the correlation between even and odd days for hours of sleep was .84.

Aptitude-by-treatment interactions (ATIs)

Definition. Interaction of individuals' characteristics with interventions.

Description. Treatments and interventions such as counseling and psychotherapy can be conceptualized as special types of situations or environments (Cronbach, 1975a; Cronbach & Snow, 1977). In an study where an experimental group is contrasted with a control group, the 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). In an ATI study researchers attempt to identify important individual characteristics or differences that would facilitate or hinder the usefulness of various treatments (Snow, 1991). A computer-based mathematics course or any type of distance learning course would probably be most beneficial, for example, to students who are comfortable and knowledgeable about technology.

From a commonsense perspective, ATIs should be plentiful in the real world. That is, it seems reasonable to assume that persons with certain characteristics should benefit more from some treatments than others. From the perspective of selection, intervention, and theoretical research, finding ATIs would seem to be of the utmost importance. ATIs offer the possibility of increased efficiency in these applied areas. For example, Domino (1971) investigated the interaction between learning environment and student learning style. Domino hypothesized that independent learners, students who learn best by setting their own assignments and tasks, might show the best outcomes in a class when paired with teachers who provided considerable independence. Similarly, conforming students who learn best when provided with assignments by the teacher might perform better when paired with instructors who stressed their own requirements. Domino did find empirical support for this interaction.

Assessment

Definition. Human judge's combination of data from tests, interviews, observation, and other sources.

Description. Assessment is a broader term than *measurement* and includes any measurement method that involves human judgment. A reading specialist, for example, might make an assessment of a child's reading problems on the basis of data from standardized test scores, observation of the child during class time as well as during test taking, and interviews with the child, parents, and teachers. In a clinical or psychological context, assessment information can include a history of the presenting problems, family background and social supports, current medical problems, education, current employment, and financial resources. Aiken (1996) listed references for assessments developed for specific groups, including children (Weaver, 1984), adolescents (Harrington, 1986), and adults (Swiercinsky, 1985), as well as settings, including education (Levy & Goldstein, 1984), psychology (Robinson, Shaver, & Wrightsman, 1991), and mental health (Comrey, Bacher, & Glaser, 1973).

Martin's (1988) work suggests that researchers must design their assessments to control unwanted factors in the information-gathering process. Ideally, assessors should use multiple settings (e.g., places where an individual or group is observed or assessed), multiple sources

(e.g., parents and teachers), and multiple methods or instruments (e.g., observations, tests, and interviews) to minimize error. Assessors must also pay attention to test taker characteristics that may influence scores, including motivation to respond accurately, age, ability to read, socioeconomic status, cultural and family values, and motor ability.

Neuropsychological assessment involves use of tests and observation for the purpose of diagnosing brain dysfunction (Gregory, 1992). Gregory (1992) described the neuropsychological assessment of a college junior who reported the onset of poor performance in a premed curriculum after 2 years of good grades. Administration of the Weschler Adult Intelligence Scale —Revised (WAIS-R) found an IQ of 122, but the student could not accurately copy a simple geometric cross, showed large differences in fine motor control (e.g., finger tapping) between the two sides of his body, and performed poorly on a measure of abstract reasoning. Gregory (1992) reported that the copying difficulty and left hand motor slowing was indicative of right hemisphere impairment and problems with spatial relationships. A CT scan confirmed a lesion in the frontal – pariental lobe of the right hemisphere. The student changed majors to history and graduated with a degree in education, a switch Gregory (1992) believed made more sense, given the student's strengths in the left hemisphere. In this instance, the combination of testing, observation, and history taking constituted *assessment*.

Change-based measurement

Definition. Tests whose primary purpose is to detect change in one or more constructs.

Description. Change-based measurement is intended not to detect stable traits, but states and other conditions such as moods or skills that change over time and in different situations. As noted previously, testing traditionally has focused on measuring traits such as intelligence that were assumed to be largely a function of heredity and immune to situational, developmental, and intervention influences. Attempts to measure traits affected how tests were constructed; reliability and validity became the central criteria for evaluating a test's quality (Meier, 1997, 1998). Efforts to develop tests whose purpose is to be sensitive to intervention and developmental effects are relatively new.

In contrast, Meier (1997, 1998, 2003, 2004) drew on the concepts described by criterion-referenced and longitudinal test developers (Collins, 1991; Gronlund, 1988; Tryon, 1991) to develop test construction rules (intervention item selection rules, IISR, described in Chapter 6) designed to select test items and tasks sensitive to intervention effects. Intervention items, like traditional items, should also be theoretically based, unrelated to systematic error sources, and avoid ceiling and floor effects. Because empirically derived items may be capitalizing on sample-specific variance, items should be cross-validated on new samples drawn from the same population. Intervention-sensitive items, however, should possess several unique properties, foremost of which is that they should change in response to an intervention and remain stable over time when no intervention is present.

Meier (1998), for example, conducted a comparison of traditional and IISR rules with an alcohol attitudes scale completed by college students in an alcohol education group and a control group. The intervention and traditional item selection guidelines produced two different sets of items with differing psychometric properties. The intervention-sensitive items did detect pre –

post change; these items also possessed lower test – retest reliability in intervention participants while demonstrating stability when completed by controls. In contrast, items evaluated with traditional criteria demonstrated greater internal consistency and variability, characteristics that enhance measurement of stable individual differences. In a study of a symptom checklist completed at intake and termination by students at a college counseling center, Weinstock and Meier (2003) found similar differences between intervention-sensitive and traditionally selected items.

Cognitive ability/intelligence tests

Definition. Tests designed to measure intelligence or other constructs related to cognitive ability.

Description. The ambiguous definition above reflects the longstanding uncertainty about what traditional intelligence tests measure. Sternberg (1984) summarized the beliefs of many when he wrote that although many psychologists "act as though 'intelligence is what intelligence tests measure' . . . few of us believe it" (p. 307). Measures of cognitive abilities can predict educational and occupational performance (e.g., Austin & Hanisch, 1990), but what these tests actually measure remains in some doubt.

On the basis of strong positive correlations among intelligence measures, Spearman introduced the idea of *g*, or a general factor of intelligence (Nichols, 1980). Subsequent work by Thurstone (1938) and Guilford (1967) led them to believe that more specific group factors accounted for the operations of cognitive abilities. Ascertaining the structure of cognitive abilities remains an important but elusive goal for those who desire to improve the measurement of cognitive abilities and skills.

Three related types of tests are (a) achievement tests, intended to measure students' current academic levels, (b) ability tests, broad tests of skills intended to estimate general intellectual ability, and (c) aptitude tests, tests designed to measure specific skills, independent of previous learning, in the hope of predicting future performance in that domain. Although often described as distinct types of cognitive ability tests, achievement and ability tests correlate very highly (Gregory, 1992). Ability or intelligence tests typically tap into verbal comprehension, reasoning, and perceptual organization (Gregory, 1992); aptitude tests focus on one of these specific areas.

Whereas many achievement and aptitude tests are administered in groups, ability tests such as the Wechsler tests of general intelligence—the Wechsler Adult Intelligence Scale—Revised (WAIS-R) and the Wechsler Intelligence Scale for Children III (WISC-III)—are individually administered. Both tests consist of a battery of tasks that produce verbal, performance, and full-scale IQ scores. Tasks include arithmetic, vocabulary, picture arrangement, object assembly, and comprehension. Groth-Marnat's (1990) review of the WAIS-R reported high reliability: split-half estimates for full scale equaled .97, for verbal, .97, and for performance, .93. Full-scale WAIS scores have been found to correlate highly with other intelligence measures such as the Stanford-Binet and Slosson Intelligence Test, as well as with years of education.

Cognitive ability tests typically produce a set of scores for each individual. The scores in Table G.1 summarize the Wechsler Adult Intelligence Scale—Revised (WAIS-R) performance of Peter, a 24-year-old male completing treatment for alcoholism (Hood & Johnson, 1991).

Table G.1. *WAIS-R Scores for Peter, a 24-Year-Old in Treatment for Alcoholism*

Verbal tests	Scaled score	Performance tests	Scaled score
Information	6	Picture Completion	8
Digit Span	10	Picture Arrangement	7
Vocabulary	7	Block Design	7
Arithmetic	12	Object Assembly	7
Comprehension	10	Digit Symbol	5
Similarities	10		
Total Verbal	55	Total Performance	34

	Sum of Scaled Scores	IQ
Verbal	55	95
Performance	34	76
Full Scale	89	85

Hood and Johnson (1991) noted that Peter's Verbal scores are in the normal range, while the Performance scores are below normal. They suggested that the Verbal –Performance difference is a result of continued alcohol abuse.

Construction

Definition. Procedures employed to create a test.

Description. The rules employed to create a test have serious implications for the interpretation of any scores produced by that test. Given its importance, it is surprising that little consensus has developed over the best procedures for test construction. In the following discussion I describe several sets of construction guidelines.

Gregory (1992) described five steps in test construction: (a) defining the test (e.g., purpose, content), (b) selecting a scaling method (i.e., rules by which numbers or categories are assigned to responses), (c) constructing the items (e.g., developing a table of specifications that

describes the test's content areas according to the methods by which the content is measured), (d) testing the items (i.e., administering the items and then conducting an item analysis), and (e) revising the test (e.g., cross-validating it with another sample because validity shrinkage almost always occurs). A researcher evaluating a new mathematics curriculum, for example, might (a) desire a test that could show changes over time in mathematics skills, (b) assign a score of 1 to each math item correctly scored, (c) create a table of specifications indicating what kinds of skills would be expected to be acquired, (d) run a study to determine which items were sensitive to change, and (e) repeat the process with the selected items with a new group of students.

Similarly, Burisch (1984) described three approaches to personality test construction representative of many domains:

1. External approaches that rely on criteria or empirical data to distinguish useful items. The content of the item is less important than its ability to meet a preestablished criterion. For example, Minnesota Multiphasic Personality Inventory (MMPI) items were chosen on the basis of their ability to distinguish between normal persons and those with a diagnosed psychopathology.

2. Inductive approaches that require the generation of a large pool of items, which are then completed by a large number of subjects, with the resulting data subjected to a statistical procedure (such as factor analysis) designed to reveal an underlying structure. Many aptitude tests, such as the General Aptitude Test Battery (GATB), were constructed in this fashion.

3. Deductive approaches that rely on a theory to generate items. Items that clearly convey the meaning of the trait to be measured and that measure specific (as opposed to global) traits are more likely to be useful. Items for the Myers – Briggs Type Indicator, for example, were originally derived from Jung's (1923) theory of types.

Burisch's (1984) review of the literature found no superiority for any of these approaches in producing reliable and valid scales. In fact, he suggests that it is more useful to simply ask individuals to rate themselves on a trait that they understand and for tasks in which they possess high motivation.

Educational and psychological researchers frequently wrestle with the question of whether they need to create a new scale for a study. In the psychological arena alone, however, estimates are that 20,000 new psychological, behavioral, and cognitive measures are developed each year (American Psychological Association, 1992). It is quite likely that a self-report scale, interview, or other operation has already been developed in your practice or research area. The question then becomes finding that operation. Most disciplines have books or databases that are good places to start. In education and psychology, for example, sources of information about published tests include *Tests in Print* (Buros Institute for Mental Measurements), *Mental Measurements Yearbook* (Buros Institute for Mental Measurements), *Tests* (Pro-Ed, Inc.), and *Test Critiques* (Pro-Ed, Inc.). Sources of unpublished tests include the *Directory of Unpublished Experimental Mental Measures* (Wm. C. Brown), *Measures for Psychological Assessment: A Guide to 3,000 Original Sources and Their Application* (Institute for Social Research, University of Michigan), and *Tests in Microfiche* (Educational Testing Service). Some of these tests have

been placed on computer; information about such applications can be found in *Psychware Sourcebook* (Pro-Ed, Inc.) and *Computer Use in Psychology: A Directory of Software* (American Psychological Association). In addition, there is a widely available database called Health and Psychosocial Instruments (HAPI), available through BRS Information Technologies, which lists more than 7,000 instruments.

Context

Definition. The setting or background in which an event or experience takes place.

Description. *Test context* refers to any circumstance or situation that test takers perceive as part of the testing process. Examples of test context include the characteristics of the test administrator; characteristics of the test takers; the specific wording of test instructions, items, and response format; and the testing method (e.g., self-report) itself. Although the idea that context affects test takers' behaviors has been recognized in the testing literature—Schwarz and Oyserman (2001), for example, suggested that responses to "self-reports are highly context dependent" (p. 128)—testing context has yet to receive the centrality it deserves.

If contextual cues are perceived similarly by a group of individuals taking a particular test, the common perception of those cues in that group is a *shared context*. For example, students applying for admission to college or graduate school may recognize the intent of transparent questions and reply with distorted information that favors their selection. Test developers implicitly depend on shared contexts when they create and administer tests that they intend to be valid measures of a construct. That is, test developers assume, but typically do not evaluate, that characteristics of the testing method, the testing situation, and the test takers all influence test scores in a manner that enhances or at least does not detract from test validity. Test developers, for example, assume that test takers understand items similarly and in the manner intended by the test developer (Walsh & Betz, 1985). Shared contexts become a source of invalidity, however, when such contexts function in a manner contrary to the test's intended purpose.

Criterion-referenced interpretations

Definition. Interpreting a test score in relation to a criterion or preestablished level instead of in relation to other persons.

Description. Suppose an individual received a score of 95% on a classroom test. What does that mean? In a norm-referenced interpretation, that would indicate that the student scored higher than 94% of the rest of the class. A criterion-referenced statement would be "correctly completed 95 of 100 questions." Criterion-referenced interpretations simply describe performance in relation to a standard other than other persons.

With criterion-referenced tests, items are retained during test development because of their relation to a criterion, regardless of the frequencies of correct or incorrect responses. However, criterion-referenced tests cost more than norm-referenced tests because they (a) require considerable effort in the analysis and definition of the performance criteria to be measured and (b) may necessitate special facilities and equipment beyond self-report materials. If one is

interested in predicting performance on a criterion—the major purpose of selection testing—then criterion-referenced approaches would seem a logical choice. If one is interested in knowing whether a person can shoot a basketball, it usually makes more sense to give that person 20 shots than to administer a test of eye – hand coordination.

In regard to item development of criterion-referenced tests, Swezey (1981) emphasized the importance of precisely specifying test objectives. Criteria can be described in terms of variables such as product or process, quality, quantity, time to complete, number of errors, precision, and rate (Gronlund, 1988; Swezey, 1981). A criterion may be a product such as "student correctly completes 10 mathematics problems"; a process criterion would be "student completes division problems in the proper sequence." Process measurement is useful when diagnostic information is required, when the product always follows from the process, and when product data are difficult to obtain.

Criterion-referenced tests should be reliable and valid to the extent that performances, testing conditions, and standards are precisely specified in relation to the criteria. For example, Swezey (1981) preferred "within 5 minutes" to "under normal time conditions" as a precise testing standard. In some respects, the criterion-referenced approach represents a move away from a search for general laws and toward a specification of the meaning of test scores in terms of important measurement facets. Discussing test validity, Wiley (1991) presented a similar theme when he wrote that the labeling of a test ought to be "sufficiently precise to allow the separation of components of invalidity from valid variations in performance" (p. 86). Swezey's and Wiley's statements indicate the field's increasing emphasis on construct explication.

Cross-situational consistency

Definition. Tendency of a person to behave consistently across situations or settings.

Description. If traits are the dominant psychological phenomena, individuals should behave consistently across situations. In contrast, *situational specificity* refers to the tendency of individuals to behave according to the specific situation in which they find themselves.

Traits are assumed to be stable across situations. Thus, persons described as honest are expected to display honest behavior regardless of the situations in which they find themselves. 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, 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). Magnusson and Endler (1977) discussed coherence, a type of consistency that results from the interaction between individuals' perception of a situation and individuals' disposition to react

consistently in such perceived situation. The factors that influence this interaction, such as intelligence, skills, learning history, interests, attitudes, needs, and values, may be quite stable within individuals.

For example, Lyytinen (1995) studied the effects of two different situations on children's pretend play. She placed 81 children ages 2 – 6 years in either a play-alone condition or with a same-gender, same-age peer. Children playing with the familiar peer displayed a significantly higher proportion of pretend play acts than when playing by themselves. Children playing with another child, however, displayed fewer play acts overall because of the time they spent looking at and talking about each other's play. Thus, situational specificity appears to be at work in the pretend play of children.

Factor analysis

Definition. A statistical method for understanding the number and type of constructs influencing a test's score.

Description. Factor analysis is a method for analysis of test data. Factor analysis has been such an important technique in the development of scoring procedures for tests that I discuss it here.

Test developers assume that any large number of items or tests reflect a smaller number of more basic factors or traits. These factors consist of a group of highly intercorrelated variables (Vogt, 1993). *Factor analysis* refers to a set of statistical procedures used to examine the relations among items or tests and produce an estimate of the smaller number of factors that account for those relations.

Two basic types of factor analysis are commonly employed: exploratory and confirmatory. In exploratory factor analysis, little or no knowledge is available about the number and type of factors underlying a set of data. Test developers employ exploratory factor analysis when evaluating a new set of items. With confirmatory factor analysis, knowledge of expected factors is available (e.g., from theory or a previous exploratory factor analysis) and used to compare factors found in a new data set. A good way to begin learning about factor analytic techniques and their output is through statistical users manuals as provided by companies like SPSSx and SAS.

Golden et al. (1984) maintained that test developers must understand the theory employed to select items in a factor analysis "since the resulting factors can only be interpreted accurately within the context of a theoretical base" (p. 27). Nevertheless, many, if not most, test developers base their item selection only loosely on theory. Gould (1981) similarly criticized the use of factor analysis in the creation of intelligence tests. Gould believed many social scientists have reified intelligence, treating it as a physical entity instead of as a construct. Gould maintained that "such a claim can never arise from the mathematics alone" (p. 250) and that no such evidence exists in the case of intelligence.

One decision that test developers must make during the course of a factor analysis is whether to rotate the factor loadings. If test developers desire their factors to be independent of

one another (i.e., orthogonal), the analysis includes a rotation (but see Pedhazur & Schmelkin, 1991, for a different perspective). Another issue is deciding how many factors should be extracted during an analysis. One approach is to examine the eigenvalues of the found factors; eigenvalues roughly correspond to the proportion of variance explained by summing the squared loadings on a factor. A general rule of thumb is that factors with eigenvalues of 1 or more should be considered useful.

For example, Blaha and Wallbrown (1996) conducted factor analyses on the WISC-III subtest intercorrelations. Subtests include arithmetic, vocabulary, picture completion, and mazes. Blaha and Wallbrown obtained two- and four-factor solutions for four age levels (6 – 7, 8 – 10, 11 – 13, and 14 – 16 years old). The two-factor results supported a general *g* factor (defined as an overlap among different assessments of intelligence) as well as two major group factors of verbal-numerical-educational ability and spatial-mechanical-practical ability. The four-factor solution suggested factors of perceptual organization, verbal comprehension, freedom from distractibility, and perceptual speed. Blaha and Wallbrown concluded that these results support the construct validity of the Full Scale IQ of the WISC-III as a measure of general intelligence.

Interpretation

Definition. Placing measurement data in a context, or making sense of test data.

Description. Test interpretation depends on all the steps that came before it. That is, the test construction process must have produced a valid test if the interpretation is to be valid; the test must have been administered and scored with a minimum of error during those processes. Because tests are never perfectly valid, interpretation should include statements about the limits of the test as influenced by demonstrated and likely sources of error. Without such statements of limitations, you may misinterpret the scores of the measurement methods you employ.

Test interpretation, particularly in educational settings, traditionally has focused on norms. In norm-referenced tests a test score is interpreted by comparing it to a group of scores. I can say, for example, that a third-grade student's score on an achievement test places him or her at the 90th percentile of performance. Norm-referenced interpretations are typically contrasted with criterion-referenced test interpretations (i.e., comparison to a standard, instead of other persons). That same third-grade student may have correctly answered 35 of 40 test items that assessed previously taught material; the teacher may have set a criterion of 30 correct answers for students to pass the course.

Other types of interpretations are also useful. With *formative* tests, interpretation focuses on an individual's performance on the components of an intervention. In a mathematics course, a formative test might provide information about the particular types of addition and subtraction problems a particular student answered correctly and incorrectly. During an intervention, formative tests provide feedback to the intervenor and participant that reveal progress and guide adjustment of the intervention. In education, Cross and Angelo (1988) described this process as a loop "from teaching technique to feedback on student learning to revision of the technique" (p. 2).

Summative tests provide an overall evaluation of an individual's performance in an intervention (e.g., a course grade). Summative tests provide data convenient for administrative decision making. Summative tests can suggest initial hypotheses relevant to interventions: for example, a standardized achievement test can describe a student's strengths and weakness (compared to other students) across subject areas, information that may be relevant to inclusion in or exclusion from an intervention (e.g., a remedial course or repeating a grade). More sensitive measures will be needed to develop and test those hypotheses, however, and it is here that formative tests can be useful (Bloom, Hastings, & Madaus, 1971; Cross & Angelo, 1988). The interpretation of summative tests focus on an aggregate score (of items and components), whereas administrators of formative tests tend to examine item response patterns (Bloom et al., 1971).

For example, much more attention has been paid in the literature to how the test administrator or researcher interprets test scores than to how test takers make sense of them. One exception to this is research on the Barnum effect. Gauging the accuracy of a particular test interpretation depends on making comparisons with other types of test interpretation. The Barnum effect occurs when individuals take a test and receive test interpretations not based on their test data, but simple generic statements that might apply to anyone, such as the statements that appear in horoscopes ("Work hard today and your efforts will pay off"). Test takers usually find such bogus feedback as accurate as real interpretations. Guastello and Rieke (1990) evaluated the accuracy of real computer-based test interpretations (CBTIs) based on 16PF scores (a personality inventory) with bogus reports. A sample of 54 college students rated the real reports as 76% accurate and the bogus reports as 71% accurate. Computer-based reports are likely to increase the Barnum effect because many people ascribe increased credibility to computer operations.

Interviews/ratings by others

Definition. Qualitative and quantitative assessments of a person or group by other persons along an educational or psychological dimension.

Description. If self-reports are subject to distortion, an obvious avenue to pursue is raters who have some experience in gathering information and who do not share the biases of test takers. Thus, the interview is the most commonly employed method other than self-report.

Interviews have been referred to as conversations with a purpose. Interviews can be categorized according to their degree of structure. *Structure* here refers to an interviewer's predetermination of such elements as the information to be obtained, order of questions, coding of questions and answers, and guidelines for probing responses. Research suggests that the addition of structure to interviews often improves their reliability and validity (e.g., Conway, Jako, & Goodman, 1995). In the realm of employment interviews, Wright et al. (1989) maintain that such structured interviews work well because they (a) are closely based on a job analysis of the employment position, thus reducing error resulting from information irrelevant to the specific job; (b) assess individuals' work intentions, which are often linked to work behavior; and (c) use the same set of questions and standards for scoring answers, thereby increasing reliability. Hoshmand (1994) summarized another set of guidelines for interviewers. She suggested, for example, that interviewers need to manage interviewees' anxiety so as to facilitate

communication. Open questions that require elaboration (e.g., "Tell me more about that experience") produce better information than closed questions that produce one- or two-word answers (e.g., "Were you satisfied with that job?").

Structured interviews begin with a set of items or questions that the interviewer poses to the participant. For example, Hood and Johnson (1991) described the SAD PERSONS scale, developed by Patterson, Dohn, Bird, and Patterson (1983) to assess suicide risk. With relevant training, researchers interested in suicide could assess risk using interview questions about the following:

S ex (Males more likely to commit suicide)
A ge (Persons under 25 or over 45 more likely)
D epression

P revious attempts
E thanol abuse
R ational thinking loss
S ocial support loss
O rganized plan
N o spouse
S ickness

One risk point is awarded for each of these 10 risk factors. Particularly with factors such as depression and rational thinking loss, interviewers would probe beyond an initial question before making a yes/no judgment.

Benes (1995) reviewed the Social Skills Rating System (SSRS), a standardized instrument whereby teachers, parents, and students can rate children's social behaviors. The SSRS is designed to provide screening and classification of students' behavior in educational and family settings. The SSRS Parent Form, completed by the mother and/or father, provides four social skills subscale scores (cooperation, assertiveness, responsibility, and self-control) and two problem behavior scores (externalizing and internalizing). Parents rate such items as "Attempts household tasks before asking for help" on a 3-point scale (never, sometimes, very often) and rate the importance of the behavior (not important, important, critical). Research with the SSRS found that coefficient alphas for scale scores were generally in the .80s and that the SSRS correlated highly with other social behavior assessments.

Item analysis

Definition. Methods for evaluating the usefulness of test items.

Description. Typically test developers perform item analysis during test construction to determine which items should be retained or dropped. Although *items* usually refers to questions or statements, here I use *items* to mean any distinct measurement measure, including an observation or behavioral performance.

A story about how Thomas Edison invented the light bulb is illustrative of the item analysis and test construction process. Edison reportedly sorted through thousands of types of materials in the search for a filament that could conduct electricity, emit light and minimize heat, and endure for a long period of time. Similarly, test developers typically sort through dozens or hundreds of items in an attempt to find a number that exhibit the characteristics desired for a particular test.

Guidelines for item selection have been proposed by numerous authors (e.g., Burisch, 1984; Dawis, 1987; Epstein, 1979; Gronlund, 1988; Jackson, 1970). For example, Jackson (1970) proposed four general criteria, suggesting that scales (a) be grounded in theory, (b) suppress response style variance, (c) demonstrate reliability, homogeneity, and generalizability, and (d) demonstrate convergent and discriminant validity. Criterion (a) can be evaluated by noting the degree to which the initial item pool was rationally constructed. The degree of response style or response set variance (b) could be assessed by correlating items with a measure of social desirability. Criterion (c) can be assessed by examining item – total correlations and by checking for ceiling and floor effects (i.e., participants' responses to an item cluster near the top or bottom of the possible range of scores). Correlations among scale items and related and different constructs can be computed to assess validity (d).

For example, Musser and Malkus (1994) employed an item analysis to develop the Children's Attitudes Toward the Environment Scale (CATES), a measure designed to assess children's knowledge about the natural environment. They administered a pool of 90 items to 232 fourth- and fifth-grade students and subjected those items to analyses that evaluated their internal consistency (seeking items with high item – total correlation), mean level (with items showing ceiling or floor effects dropped), and variability (with items showing low variability dropped). The 25 selected items were then administered to a new sample of 90 third-, fourth-, and fifth-grade students; these items together displayed a coefficient alpha of .70. Finally, the 25 items were administered twice, from 4 to 8 weeks apart, to 171 third, fourth, and fifth grade students. Test – retest reliability was calculated at .68; coefficient alpha for the two administrations was .80 and .85. These repeated waves of item administration, analysis, and item selection typify most item analyses. Also notice that the analyses Musser and Malkus employed, although standard, are best used to select items that measure stable constructs. The resulting items are likely to be less useful for studying constructs that change.

Measurement

Definition. The process of assigning numbers or categories to phenomena according to agreed-upon rules.

Description. Measurement is a more specific term than *test* and begins to move us toward a discussion of what constitutes a better or worse test. Krantz et al. (1971) defined measurement as assigning numbers to objects "in such a way the properties of the attributes are faithfully represented as numerical properties" (p. 1). In other words, data that result from the measurement process should reflect the characteristics present in the phenomenon we are interested in measuring.

A key idea here is that tests, assessments, and measurements measure constructs, which are abstract summaries of natural regularities indicated by observable events. Construct explication is the process by which constructs are connected to observable events (Torgerson, 1958). Construct explication is important because most social science constructs usually cannot be sufficiently defined through a single operation. However, many researchers and clinicians behave as if their choice of method for measuring a construct is unimportant. Researchers and practitioners who default to traditional measurement methods are ignoring what I call "the explication hypothesis." With any construct, there exists three questions related to explication:

1. Is the construct useful enough to measure? If I have some reason to believe that the construct has potential or demonstrated value, then I should ask:

2. Can any existing method measure the construct to the extent necessary for our purpose? It is possible that the construct is useful, but I have no available method of adequately measuring it. If such methods are available, the next question is:

3. Which of the available methods best measure the construct?

Unfortunately, many professionals ignore these admittedly difficult judgments and default to traditional methods. For example, researchers in personality psychology typically resort to self-reports, and many qualitative researchers assume that interviews are their only viable method.

For example, Henslin (1993) summarized Merton's (1956, 1968) strain theory, designed to explain individuals' reactions when they are socialized to desire cultural goals (e.g., material goods) while systematically prevented from reaching those goals (e.g., because of racism or sexism). Merton suggested that individuals will have one of five reactions: (a) conformity, continuing to use legitimate means to attain the goals, (b) innovation, devising illegal means, (c) ritualism, giving up on the goals but continuing to conform, (d) retreatism, rejecting the goals and the standard means, and (e) rebellion, rejecting the goals and the means, and attempting to replace both with new goals and means. Given these constructs, how might I measure them? For example, how might I measure the distance between desired cultural goals and achieved goals? I might expect measures of constructs such as socioeconomic status, ethnicity, and gender to be related to this distance.

Measurement error

Definition. Phenomena that affect scores on tests that are not intended to be reflected in those scores.

Description. In classical test theory, test scores are a combination of *true scores* and *error*. This traditionally has been represented by the following formula:

$$Y = X +/- e$$

where Y is the score that reflects the test taker's true score on the phenomenon, X is the score the test taker actually receives on a test, and e is error.

The term *bias* is sometimes used to refer to systematic errors associated with membership in a group. For example, socioeconomic status or ethnicity of test takers may interact with test items to over- or underestimate their true performance on the items (cf. Helms, 1992). One of the central controversies with intelligence tests, for example, is whether intelligence tests underestimate the ability levels of persons of color. Such bias can be checked, however, by assessing whether:

1. The content of the test is more familiar to certain groups than others. First, select test takers from different groups who have similar total scores on a test. Next, determine whether any individual items are passed or failed by different proportions of individuals in each group. If so, that item is biased.

2. The test does a better or worse job of predicting a criterion for different groups. The relation between the test and the predictor can be expressed with a regression line. If the slope of the regression lines per group differ, then bias is present. In this case scores on the test do not indicate equal performance on the criterion.

Stone et al. (1990) noted that test researchers rarely study the ability of test takers to understand test instructions, item content, or response alternatives. If test takers cannot adequately read and understand such content, they may respond to tests in unintended ways—that is, error is introduced. Stone et al. proposed that if respondents lack the cognitive ability to read and interpret questionnaires, their motivation and ability to complete a questionnaire will be impaired, and that such effects could be detected by comparing the psychometric properties of questionnaires completed by groups with different levels of cognitive ability.

Stone et al. (1990) used the Wonderlic Personnel Test to classify 347 Army Reserve members into low-, medium-, and high-cognitive-ability groups. Subjects also completed an additional 203 items in a test battery of 27 measures that included the Job Diagnostic Survey, which measures such constructs as job satisfaction and organizational commitment. Stone et al. found significant differences in coefficient alpha for 14 of the 27 constructs. In 12 of those cases, alpha rankings were as predicted: Scales' reliability estimates matched low- to high-cognitive-ability groups. Stone et al. also found a significant negative correlation (r = -.23) between cognitive ability and the number of missing questionnaire responses; that is, persons with lower cognitive ability left more items unanswered. Thus, it appears that respondents' cognitive ability can introduce error with some tests.

Method variance

Definition. Refers to the observation that the variability of a group of educational or psychological test scores results, at least in part, from the method employed to collect those data.

Description. An *operation* is a specific, single activity designed for measurement. In contrast, *method* refers to a group of similar measurement operations. For example, you might have two operations (e.g., the State Anxiety Inventory and the Trait Anxiety Inventory) that share a single method (i.e., self-report).

Resource problems frequently create *mono-method and mono-operations biases*. Researchers, for example, frequently find themselves in a situation where they must conduct a study as quickly and efficiently as possible, and consequently use only self-report measures or interviews. Mono-operation bias refers to the collection of data through a single operation, and mono-method bias occurs when only a single method is used. Mono-method and mono-operation biases result from the fact that how data are collected—the method—strongly influences the data themselves (Campbell & Fiske, 1959). For example, you might avoid a mono-operation bias by using two separate self-report instruments. You would still, however, have a mono-method problem, because you employed only a single method, self-report. Employing multiple operations and multiple methods, in general, increases the chance that the resulting data will reflect the constructs of interest more than the measurement methods.

It is an axiom of measurement that no single operation totally reflects any single construct. Contemporary researchers generally embrace a philosophy of *multiple operationalism* (i.e., the use of multiple measures or methods to measure any construct; Cook & Campbell, 1979), yet this approach creates problems of its own. Which tests, for example, should be employed? In general, test users tend to employ operations that require the least resources (e.g., self-reports). When operations are measured via different methods, the methods themselves will influence scores. For example, observation and self-report of any single construct will yield at least somewhat divergent scores. Which one is more valid? The default solution in many cases is to aggregate across operations and methods, hoping that the scores on the construct of interest will aggregate while the influence of irrelevant factors (such as method) will be balanced or cancelled.

Meier (1988b) presented 31 college undergraduates with a self-report alcohol attitudes scale before and after they viewed a computer-assisted instruction (CAI) program on alcohol education. Statistical analysis of differences between pre- and posttest scores found a significant difference, indicating that students reported more responsible attitudes toward alcohol after the intervention. This study exemplifies both types of biases: (a) mono-operations bias is present because only one measurement device was employed to detect changes resulting from the intervention, and (b) mono-method bias is evidenced by the use of self-report only. Any study with a single measurement device displays both mono-operations and mono-method biases. More typical in the literature are studies that employ multiple measurement devices (thus avoiding mono-operation biases) but only one method, such as self-report (i.e., mono-method bias).

Norms

Definition. Data about a distribution of scores for a particular test.

Description. In norm-referenced interpretations the purpose of testing is to compare scores among individuals. Thus, the test is intended to detect individual differences. Gronlund (1988) indicated that developers of norm-referenced tests seek items with the greatest possible variability. With achievement tests, these items are pursued through a selection process that retains items of average difficulty. Easy and difficult items, which everyone passes or fails (cf.

Collins, 1991), are likely to be discarded. Aggregation of items with greater variability increases the possibility of making valid distinctions among individuals.

Norm-referenced testing has been the predominant approach in selection testing (Murphy & Davidshofer, 1994). Besides having a lower cost, norm-referenced tests also seem more applicable when the test administrator desires to select some portion of a group (e.g., the top 10% of applicants) as compared to all applicants who could successfully perform a function. Thus, norm-referenced tests are useful in selection situations where individuals are chosen partially on the basis of scarce resources. Suppose you conduct a research study and find that 95% of all graduate students who score 600 or above on the GRE Verbal scale are able to pass all required graduate school courses. From the perspective of criterion-referenced testing, everyone scoring 600 or above should be admitted. In many graduate departments, however, that would mean admitting more students than can be accommodated by the available courses, instructors, or financial supports. Such a situation certainly occurs in other educational, occupational, and clinical settings with fixed quotas. Norm-referenced testing, then, provides a solution: identify the top-scoring number who match the available resources.

If a test is intended to function as a selection device, its items should be developed with a sample representative of the population for whom the test is intended. Thus, the selection of a norm group for test development has serious consequences for the interpretation of future scores compared to the norm group. Much controversy has occurred over the widespread use of intelligence tests or vocational interest inventories, for example, that were developed and normed on predominantly white, middle-class persons.

Observational strategies

Definition. Assessment methods involving the direct observation of behavior.

Description. Behavioral assessment is a major type of observational strategy in psychology and counseling. In most cases *behavioral assessment* refers to the practice of employing a trained rater to observe another person's (usually someone completing behavior therapy) overt behaviors.

Although developed for use in inpatient mental health settings, the approach described by Paul and colleagues generalizes to a wide range of counseling and research purposes. Paul, Mariotto, and Redfield (1986b) suggested that the units of observation be established before the observation period so that observers are able to focus on important elements. Such units should be discrete samples of behavior, as opposed to global signs, insomuch as greater amounts of interpretation by observers are more likely to reflect characteristics of the observer. In a clinical setting, examples of discrete (and inappropriate) behavior include talking to oneself and hitting another person.

Error arising from such factors as carelessness or fatigue of the rater will be minimized when measurement data can be aggregated from multiple occasions. Paul et al. (1986b) concluded that the accuracy and relevance of observations can be maximized using multiple,

discrete, and scheduled observations made by trained observers as soon as possible following a behavioral event.

Paul et al. (1986b) described their chief assessment tools as Direct Observational Coding (DOC) procedures. DOCs require explicit sampling of individuals and occasions by trained observers. Paul et al. (1986b) noted two important sources of error that should be monitored with observers: (a) decay, random changes in the observer's reliability or consistency of observation, and (b) drift, systematic changes in the definition or interpretation of coding categories. A rater evidencing decay might pay close attention to observing initially, then tire over the course of several hours. A drifting rater might forget the initial rules for what constitutes "shouting," for example, and begin to count in that category any time a client simply raises his or her voice. Paul et al. (1986b) maintained that such errors could be minimized by obtaining converging data from different assessment procedures, conditions, and operations. Such observer biases have been linked to fatigue, knowledge of hypotheses, and observer's expectancies (Hoshmand, 1994).

Licht et al. (1986) reported that such DOC systems have been implemented with more than 600 clinical staff members in 36 different treatment programs in 17 different mental health institutions. The resulting flood of data has produced results of interest to researchers as well as to clinicians and administrators in the studied agencies. Data from DOC systems have produced evidence of substantial differences in the behavior of different clinical staff members and treatment programs. For example, staff – client interactions in 30 studied agencies ranged from 43 to 459 interactions per hour; over a full week, staff members were responsible for as few as 4 clients or as many as 33. Licht et al. (1986) found that how staff members interact with clients—that is, specific intervention programs—was highly correlated with client functioning and improvement (*r*s ranged from .5 to .9). In addition, the quality of staff – client interaction was more important than the quantity of that interaction. Licht et al. (1986) noted that DOC information may not only aid in the monitoring of treatment implementation but may be employed as feedback to adapt treatment for improved effectiveness.

Outcomes

Definition. The effects of a psychosocial intervention.

Description. Individuals seeking psychosocial interventions typically describe one or more identified problems or *target complaints*, a set of problems that becomes the initial focus of efforts at psychotherapeutic change. Researchers and clinicians typically start with these problems when trying to assess the outcomes of psychotherapy. That is, they assume that the client's presenting problems should be the focus of assessment at a later point for the purpose of evaluating whether change occurred. For a variety of reasons, however, during the course of counseling clients may alter the problem(s) that they wish to address. This is the issue of *persistent relevance*: Do the key problems reported at the beginning of therapy remain the chief issues throughout the course of therapy?

Clinicians and researchers typically assume that client outcomes have causes or processes that influence these outcomes. Creation of the link between process and outcome with a particular client is called a *case conceptualization* (Meier, 2003). This conceptualization then

becomes the basis for tailoring a particular treatment plan for that client. An alternative approach, employing *empirically supported treatments* (ESTs), indicates that once a desired outcome for a client is identified, one or more interventions should be employed that have been demonstrated to be effective in randomized clinical trials with that particular problem.

Conceptualizing outcomes with any particular client can be difficult when intermediate outcomes are necessary before a longer-term change is possible. For example, a student with failing grades may need her family to participate in family therapy (to stabilize the family's environment) before the student can turn her attention consistently to studying for school. In addition, for clients evidencing *treatment failure*, it may be useful to provide regular feedback to the therapist and client so that the therapy can be adjusted or changed for improved outcomes (Gray & Lambert, 2001).

Personality and interest tests

Definition. Tests assessing individual differences in personal and vocational traits.

Description. Most personality and interest tests are self-reports. Historically, developers of personality tests believed that personality, like intelligence, was consistent across persons and independent of situations (Danziger, 1990). On the basis of 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 (a) stability, declines across longer measurement intervals, (b) is lower in children, and (c) 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 in scores, leaving considerable room for environmental and person – environment influences.

Swanson and Hansen (1988; see also Campbell, 1971) found similar results with the stability of vocational interests: Although individual variability and environmental influences existed, trait stability could be demonstrated over time. Funder and Colvin's (1991) laboratory study with 140 undergraduates found behavioral consistency across laboratory and real-life settings, although consistency varied by type of behavior. Staw and Ross (1985) found that job satisfaction remained stable in a sample of 5,000 middle-aged men even when they changed jobs and occupations.

Interests are generally considered distinct from such constructs as ability and aspirations. Vocational interest tests ask individuals to report their likes and dislikes among various activities (e.g., working outdoors, working with people, doing clerical tasks). Developers of interest tests must create extensive norms of interests for persons in a wide variety of occupations. An individual's interests are then matched to these groups, with the assumption being that the field of closest match is likely to hold the greatest job satisfaction for the test taker (Gregory, 1992). Examples of current interest inventories include the Strong Interest Inventory, the Kuder Occupational Interest Survey, and the Self-Directed Search.

Gregory (1992) reviewed research evaluating the Strong Interest Inventory (SII). Test – retest reliability for 1- and 2-week periods exceeds .90, but drops into the .60s and .70s when the retest interval exceeds a year for respondents under 25 years of age. Gregory noted that the SII has proven useful in predicting which occupations individuals do and do not enter. Similarly, Cronin (1995) investigated the relations between the Sensation-Seeking Scale (SSS) and the SII with 55 undergraduate women. He found that women who scored high on the SSS also scored high on the SII's Adventure Basic Interest subscale. Cronin suggested that such women may become bored with traditional occupations and may seek out nontraditional choices.

Physiological measures

Definition. Tests designed to measure biological states.

Description. The intent of most of this type of testing is to link the physiological state with an educational or psychological measure. For example, occupational stress might be correlated with constructs such as heart rate or blood pressure.

Expecting broad classes of psychological and physiological phenomena to correlate, however, may represent a contemporary extension of the mistake committed by early psychologists. They expected to find relations between many different types of physical tasks, physiological activities, and intelligence, but discovered that such correlations were largely absent. More than 100 years after early psychologists began the task, Cacioppo and Tassinary (1990) found that attempts to link physiological states to psychological operations remain problematic because of confusion about the relations among the categories of events measured. They proposed that such relations be conceptualized as:

1. Outcomes, where many physiological events vary as a function of a single psychological operation with certain individuals or situations;

2. Markers, where a single physiological event varies with a single psychological operation with certain individuals or situations;

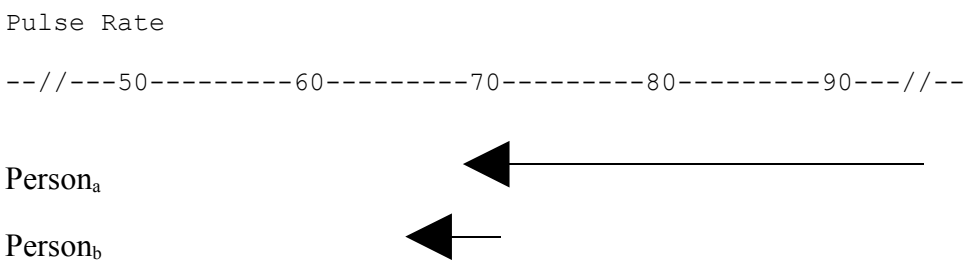
3. Concomitants, where many physiological events vary with a single psychological operation across a broad range of situations and individuals;

4. Invariants, where a single physiological event varies with a single psychological operation across a broad range of situations and individuals.

A practical contribution from physiological studies to the theory of measurement and assessment is the law of initial values (LIV; Wilder, 1957, 1967). The LIV indicates that physiological responses to stimuli depend on the prestimulus value of the physiological system. As shown in figure G.1, the higher the initial level, the smaller will be the response to stimuli that increase responding and the larger the response to stimuli that decrease responding. A person with a high pulse rate, as shown in Figure G.1, should likely evidence a greater change to a relaxing stimulus than a person with a moderate pulse rate.

Figure G.1.

Law of initial values.



Gould (1981) noted that attempts to correlate intelligence to physiological structures have been largely unsuccessful. Matarazzo (1992) predicted that intelligence testing would become increasingly linked with measures assessing brain activity. He reviewed studies that found moderate to high correlations between brain activity and intelligence scores. An implication of studies such as those cited by Matarazzo (1992) is that physiological measures represent the most valid measurement of the constructs in question. Yet research has shown that physiology can be altered by both medication and behavior (Schwartz, Stoessel, Baxter, Martin, & Phelps, 1996). Goldstein and Hersen (1990) indicated that efforts to identify biological markers of most forms of psychopathology have been unsuccessful, and Babor et al. (1990) reported on a similar lack of success in identifying biochemical markers of alcoholism.

Precision

Definition. The ability to detect small differences in a phenomenon, or the ability of a test to produce data closely reflecting the natural ordering and range of a phenomenon.

Description. A number of terms and definitions similar to *precision* have been offered. For example, Boyce, Meadow and Kraft (1994) defined (a) *resolution* as the finest interval of an instrument's measurement scale that can be distinguished by an observer (e.g., degrees on a thermometer); (b) *accuracy* as comparing values from a measurement process with measurement from other processes (e.g., comparing newly made thermometers with one known to be valid); and (c) *calibration* as checking an instrument against a known standard (i.e., the process of making a particular instrument accurate).

Stevens (1951) described the concept of *scale types*. *Nominal* scales are those that contain qualitative categories (e.g., red, blue, or green) but do not have information about differences in amount. *Ordinal* scales describe objects that differ from each other by some amount and that may be ranked in terms of that amount. *Interval* scales describe objects whose differences are marked by equal intervals, and *ratio* scales are interval scales that possess a zero point. Ordinal scales provide more precise information than nominal, interval more than ordinal, and so on. *Precision* thus refers to the ability of a measurement device to produce data that reflect the ordering, range, and distinctions of a phenomenon at a level sufficient for a particular purpose (cf. Nay, 1979).

The validity of a test depends on naming it correctly and its possessing adequate precision for its intended purpose. The naming aspect refers to the extent to which the test developer and test user understand the multiple constructs (i.e., validities and invalidities) that influence test scores. For example, I might wish a mathematics test score to reflect addition and subtraction abilities rather than reading ability. And to the extent that scores reflect the distinctions of the particular construct I wish to measure—that is, a test's precision—they reflect the phenomenon I wish to measure. For many purposes (e.g., grading) I would like that mathematics test to reflect the full range of ability levels rather than a simple high or low classification.

Projective devices

Definition. Measurement procedures that present respondents with ambiguous material for the purpose of producing information about unconscious processes and structures.

Description. Published in 1921, the most well-known projective device, the Rorschach, was developed to assist in differentiating between normal and clinical groups (Groth-Marnat, 1990). The Rorschach consists of 10 cards with symmetrical inkblots. The examiner hands a card to the subject and asks, "What might this be?" The examiner continues through all 10 cards and records the free association the respondent makes with each card. After this initial sequence the examiner again goes through each card, asking the respondent to indicate the material on the card that stimulated the particular responses.

Although the Rorschach is the predominant projective technique, mainstream testers commonly hold it in disrepute. Even Rorschach advocates sometimes attempt to deflect criticism by referring to the Rorschach as a technique and not a test (cf. Aronow & Moreland, 1995). The Rorschach has enjoyed a resurgence, however, as a result of efforts by Exner (1978, 1986) and colleagues to establish more standardized procedures for administering and scoring the instrument. Exner (1986) provided an excellent summation of the key projective assumption of the test:

It is important to remember that Rorschach answers are, in microcosm, a unique and valuable behavioral sample reflecting the way the individual is most likely to respond in a problem solving situation where there are few rules or principles directing the “psychological traffic.” In the Rorschach, the individual is “on his own,” forced to use the behaviors with which he is most comfortable, which are easiest for him to display, and which, in his judgment, will lead to acceptable performance. One of the most important features of the Rorschach is that it is “nondirected” and does force the individual to display his “psychological wares” in coping with the situation. When the Comprehensive System was developed, one point became clear above all others: the importance of keeping the task as free as possible from externally induced direction. (p. 59)

Parker, Hanson, and Hunsley (1988) conducted a meta-analysis to compare the published psychometric properties of another well-known personality test, the Minnesota Multiphasic Personality Inventory (MMPI), with the Rorschach. They collected data about test reliability (including internal consistency and rater agreement estimates), stability (test – retest), and convergent validity (correlations with relevant criteria). Interestingly, Parker et al. found an

insufficient number of discriminant validity reports to be able to report a comparison of the two instruments in this category. Parker et al. (1988) combined test subscales to produce the following psychometric estimates: (a) for reliability, an overall *r* of .84 for the MMPI and .86 for the Rorschach; (b) for stability, an overall *r* of .74 for the MMPI and .85 for the Rorschach; and (c) for convergent validity, an overall *r* of .46 for the MMPI and .41 for the Rorschach. Parker et al. concluded that despite the MMPI's reputation as the superior instrument, the MMPI and the Rorschach appear to possess comparable psychometric values.

Ornberg and Zalewski (1994) reviewed 48 studies examining the usefulness of the Rorschach with adolescents. They found evidence that the Rorschach could provide valid information about such constructs as depression, psychological distress, reality testing, and psychotic thinking. However, Ornberg and Zalewski noted that many of the Rorschach studies were limited by small sample sizes and highly variable scoring systems.

Psychophysics

Definition. Study of individuals' perceptions of sensory stimuli.

Description. Gracely and Naliboff (1996) described the two central measurement tasks in psychophysics as sensory detection (in which a judgment is made about whether a stimulus is present, such as the presence of a light or a tone), pain thresholds (where a sensation is always present), or some combination of the two. Thus, psychophysical judgments involve a response criterion that results in the detection of a stimulus or the labeling of a stimulus as painful (Gracely & Naliboff, 1996). In comparison to the more complex rating tasks seen with self-reports and ratings by others with psychological constructs, these psychophysical judgments present a simpler phenomenon for study of measurement issues.

In the *Method of Constant Stimuli*, an experimenter randomly presents a range of stimulus intensities around the expected detection threshold; the method can be employed to detect sensations as well as when sensations become painful. The threshold is defined as the point where the stimulus is detected 50% of the time. Gracely and Naliboff (1996) reported that research with this method, however, indicates that the "transition between no sensation, nonpainful sensation and pain sensation . . . is not distinct and vary over trials" (p. 244). In the *Staircase Threshold Method* for determining pain threshold, a series of stimuli are presented and the subject describes each as not painful or as producing mild pain. If a stimulus is perceived as not painful, its intensity is increased; if painful, then intensity is decreased.

Purpose

Definition. The intended use of a test, measurement, or assessment.

Description. Tests are employed for many purposes, but most of these can be classified under one of three headings: theory building, selection, or detecting change. Tests designed for theory building are intended to provide information to test, evaluate, and modify the hypotheses and models derived from a theory. Historically, selection tests are the dominant use: Test data are employed to make a decision about whether or not the test taker is selected for a job, school,

service in the armed forces, or other position. Tests designed to detect change typically attempt to find effects resulting from interventions of some type or from developmental processes.

The key issue is this: Problems may arise when tests are employed for purposes for which they were not explicitly intended. For example, selection tests are constructed with items designed to measure presumably stable individual traits (e.g., intelligence). Many researchers and practitioners, however, then employ these tests in an attempt to gauge the effects of interventions and developmental processes. Scores on standardized achievement tests employed in schools, for example, may partially reflect such constructs as socioeconomic status and general cognitive ability. However, they are less likely to show the effects of what is learned in the classroom than mastery or criterion-referenced tests specially created for evaluating the effects of classroom learning.

An analogy is the case of a meteorologist who wants to study the effect of temperature on plant growth but uses a barometer to measure temperature. Now, measurements using a barometer for some periods might actually correlate roughly with temperature; during the summer, high barometric pressure is more likely to be associated with warmer temperatures. Consequently, the meteorologist might even find some weak relation between barometric pressure and plant growth. That relation, however, will be weaker than the one found with an instrument whose primary purpose is to measure temperature, the thermometer.

Qualitative assessment

Definition. Method employed to collect nonnumerical data such as text and speech.

Description. Qualitative assessors typically observe individuals or present them with open-ended queries designed to elicit samples of the phenomenon in question. This material is subjected to a coding scheme designed to organize it conceptually. For example, a graduate student observed group counseling sessions of students who were in the process of adapting to a new school. Table G.2 summarizes those qualitative data.

Table G.2. *Example of Qualitative Notes*

Qualitative Progress Notes—Middle School Group

Session 1

Theme: Getting acquainted.

Leader described the group as sixth graders who are new to the school. Leader gave the rules.

All played Introduction game—pressure to remember each person's name and former school/location. All succeeded.

All played To Tell the Truth—spotlights each person and encourages dialogue to solve which one of four statements is false.

Session ended before end of game.

Session 2

Theme: Structured dialogue leads to cohesion.

Presence of new person led to reintroductions by group.

Counselor-as-group member role was lessened.

To Tell the Truth game continued:

- •Focus on all the rest of group (including new person).
- •Students wanted to play again.
- •All asked questions; only the students who chose to be "it" made true – false statements.

Session 3

Theme: Cohesion in spite of dominant member of group.

Girl brought family and vacation pictures; interest was high. Too many pictures and excessive descriptions led to waned interest of rest of group.

Counselor steered to group activity, "Compared to my last school, something I find different at GMS is" and "something I find the same at GMS is":

- •Group members concentrated on school and personnel characteristics at former/present locations.
- •More focus on differences than similarities.
- •All participated; one quiet boy remained on the edge of discussion but was drawn in by leaders.

Session 4

Theme: Increased group intimacy.

Passes to group had not been distributed; group members reminded counselor it was group day. Susan, quiet girl, was one who reminded.

Sudden death of the father of another sixth-grade girl led to discussion of death and loss. Group members spoke with more candor and intimacy:

- •Stuart is adopted, doesn't know birth parents.
- •Dahlia tried to dominate, but others did not allow it. She seems very emotionally needy.

a. She had a friend who was beaten to death by relative's live-in boyfriend.

b. She can't understand why her father is fighting giving child support.

All participated discussing: wakes, funerals, heaven, what their reactions would be to lose someone close.

Session 5

Theme: Cohesion threatened by intragroup cliquing.

Clique—led by Stuart—tried to get the leader to join. Initial behavior was giddy.

Structured conversation topic led to involvement of all group members.

Session 6

Theme: Increased group intimacy.

Enthusiasm was very high as group members talked about events since last session:

•Olga told us about throwing a rock and hitting a gull (witnessed by Jose).

•Jane told us she got new skis.

•Dahlia said her grandfather came home from a business trip.

Dahlia said her home room had won the Drug Free Door Contest. This led to heavy discussion about other doors.

Jane made curt comment that Jose never says anything.

Leaders introduced a proposed group video project:

•Group members talked excitedly.

•Jill proposed an accompanying book.

•Dahlia wants to star in the video.

•Jose spoke up without coaxing.

Students took leaders on a tour of the school to see the doors.

As shown in Table G.3, these notes were then examined by session to abstract the important topics and themes for the group as a whole, individuals in the group, and the group leaders.

Table G.3.

Qualitative Analysis

Session No.	Level of analysis for themes: Group	Individual	Leader
1	Starting the group; group game starts.		Leaders provide rules.
2	Cohesive; reintroductions; game continues.		Leader structure aids cohesion.
3		Dahlia is the star.	Structure to limit Dahlia.
4	Increased intimacy while discussing death and loss issues.	Dahlia is the star; members remind leaders about needed passes.	
5		Three members Subgroup.	Structure to decrease subgrouping.
6	Cohesive; increased intimacy.	One member attacks another; Dahlia wants to star in group's video.	Introduce video project.

Several trends are apparent from this analysis. First, the group shows cohesiveness relatively quickly, perhaps because of its members homogeneity (i.e., they share the same problem of adapting to a new school) and the leaders' initial structure. In Session 4, they feel comfortable enough with each other to discuss their feelings about death and loss. At the individual level, some members attempt to break away from the main group (Session 5), and one member, Dahlia, frequently attempts to monopolize the group. The leaders consistently respond to such difficulties by attempting to increase the structure of the group (e.g., by providing group tasks).

As in survey research, many qualitative assessors ignore issues around reliability and validity. Should qualitative assessments be considered a measurement method? I would suggest that the answer depends on the purpose of the assessment. If your goal is simply to explore a phenomenon, the answer is no. But once the purpose turns to testing hypotheses or making applied decisions, qualitative assessments can be considered to be measuring constructs and thus subjected to reliability and validity questions.

As examples of qualitative assessments that could be of use to counselors, Goldman (1992) described the Life Line and the Vocational Card Sort (VCS). For the Life Line, clients first draw a line vertically down a sheet of paper and then begin to list important life events along the line chronologically. The VCS consists of a set of cards containing occupational names, which the client sorts in two stages. First, the cards are sorted into three piles: occupations the person would consider, those he or she would not consider, and those about which the person has doubts. Next, the client takes the piles, beginning with the "No" category, and sorts them into smaller piles containing similar reasons the person would not consider them. This exploration process enables counselor and client to get an in-depth sense of the factors important to the client's career decision making.

Random and systematic errors

Definition. Random errors are irrelevant effects that influence measurement unpredictably, and systematic errors display some pattern or order.

Description. All measurements are presumed to be influenced by error sources, both random and systematic. Random errors reflect sources that are unrepeatable or haphazard (Abelson, 1995). In contrast, systematic errors can be identified and investigated. Table G.4 displays a partial list of systematic error sources that have been studied in the educational and psychological literature.

Table G.4. *A Partial List of Measurement and Assessment Error Sources*

Thorndike (1949) (in Murphy & Davidshofer, 1988)	Paul (1986)	Nelson (1977a)
Test-taking skills	Carelessness	Motivation
Ability to comprehend instructions and items	Fatigue	Valence
	Boredom	Instructions
	Information overload	Type of behavior
Response sets		Timing
Health	Emotional strain	Schedule of self monitoring
Fatigue	Attention shifts	
Motivation	Equipment failure	Type of recording device
Stress	Variations in lighting and temperature	Number of behaviors concurrently monitored
Set for a particular test		
Examiner characteristics	External distractions	

One way to consider systematic error in measurement is to think of instances when measurement method and participant mismatch, that is, interact in an undesired fashion. Such mismatches can be characterized in terms of cognitive, affective, and behavioral categories.

Cognitive mismatches occur when there are differences between the test language and cultural assumptions and the test taker, the test taker has no experience in the content area, or the test taker lacks sufficient cognitive skills (e.g., reading ability), memory skills, or education to be able to understand and complete the test. For example, an interviewer may read complex questions in English to a person who is a nonfluent speaker of English. When such mismatches occur, for example, the test taker may respond randomly to items; it may be worth repeating a subset of items to detect such responding. Such problems may be prevented by pilot testing methods with a small subset of persons and by rewriting items and tasks to enhance clarity and understanding.

Affective mismatches occur when test takers become fatigued or bored during the testing process, have strong concerns about the consequences of testing, or possess anxiety or other emotional problems that interfere with test taking. For example, research participants who do not believe that their answers will be treated confidentially may answer in a socially desirable manner; such an instance might arise when teachers administer, collect, and score course evaluation forms from their students. It may be possible to check for and minimize such mismatches by looking for different responses between the first and second half of the test (to detect fatigue and boredom effects), developing rapport with test takers and exploring their testing concerns, and asking sensitive questions at the end of the test.

Mismatches resulting from behavioral and environmental factors may occur when observers are present, when test takers have insufficient time to adapt to testing conditions (particularly when special apparatus are required), and when inappropriate testing apparatus is used (e.g., requiring extensive computer keyboard use by persons with no computer experience or who have physical disabilities that interfere with such activity). To minimize these factors, make observers as unobtrusive as possible and provide sufficient time to adapt and practice responding to tests, tasks, and special apparatus.

Fowler (1992) investigated how ambiguous item wording could affect responses to a national health survey. A preliminary set of interviews found seven questions that contained one or more poorly defined terms. When these items were clarified and used in a second set of interviews, significantly different results were obtained. Thus, item ambiguity represents a source of systematic error in measurement (cf. Angleitner et al., 1986).

In practice, it may be difficult to separate random errors from systematic errors. For example, individuals who are uninterested in completing a test may begin to respond randomly to items or tasks. Berry et al. (1992) investigated such 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 567 items. Seven percent reported random responding to many or most of the items; students who acknowledged some random responding averaged 36 such responses. In a second study, Berry et al. (1992) found that most subjects who admitted to random responding reported having done so at the end of the test, although another sizeable group scattered responses throughout. Finally, a study of 32 applicants to a police training program found that 53% indicated that they had randomly responded to some test items.

Rater errors

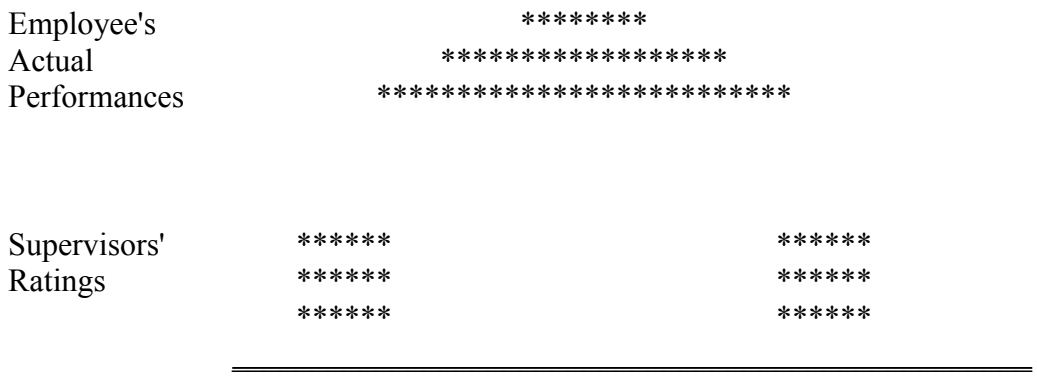
Definition. Judgments produced by raters that are irrelevant to the purpose of the assessment.

Description. Given the prevalence of ratings in counseling, occupational, and educational settings, it is no surprise that investigators have studied a number of different types of rater errors. I summarize the most important types below.

Murphy and Davidshofer (1994) described (a) halo errors, in which a rater's overall impressions about a ratee influences ratings about specific aspects of the person, (b) leniency errors, overestimates of ratee performance, and (c) criticalness errors, underestimates of the performance of ratees. To illustrate the latter two errors, suppose you are an employee who has two supervisors. Figure G.2 displays a frequency count of your actual performance; that is, it summarizes the quality of a large number of your performances. You can see that you have relatively few low- or high-quality performances, and that most of your work would be rated as being of moderate quality. In contrast, Supervisor A's ratings are below your actual performances, whereas all of Supervisor B's ratings are above your actual work quality. Your supervisors are displaying criticalness and leniency errors.

Figure G.2.

Leniency and Criticalness Errors



Hypothesis confirmation bias is a special type of error committed by test users, researchers, counselors, educators, and laypersons—in other words, everyone. It refers to the tendency to crystallize early impressions and ignore later information that contradicts the initial hypothesis (Darley & Fazio, 1980; Jones et al., 1968). Overshadowing occurs when a rater focuses on a particularly salient aspect of a person or situation (e.g., mental retardation) while ignoring other aspects that may also be important (e.g., mental illness) (cf. Reiss, 1994).

Among the solutions to rater errors are providing thorough training, calculating interrater reliability and redoing ratings if reliability is low, and rechecking raters' reliability randomly (cf. Paul, 1986).

Haverkamp's (1993) research provides an example of the hypothesis confirmation bias in a clinical context. She asked 65 counseling students to view a videotape of a counselor and client interaction. Students were provided with problem descriptions generated by the client and also asked to generate hypotheses themselves about the client's problem. After viewing the videotape, students were presented with a series of tasks (e.g., what further questions would you ask?) designed to determine the frequency of the type of information they were seeking (e.g., confirmatory, disconfirmatory, neutral, other). Haverkamp found that student counselors did not seek to confirm the hypotheses provided by the client, but did attempt to confirm their own hypotheses about the client. Such an approach, Haverkamp maintained, means that the counselor may ignore information that could support an equally plausible explanation and intervention for the client's problem.

Reliability

Definition. The consistency of measurement.

Description. The usual definition of *reliability* refers to a measurement method's ability to produce consistent scores. Thus, one might check the reliability of a measure of a trait by administering it twice to the same group of individuals one week apart and then correlating those scores. If the correlation is high (e.g., above .80), this means that the measure has good test – retest reliability (cf. Meier & Davis, 1990). A low estimate (e.g., below .70) presents a problem for subsequent interpretation of the meaning of these scores.

You can evaluate a measurement method's reliability in many different ways (Cronbach, 1984; Crocker & Algina, 1986; Murphy & Davidshofer, 1988). As summarized in Table G.5, you could calculate split-half reliability (the extent to which two halves of the same test correlate), internal consistency (the average correlation between any item and the sum of the items), alternate-form reliability (the correlation between two forms of the same test), test – retest reliability (the correlation between two administrations of the same test given to the same persons), or interrater reliability (the correlation between two raters who observe the same phenomenon). Coefficient alpha, a measure of internal consistency, currently is the most frequently used method for quantitative data because it requires only a single administration of the measurement method and is easily computed using programs such as SPSSx and SAS.

Table G.5.

Types of Reliability and Their Advantages

Type	Advantage
Split-half	Requires only a single administration of a test.
Internal consistency	Requires only a single administration of a test; easily computed.
Alternate form	Once demonstrated, provides two forms that can be employed at different intervals with minimal practice effects.
Test – retest	Provides evidence of stability over time, a major issue with trait-based tests.
Interrater	Provides evidence of stability across observers, a major issue with social science constructs.

Your theoretical understanding of the construct and its measurement method—not the ease of calculation—should guide the selection of the reliability analysis. For example, you might consider to what extent the construct resembles a trait or a state. A *trait* is a phenomenon assumed to be relatively stable, enduring, and unresponsive to environmental influences. A *state* is a transitory psychological phenomenon that changes because of situational, developmental, or psychological factors. If the construct you are measuring has significant state components, then you would expect test – retest reliability to be relatively low. It would make more sense to evaluate tests of states with a measure of internal consistency such as coefficient alpha. It may also be the case that the construct can be considered in terms of trait and state components. For example, Speilberger et al. (1970) have discussed and developed well-known measures of state and trait anxiety.

Because reliability depends partially on the sample of persons who complete a test under certain conditions, use the term *reliability estimate* when referring to the results of a reliability analysis with a set of test scores. For example, scores on Test A, when completed by a group of college students, may result in a coefficient alpha of .95. However, alpha may be considerably reduced when Test A is completed by fifth-grade students who experience difficulty comprehending Test A's items. You should not assume that a test that has been previously reliable will be so under your research or practice conditions. You should also not assume that a test you have devised will be reliable. Any such homemade tests should be evaluated for reliability and validity, at least during pilot testing (cf. Meier & Davis, 1990).

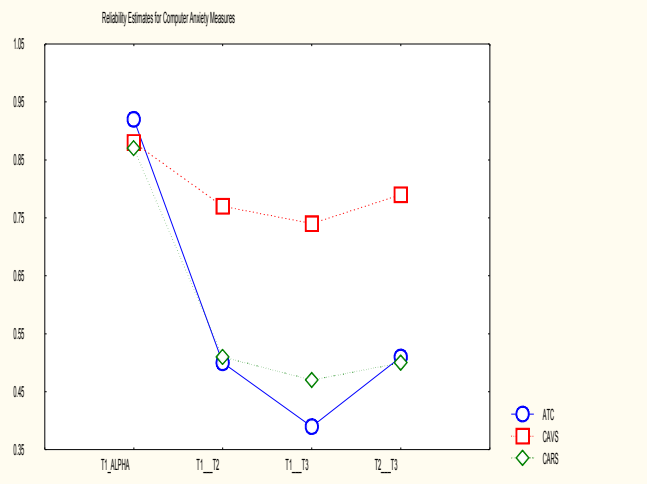
Test users typically pay the most attention to reliability when a measurement method is first developed or when methods are compared. Meier and Lambert (1991) compared three scales developed to measure individuals' comfort with computer use. They administered the Attitudes Toward Computer (ATC) scale, the Computer Aversion Scale (CAVS), and the Computer Anxiety Rating Scale (CARS) to 1,234 college students during weeks 1, 8, and 15 (Times 1, 2, and 3, respectively) of a semester. Table G.6 summarizes the reliability results and the accompanying Figure G.3 displays the results graphically.

Table G.6.

Reliability Estimates for Computer Anxiety Measures

Scale	Time 1 alpha	Test-Retest Reliability: Time 1 – Time 2 correlation	Time 1 – Time 3 correlation	Time 2 – Time 3 correlation
ATC	.92	.50	.39	.51
CAVS	.88	.77	.74	.79
CARS	.87	.51	.47	.50

Figure G.3. *Plot of all reliability estimates.*



Whereas the three Time 1 alphas are approximately equal, the CAVS shows much higher test – retest reliability. Thus, if you sought a more stable measure of computer comfort, the CAVS would be the clear choice. On the other hand, if you were interested in a measure more likely to be responsive to environmental or treatment effects, the ATC and CARS would be preferable.

Reliability can also be described in terms of agreement among raters. Research comparing multiple raters asked to observe the same phenomenon, however, often finds some degree of inconsistency (Lambert & Hill, 1994). Christensen et al. (1992) found considerable differences in mothers' and fathers' reporting on the Child Behavior Checklist about their children ages 3 – 13. Mothers reported more negative behaviors than did fathers, and parents

disagreed about the occurrence of a behavior twice as often as they agreed. Christensen et al. (1992) found more consistency with behaviors described as more disturbed, overt, and specific.

Response strategies

Definition. Processes individuals use to complete test items, problems, and tasks.

Description. Test takers sometimes employ response strategies that involve the creation or distortion of information. Examples of generative strategies include random responding, dissimulation, malingering, and social desirability. When individuals randomly respond to a measurement device they enter answers by chance. With malingering, respondents simulate or exaggerate negative psychological conditions (e.g., anxiety, psychopathology). Respondents who dissimulate attempt to fake good or bad on tests. Socially desirable responses are those that are socially acceptable or present the respondent in a favorable light.

Response sets and response styles represent similar concepts that focus more on motivational than cognitive factors (Lanyon & Goodstein, 1982). With response sets the test taker distorts answers in an attempt to generate a specific impression (e.g., "I have good work habits for this job").

Response Set: Distort toward **SPECIFIC IMPRESSION**

Social desirability is an example of a response set because the test respondent is attempting to answer items in such a way that leaves a positive impression. In contrast, with response styles there is a distortion in a particular direction regardless of item content.

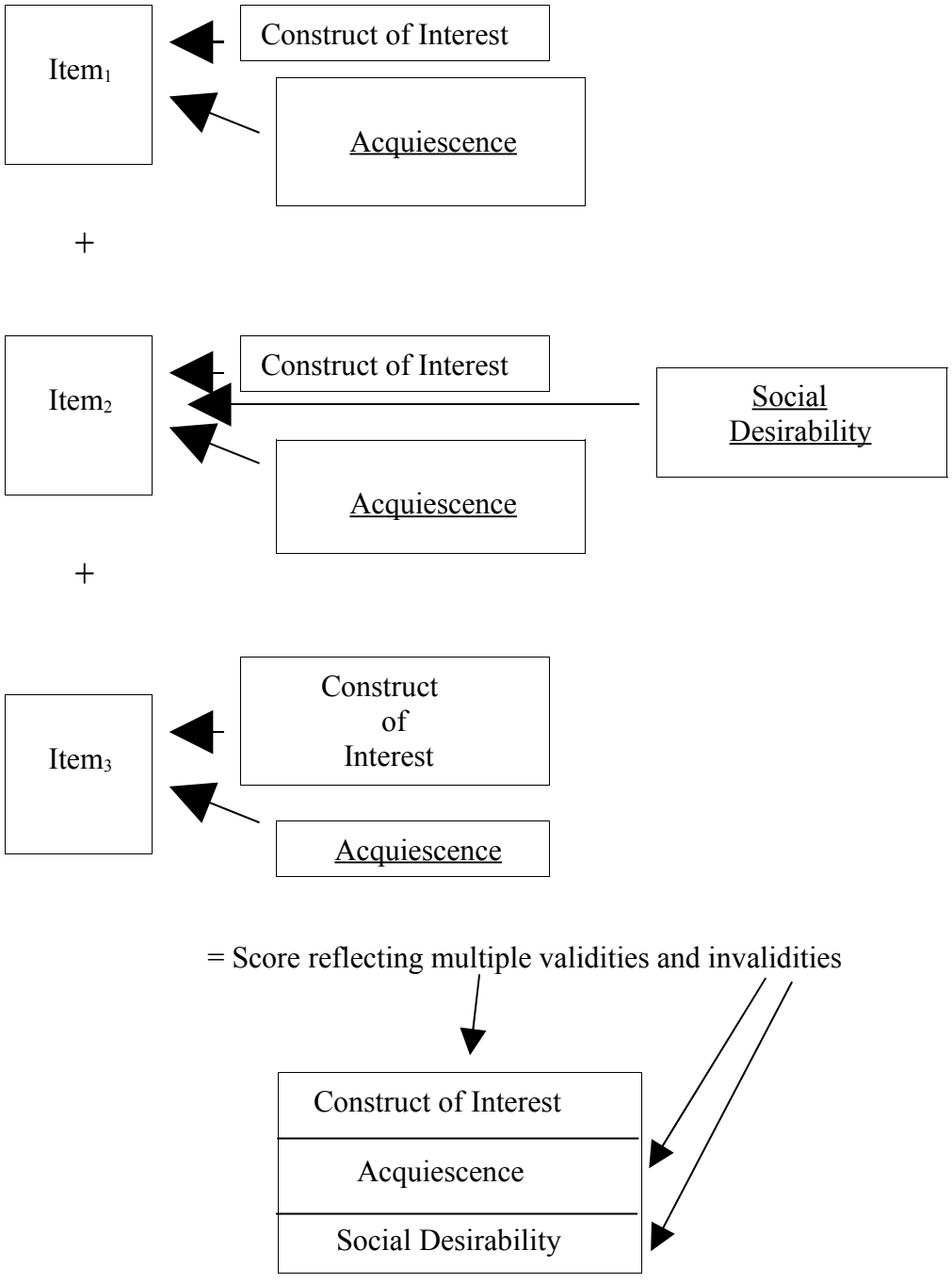
Response Style: Distort toward **PARTICULAR RESPONSE DIRECTION**

Examples of response styles are acquiescence (i.e., tendency to agree regardless of content) and criticalness (i.e., tendency to disagree regardless of content). This can be an issue with a test such as the Career Maturity Inventory (CMI) where false responses to 43 of the 50 attitude items are scored as indicating career maturity.

As shown in Figure G.4, it is possible for multiple sources of error such as acquiescence and social desirability to be influencing a single measurement or assessment method. If the method is a test, for example, summing items that contain systematic error scores produces a total score reflecting the construct *and* error sources (i.e., invalidities):

Figure G.4.

Item response and systematic errors



Response sets partially result from the clarity of item content: The more transparent the item, the more likely that a response set such as social desirability will occur (Martin, 1988). For example, Murphy and Davidshofer (1994) suggest that an item like "I hate my mother" is very clear and invites a response based on its content. If the item is ambiguous, however, then the

probability of a response style such as acquiescence increases. 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, whose intent may be less transparent to test takers, but which may also invite acquiescence or criticalness because individuals have little basis on which to respond.

An item 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 "I think Washington was greater than Lincoln." Research tends 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.

Generative responses would seem more likely with reactive or transparent tests. *Reactivity* refers to the possible distortion that may arise from individuals' awareness that they are being observed or are self-disclosing.

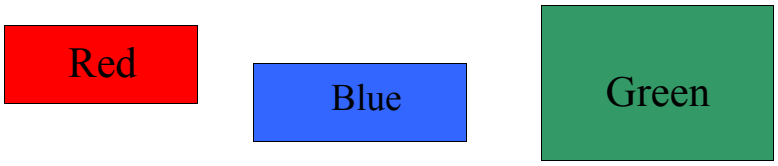
Wetter and Deitsch (1996) investigated the consistency of response to the MMPI-2 by persons faking posttraumatic stress disorder (PTSD), faking closed-head injury (CHI), or controls. The researchers asked 118 undergraduate students to imagine they were part of a lawsuit in which their faking of psychological symptoms would increase the chances of a large financial award. After reading descriptions of the disorder they were told to fake, participants completed the MMPI-2 twice (at a 2-week interval). Significantly lower reliability coefficients were found for scales completed by individuals faking CHI than obtained for controls or persons faking PTSD.

Scale types

Definition. The type and amount of information contained in test scores.

Description. Traditionally, four types of measurement scales are commonly described:

1. *Nominal* scales, which contain qualitative categories:



2. *Ordinal* scales with rank information:

Employee's Performances **** *** ****

Low Moderate High

3. *Interval* scales containing rank information with equal intervals:

Performance level

1 2 3 4 5 6 7

Poor Average Superior

4. *Ratio* scales, which contain equal intervals with a meaningful zero point:

Check the number of cigarettes smoked today:

0 1 2 3 4 5 6 7 8 9 10

11 12 13 14 15 16 17 18 19 20

Each successive type contains more information than the previous. Ratio scales, for example, provide more information about a construct than interval, ordinal, or nominal scales. Ratio scales should be the most precise if they reflect the actual values present in a phenomenon.

For example, diagnostic categories typically contain nominal information; that is, they distinguish between different types of phenomena but provide no information about differences within a particular phenomenon. Dihoff, Hetznecker, Brosvic, and Carpenter (1993) developed ordinal diagnostic criteria with 20 autistic children ages 2 – 3 years. Subgroupings of the children were identified and found to differ on behavioral measures, standardized tests, and school achievement. Dihoff et al. (1993) reported that use of the ordinal criteria promoted diagnostic agreement among therapists.

Scoring

Definition. Method by which test data are assigned to produce a score or category.

Description. Aggregating or summing individual test responses or items is the predominant method of scoring tests. For example, Luzzo (1995) summed college students' answers to the 50-item attitude scale of the Career Maturity Inventory (CMI) and found an average score of 36.84 for the 401 persons who completed it. This means that on average this group answered about 37 of the 50 items in a manner indicating a mature career attitude.

Items, tasks, and ratings can also be weighted (e.g, counted more or less in relation to other items) prior to aggregation. If you were creating a measure of aggression in children, for

example, you might possess a theoretical reason for assigning more weight to physical acts of violence (e.g., hitting, kicking) than to verbal acts (e.g., insults, threats).

Some test items are not scored per se but employed as decision trees whereby answers direct the tester toward some final decision, typically about diagnosis. Versions of the *Diagnostic and Statistical Manual of Mental Disorders* (e.g., American Psychiatric Association, 2000, or earlier versions) contain decision trees whereby diagnosticians can follow a set of branching questions that lead to a specific diagnosis. For example, the tree of differential diagnoses of Organic Brain Syndromes begins with the question, "Disturbance of attention, memory and orientation developing over a short period of time and fluctuating over time?" A “yes” answer leads to a possible diagnosis of Delirium, whereas “no” branches to the next question, and so forth, through the set of possible related diagnoses.

For example, computer scoring of tests generally eliminates errors. However, some procedures require the participant or experimenter to score a test, and here research suggests that a surprisingly high percentage of mistakes can be made. For example, Ryan, Prifitera, and Powers (1983) asked 19 psychologists and 20 graduate students to score WAIS-R (Wechsler Adult Intelligence Scale—Revised) information that had been administered to two vocational counseling clients. They found that regardless of professional experience, participants' scoring of the identical materials produced scores that varied by as much as 4 – 18 IQ points. Other examples of scoring errors with seemingly straightforward procedures abound (Worthen et al., 1993). Scoring becomes even more problematic when human judgment is introduced into the scoring procedures, as with many projective tests and diagnostic tasks.

Self-reports

Definition. Judgments made by individuals about personal attributes.

Description. Self-reports constitute one of the most frequently employed assessment methods in practice and research. Kagan (1988), for example, cited research indicating that most personality research was based on self-report questionnaires. Noting that self-reports have been employed in alcohol research since the beginning of the 20th 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).

A common tactic to assess the effectiveness of interventions is to assess change with a self-report scale. For example, Meier (1988b) created an alcohol attitudes scale to assess how drinking-related attitudes in high school and college students changed following an alcohol education program. The scale's instructions, scoring procedures, and first four items are reprinted in Table G.7.

Table G.7.

Alcohol Attitudes Scale

In this section you are asked to indicate your values and beliefs about alcohol and drinking. The most important instruction is that you be completely HONEST with your response. Rate each of the following statements according to this scale:

	Disagree strongly		Uncertain		Agree strongly
1. It's okay to have a party where drinking is the main reason people are there.	1	2	3	4	5
2. Drunk people can be funny.	1	2	3	4	5
3. Food should be provided whenever alcohol is served.	1	2	3	4	5
4. The cocktail hour before dinner should be as long as people wish.	1	2	3	4	5

Note. Items 1, 2, and 4 are reversed scored (i.e., a 1 becomes a 5, and so forth). Higher scores indicate more responsible attitudes toward alcohol use.

In addition to their use in intervention research, self-reports are frequently employed in surveys (i.e., in which the questions are delivered via mail, telephone, or personal interviews). Surveys can be used to collect information about the nature or frequency of a phenomenon in a low-cost manner.

Despite the widespread use of self-reports, test users often adopt one of the following beliefs: (a) Because individuals can self-report, self-reports must be valid, or (b) because self-reports can be easily distorted, self-reports are useless. The first position represents that taken by most early measurement theorists. In contrast, self-report critics espousing the second position have pointed to studies comparing self-reports with what the critics see as a more objective criterion, that is, overt behavior. For example, researchers consistently find some discrepancies between self-reports of psychological phenomena and overt behavior indicative of or related to the phenomena (e.g., Doleys et al., 1977; Schroeder & Rakos, 1978).

One of the most interesting problems with self-reports is the assumption that items have the same meaning across individuals (cf. Schwarz, 1999). That is, when researchers use self-report measures, they assume that all participants understand an item in the same way. A simple exercise can demonstrate that this is often not the case. Take any self-report scale (the Alcohol

Attitudes Scale shown in Table G.7, for example) and administer it to a group so individuals complete it privately and separately (e.g., in a class). When all are finished, ask them to write, beside each item, the basis on which they answered the item. Next, go through each item to determine how individuals answered. You are likely to see that individuals understood the meanings of items in quite different ways (cf. Kahn & Meier, 2001).

Standardization

Definition. Establishment of identical or similar test procedures for each respondent.

Description. Standardization is designed to reduce error by making the test conditions and environment as similar as possible for everyone who takes the test. Conditions could include such factors as the time to complete the test, the legibility of the test, and the order of administration of various subscales or tests.

When students take the GRE or LSAT, for example, no differences should exist in the testing environment. Lighting should be adequate, the temperature should be comfortable, the room should be quiet, and so forth. The use of computers with such tests, raises an interesting issue about standardization. Most test takers are likely to be familiar with paper-and-pencil media, but the introduction of computers into such testing may represent a significant change in testing conditions for a subgroup of students unfamiliar with computers.

Gay (1990) investigated irregularities in the administration of standardized tests given to grade school and high school students. She surveyed 265 teachers and eight test coordinators and found irregularities in such areas as inaccurate timing, coaching, altered answer sheets, and student cheating. Gay recommended that test administrators review a testing code of ethics and be monitored for proper administration.

States

Definition. Transitory psychological phenomena that change because of psychological, developmental, or situational factors.

Description. States are internal or external psychological characteristics that vary. Even theorists interested in measuring traits acknowledge the presence of state effects in psychological testing. For example, many cognitive abilities such as reading and mathematics skills may have a genetic component, but some aspects of those skills may still change as a result of development (e.g., improvement with age) and interventions (e.g., education).

Collins (1991; Collins & Cliff, 1990) described a test construction method appropriate for measuring development. Collins (1991) was interested in predicting and measuring patterns of change in grade school students' acquisition of mathematical skills. She proposed that children first learned addition, then subtraction, multiplication, and division, in that order. Such a sequence can be characterized as *cumulative* (i.e., abilities are retained even as new abilities are gained), *unitary* (i.e., all individuals learn in the same sequence), and *irreversible* (i.e., development is always in one direction; Collins, 1991). This sequence can be employed to search

for items and tasks that do and do not display the expected sequence of mathematics performance over time.

The State – Trait Anxiety Inventory (STAI; Spielberger et al., 1970) is one of the most widely used state – trait measures. 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," and 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 taker "generally" feels.

The instructions do seem to produce the desired difference: Test – retest reliabilities for the state scale are considerably lower than for the trait (Spielberger et al., 1970). The STAI typically correlates at moderate to high levels with other measures of anxiety (e.g., Bond, Shine, & Bruce, 1995; Kaplan, Smith, & Coons, 1995). For example, Bond et al. (1995) asked patients with anxiety disorders and normal controls to complete the STAI and a visual analogue scale rating of anxiety. In this approach participants mark along a 100 mm line to indicate their levels of anxiety; such visual measures are useful when frequent measures of mood are necessary and when reading is a problem. Bond et al. (1995) found correlations in the .50s and .60s between the two scales, suggesting a modest degree of overlap.

Statistics related to measurement

Definition. Statistics employed to facilitate the interpretation of test scores.

Description. Making sense of test scores often depends at least partially on understanding a number of statistical indices normally computed with tests. For example, test developers usually examine (and present information about) the frequency distribution of all test scores to determine if it is normally distributed. Similarly, developers may present information about the range and standard deviation of scores to examine whether sufficient individual differences exist. Below, I describe statistics commonly used during the test interpretation process.

A mean or average is a measure of central tendency; that is, in a group of scores, where is the middle or most representative value? The mean is found by summing the scores in a group and dividing by the number of scores. Other measures of central tendency are the *median* and the *mode*. These measures provide a typical score that characterizes the performance of the entire sample. A mean, along with the other measures of central tendency, is particularly useful for comparing different groups (such as children of different ages) who take the same test as well as describing individuals in relation to a group's set of scores (where does one individual score on a course quiz in relation to the whole class?).

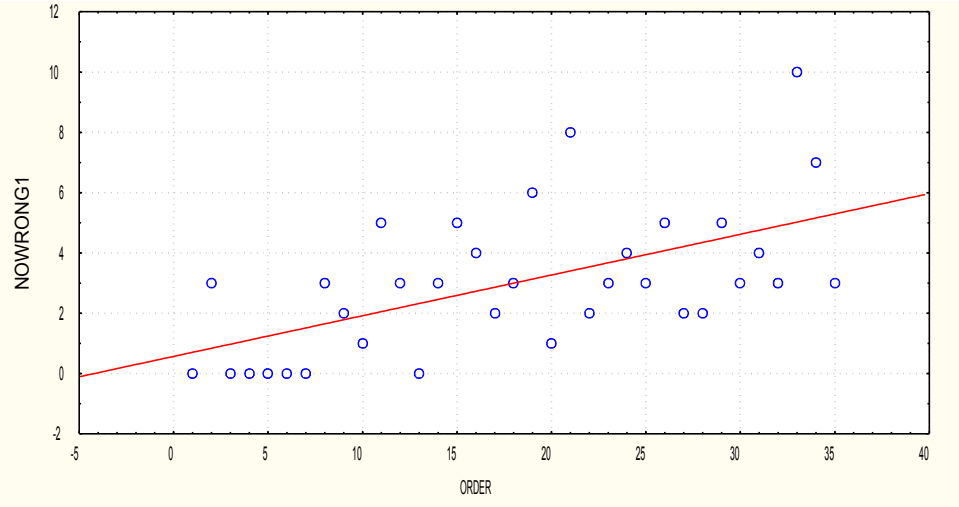
Besides knowing the central tendency in a group of scores, it is often useful to know how dispersed the scores are. One such index of dispersion, the standard deviation, refers to the average deviation of scores from the mean. The larger the standard deviation, the more widely spread the distribution of scores.

A *correlation* refers to the extent to which two variables covary. A correlation coefficient expresses the degree of relationship between two sets of scores. For example, if the highest-scoring individual on Test 1 has also obtained the top score on Test 2, and the second-best individual on Test 1 is also second-best on Test 2, as so on down to the lowest-scoring individual on each test, a perfect positive correlation would exist (+1.00). If there is a complete reversal of scores, so that the highest-scoring individual on variable 1 obtains the lowest score on variable 2 and so forth, there would be a perfect negative correlation (-1.00). A zero correlation indicates the absence of a relationship between two variables, such as might occur by chance. Thus, correlation coefficients fall between the range of -1.00 and +1.00.

The data that form the basis of a correlation coefficient can also be graphed. The graph shows the relation between the number of quiz questions students answered incorrectly in relation to the order in which they completed quiz and turned it in:

Figure G.5

Scatterplot of the Number of Incorrect Answers with Order in Which Quiz Completed



.

Figure Caption. A scatterplot showing the relation between the number of incorrect responses to a quiz (on the vertical axis) and the order in which the quiz was turned into the instructor (horizontal axis). The line through the scatterplot suggests that persons who had fewer wrong answers turned in their quizzes earlier.

As the scatterplot shows, students who completed the quiz sooner generally had a fewer number of incorrect answers. However, the relationship is not perfect; for example, the second student to turn in the quiz had three incorrect answers. The correlation computed for these data is .51, with a mean of 1.66 and a standard deviation of 2.29. Although the reason to present these (actual) data is to explain the idea of correlation, do you have any substantive idea about why this relation should exist? In other words, how would you explain why students who finished a quiz faster generally had better grades?

A standard score or *z* score is a transformation of a raw score to show how many deviations from the mean that score lies. The formula is:

z = (Raw score - mean) / standard deviation

Thus, *z* equals the person's raw score minus the mean of the group of scores, divided by the standard deviation of the group of scores. Frequently, the best information that a test score can give us is the degree to which a person scores in the high or low portion of the distribution of scores. The *z* score is a quick summary of the person's standing: Positive *z* scores indicate that the person was above the mean, and negative scores indicate that the person scored below the mean.

Other types of standard scores have also been developed, including stanines, deviation IQs, sten scores, and *T*-scores. *T*-scores, for example, allow us to translate scores on a test to a distribution of scores of our choice. *T*-scores use arbitrarily fixed means and standard deviations and eliminate decimal points and signs. The formula is:

$T = (SD * z) + X$

where *SD* is the chosen standard deviation, *X* is the chosen mean, and *z* is the standard score for a person's score on a test. For example, I might find it simpler to give feedback using a distribution of scores whose mean is 50 and whose standard deviation is 10. If a person had a score on a test whose *z* equaled -.5, the *T*-score would be:

$(10 * -.5) + 50 = 45$

Tests such as the Analysis of Learning Potential use a fixed mean of 50 and a standard deviation of 20, and the SAT and GRE use 500 as the mean and 100 as the standard deviation. Again, the *T*-score provides a convenient translation of scores so that they might be more understandable during test interpretation.

Acknowledging that error influences any particular testing occasion, the standard error of measurement (*SEM*) is the standard deviation that would be obtained for a series of measurements of the same individual if the individual did not change on the measured construct over that time period. For example, assume that I administered a test measuring a stable trait 10 times to a particular person. If that person received the same score for each test occasion, there would be no error of measurement. In reality, however, the test score would vary for each testing, and SEM is a statistic designed to summarize the amount of variation. If you have an estimate of a test's reliability, SEM can be calculated as follows:

SEM = Standard deviation * SqRt (1 - *r*)

Thus, *SEM* equals the standard deviation of the group of scores times the square root of 1 minus the reliability estimate. *SEMs* help us know the extent to which an individual's particular test score can be trusted as indicative of the person's true score on the test.

Finally, the standard error of estimate (*SEE*) helps us know the trustworthiness of a test score's ability to predict a criterion of some sort. Just as no test produces the same score when administered repeatedly to a person, no single score will be associated with the identical score on a criterion. Thus, the SEE refers to the spread of scores around a criterion, or more precisely, the standard deviation of criterion scores for individuals who all have the same score on the predictor test. The formula for SEE is:

SEE = Standard deviation * SqRt (1 - v^2)

SEE equals the standard deviation for the group of criterion scores times the square root of 1 minus the squared validity coefficient (*v*). The validity coefficient is simply the correlation between the predictor test and the criterion that one is attempting to predict. For example, graduate schools frequently screen candidates on the basis of their GRE scores because GRE scores (the predictor test) have been shown to have a modest correlation with first-year GPA (the criterion). SEE helps us gain a sense of how large the variation is likely to be around the criterion, given an individual's particular test score.

For example, let's walk through simple computations of the standard score, SEM, and SEE.

Let's start with the z or standard score. Assume that the following represents a group of test scores. To compute a *z* score, I need the mean (which equals 87.95) and standard deviation (6.82) for this group of scores.

78	90	95	70	85
88	85	85	90	83
94	95	88	91	99
93	81	94	91	84

If your score on this test was 90, your z score would be:

(90-87.95)/6.82 = .30

A z of .30 indicates you scored slightly above the mean in this group of scores.

However, if your score was 70, your z score would be:

(70-87.95)/6.82 = -2.63

This z indicates your score was well below the mean.

SEM depends on the standard deviation and the reliability of the particular test. If I have a test with a reliability estimate of .90 (high) and a standard deviation of 15, then SEM equals:

15 * SqRt (1-.9) = 4.7

Thus, 4.7 represents 1 standard deviation unit for the distribution of scores around the individual's true score. However, if the test's reliability estimate was .7, SEM increases:

15 * SqRT (1-.7) = 8.21

Thus, the lower the reliability of the test, the less confidence I have that an individual's true score is close to the actual score obtained.

Finally, with SEE, I need the correlation between the test and criterion as well as the standard deviation for the group of criterion scores. If the correlation between test and criterion equaled .61, and the standard deviation for the criterion scores equaled 100, then SEE would be:

100 * SqRt (1-[.61*.61]) = 79

Thus, 79 represents 1 standard deviation unit around the criterion score. However, if the correlation between predictor and criterion dropped to .30, the SEE would increase:

100 * SqRT (1-[.30*.30]) = 95

Thus, the lower the correlation, the less confidence I have that the predicted criterion score is the true score the individual would actually obtain.

Testing and instrumentation effects

Definition. Change on test scores that result from the use of a particular measurement method.

Description. Repeatedly administering any type of measurement device or assessment procedure can produce changes on test scores. For example, participants who take a test more than once may evidence *practice effects*, improving their scores upon repeated administrations without the presence of any intervention. They may better their performance on a standardized test, for example, by starting to employ memory strategies that improve their recall of information. *Pretest sensitization* refers to instances when pretesting influences participants' behavior during and after an intervention. Both pretest sensitization and practice effects can be grouped as testing effects: Something changes simply because the person completed the test.

Cook and Campbell (1979) suggested that *instrumentation effects* refers to pretest – posttest differences that result from changes in how respondents view the measuring instrument, not as a result of the intervention. *Response-shift bias* occurs when respondents' understanding or awareness of the measured construct changes as a result of an intervention or other experiences (Howard et al., 1979; Sprangers & Hoogstraten, 1987). Essentially, respondents experience a change in their frame of reference. Similarly, Golembiewski (1989) proposed that *alpha change* indicates that changes in pre – post scores correspond to actual changes produced by an intervention, whereas in *beta change* respondents alter the intervals of the scale. Finally, *gamma change* involves a shift in the entire meaning of the scale.

Note that these terms are sometimes used interchangeably in the testing literature and that some overlap in meaning is present. All of these effects take place in the context of repeated administrations of a test, usually with an intervening treatment.

Casey, Ferguson, Kimura, and Hachinski (1989) investigated whether carotid artery surgery would result in cognitive improvement in 36 patients without stroke. They evaluated WAIS (Wechsler Adult Intelligence Scale) and other memory performance measures with two comparable groups of patients before and 6 – 8 weeks after surgery. Casey et al. found that significant improvements on some measures were equivalent across groups, indicating that the changes were due to practice and not the surgery.

Tests

Definition. Tools or systematic procedures employed to observe some aspect of human behavior and describing it with a numerical scale or category system.

Description. Tests are employed to produce a description of some aspects of individuals or groups. Historically, most tests have been developed with the idea of selection, that is, to classify or make decisions about large groups of individuals (Haywood, Brown, & Wingenfeld, 1990). For example, tests like the SAT and GRE are intended to help administrators make admission decisions. Historically, many intelligence tests were intended to measure *g*, an intelligence factor that presumably could influence a person's performance on a wide variety of tasks.

In the book I employ the term *tests* to refer to any type of measurement or assessment method. Synonyms for *tests* include *scales*, *inventories*, *questionnaires*, *checklists*, and *rating scales* (Aiken, 1996). Similarly, researchers sometimes use the term *operation* to refer to any single method of gathering data.

Traits

Definition. Consistent personal characteristics often assumed to be of biological origin and resistant to environmental influences.

Description. The idea of individual differences indicates that individuals can behave differently on the same tasks or in the same situations (Dawis, 1992). Stable individual differences are traits. Theorists usually assume that traits are normally distributed in the population; that is, a frequency distribution of any trait should resemble a bell-shaped curve.

Selection testers typically treat measurement as nomothetic. That is, they are measuring traits—such as neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness (McCrae & Costa, 1987)—presumed to be present in every person. In contrast, idiographic assessors believe that individuals are unique and that traits may or may not be present in different individuals. In addition, many test theorists belief that traits are *latent*, that, is, unobservable characteristics that may be indicated by clusters of behaviors. If no single behavior can define a construct (i.e., no single operational definition exists), then clusters of

behaviors may be able to do so. For example, no single behavior is assumed to be indicative of intelligence.

For example, the most significant contemporary work in the area of traits has to do with research on the Big Five. The *Big Five* refers to the consensus reached by personality researchers about five traits considered the basic structure of personality. These traits are neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. Although this research remains open to alternative explanations (cf. Almagor, Tellegen, & Waller, 1995; Block, 1995), support for the Big Five interpretation (John et al., 1988; McCrae and Costa, 1987) has been bolstered by factor analyses of trait descriptions produced by different methods (such as ratings by others and self-report) and different samples (i.e., cross-cultural).

Unobtrusive measurement

Definition. Measurement that occurs when the participant is unaware of being observed.

Description. Webb et al. (1981) described four types of unobtrusive measurement: archival records, observation in contrived settings, physical traces, and simple observation. *Archival records* refers to stored information, typically put to a use for which the data were not originally intended (e.g., school records used in a research study). *Observation in contrived settings* indicates that an observer collects data in a setting, such as a laboratory, established by the researcher for unobtrusive observation; alcohol researchers, for example, have set up fake bars to observe participants' drinking under varied conditions. *Physical traces* are artifacts left by persons indicative of some activity or characteristic (e.g., garbage). *Simple observation* indicates that an observer watches and records behaviors of interest to the observer. In all cases, such data collection occurs without the awareness of the participant, thereby decreasing problems with *reactivity*, the tendency for individuals to change their behavior when they are aware of being observed.

The advent of microcomputers provides another avenue for unobtrusive measurement. Meier and Wick (1991), for example, investigated the use of a computer-assisted instruction (CAI) program for alcohol education as an unobtrusive measure of alcohol knowledge and drinking behavior. The CAI program includes a simulation, designed like a computer game, to demonstrate blood alcohol levels (BALs) for participant-selected drinking experiences. Participants first enter information about a drinking experience, including their weight, number of drinks, and time during consumption. All of this information—weight, number of drinks, time period for consumption, and BAL estimate—is recorded unobtrusively on the computer's disk. Meier and Wick (1991) found that participants' unobtrusively recorded reports of consumption in the simulation were significantly correlated with other self-report measures of recent drinking, intent to drink, and attitudes toward alcohol use.

Four problems accompany many attempts at unobtrusive measurements (Webb et al., 1981; Kazdin, 1992; Meier & Wick, 1991). First, the behavior of individuals in naturalistic or contrived situations, for example, is unlikely to be a direct reflection of a single construct. Alcohol consumption in a bar, for example, will be influenced by one's physiological and psychological states as well as the social context. Second, researchers must expend considerable effort to obtain an unobtrusive measurement; administering a self-report scale to alcohol

treatment participants is much easier than creating a simulated bar or observing subjects drink on weekend nights. Third, collecting unobtrusive measurements without arousing participants' suspicions may be difficult; in such situations (e.g., a simulated bar), participants are likely to be guessing at hypotheses, which they will not share with experimenters. Fourth, ethical questions are frequently raised with unobtrusive measurement; any type of direct or indirect deception requires substantial justification, particularly with institutional review boards.

Abler and Sedlacek (1986) reviewed examples of unobtrusive measurement. In one study, researchers examining 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 a link between errors made in filling out college orientation applications and subsequent dropouts (Sedlacek et al., 1984). Similarly, Epstein (1979) reported a study in which students' self-reported stress was significantly correlated with the number of erasures on exam answer sheets, number of absences, and number of class papers that were not turned in.

Validity

Definition. What a test measures, or what inferences can be drawn from test scores.

Description. After reliability, validity is the second major concept used for evaluating tests. Although many different types of validity have been described in the measurement literature, little consensus exists about which validity analyses are most useful. Murphy and Davidshofer (1988) concluded that "it would be fair to say that any type of data or statistic might be useful in determining" validity (p. 103). Anastasi (1986) reached a similar conclusion: "Almost any information gathered in the process of developing or using a test is relevant to its validity" (p. 3).

Murphy and Davidshofer (1988) observed that when assessing validity, test developers typically have no standards to compare tests against. It is for that reason that measurement theorists seldom use the term *accuracy* when discussing tests. Instead, test developers gather evidence from a variety of sources to demonstrate validity. Contemporary measurement theorists also note that all tests have *multiple validities*, that is, various sources that influence test performance and scores (Wiley, 1991). A universal, usually undesired influence on all tests is method variance, that portion of the test score attributable to the method of obtaining data (Campbell & Fiske, 1959).

Among the types of validity discussed in quantitative testing (and summarized in Table G.8) are:

1. A test has *face validity* when its item content appears to match the purpose of the test. A cynical synonym is *cash validity*: The more a test appears to measure what it is supposed to measure, from the perspective of test purchasers, the more cash the test accrues for its publishing company. Although professionals sometimes choose tests on the basis of their face validity, test content does not ensure *construct validity*.

2. *Content validity* refers to whether the content of a test is representative of the universe of relevant content. A test may or may not tap into all of a construct's important characteristics.

3. *Criterion validity* refers to the correlation of test scores with a relevant criterion. Similarly, *predictive validity* refers to the degree to which a test can predict future performance on a criterion. Concurrent validation occurs by correlating a test and a criterion administered at the same time point.

4. *Incremental validity* refers to a test's ability to increase the level of a prediction. For example, if undergraduate GPA correlates .3 with graduate school grades, can a test like the GRE improve the prediction of school performance above .3?

5. Tests of *convergent validity* (i.e., high correlation between two similar tests) and *discriminant validity* (i.e., low correlation between two tests of related, but dissimilar constructs) are conducted to assess *construct validity*, whether a test measures the construct it is intended to measure. C*onstructs* are abstract summaries of natural regularities indicated by observable events. All types of validity evidence ultimately relate to the construct validity of a test. For example, predictive validity depends on construct validity, because it is the phenomenon that test and criterion measure that determines the relation between the two.

Table G.8.

Types of Quantitative Validity Concepts and Their Advantages

Type	Advantage
Face	Tests with face validity make sense to test takers, usually increasing their cooperativeness.
Content	Tests with content validity ensure practical relevance for the test administrator and test taker.
Criterion/ predictive/ concurrent	Tests with such validity typically enable prediction of important behaviors.
Incremental	Enhances maximum prediction of important behaviors.
Construct	Tests with construct validity enable an understanding of relations among constructs; can assist test developers to increase content and criterion/predictive/concurrent validity.

In the beginning decades of educational and psychological measurement, predictive or criterion validity was viewed as the most important type of validity. That is, test administrators

gave tests for the purpose of selecting individuals in and out of settings such as schools and jobs. Ceci (1991) summarized the predictive validity of the first major type of test, the IQ test:

> Although it takes little more than 90 min to administer, an IQ test is alleged to capture much of what is important and stable about an individual's academic, social, and occupational behavior. In addition to their well-documented prediction of school grades (r = .55, on average; Anastasi, 1968; Matarazzo, 1970), IQ scores have been reported to have impressive validity coefficients for predicting everything from mental health and criminality to marital dissolution rates and job performance (Gordon, 1976, 1980, 1987; Gottfredson, 1986; Hunter, 1983, 1986). For example, IQ scores have been shown to predict postal workers' speed and accuracy of sorting mail by zip code, military recruits' ability to steer a Bradley tank through an obstacle course, mechanics' ability to repair engines, and many other real-world endeavors (see Hunter & Schmidt, 1982; Hunter, Schmidt, & Rauschenberg, 1984). Moreover, IQ has been touted as a better predictor of such accomplishments than any other measure that has been studied thus far. (p. 703)

However, there is a paradox: Although cognitive ability tests have substantial predictive validity, their construct validity remains in question. That is, there exists little consensus about what such tests actually measure.

Among the types of validity proposed in the literature about qualitative assessment are five described by Maxwell (1992):

1. *Descriptive validity*, which has to do with the factual accuracy of what assessors observed in a particular situation, person, or event (e.g., you accurately report what a person actually said).

2. *Interpretive validity*, reporting accurately the meaning of events as individuals perceive them (e.g., conscious and unconscious thoughts, feelings, and beliefs). These data can be accessed directly but must be constructed.

3. *Theoretical validity*, the accuracy of the theoretical constructs a researcher develops during the course of a study. This includes a description as well as an explanation of a phenomenon.

4. *Generalizability*, the extent to which an account can be accurately extended to other persons, settings, or times.

5. *Evaluative validity*, the accuracy of an evaluative or judgmental framework of a phenomenon. For example, an observer may see a student throw an eraser at a teacher and write, "The student acted out and threw an eraser at the teacher." Evaluative validity refers to whether the researcher's evaluative judgment was accurate (i.e., did others in the social context view the act as inappropriate?).

References

Note: These are cited in the Glossary only; the remainder can be found in the References section of *Measuring Change*.

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.

Almagor, M., Tellegen, A., & Waller, N. G. (1995). The Big Seven model: A cross-cultural replication and further exploration of the basic dimensions of natural language trait descriptors. *Journal of Personality & Social Psychology, 69*, 300-307.

Aronow, E., & Moreland, K. L. (1995). The Rorschach: Projective technique or psychometric test? *Journal of Personality Assessment, 64*, 213-228.

Austin, J. T., & Hanisch, K. A. (1990). Occupational attainment as a function of abilities and interests: A longitudinal analysis using Project TALENT data. *Journal of Applied Psychology, 75*, 77-84.

Benes, K. M. (1995). Review of the Social Skills Rating System. In J. C. Conoley & J. C. Impara (Eds.), *The Twelfth Mental Measurements Yearbook* (pp. 965-967). Lincoln, NE: Buros Institute of Mental Measurements.

Blais, M. A., Norman, D. K., Quintar, B., & Herzog, D. B. (1995). The effect of administration method: A comparison of the Rapaport and Exner Rorschach systems. *Journal of Clinical Psychology, 51*, 100-107.

Blaha, J., & Wallbrown, F. H. (1996). Hierarchical factor structure of the Wechsler Intelligence Scale for Children--III. *Psychological Assessment, 8*, 214-218.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Bond, A. J., Shine, P., & Bruce, M. (1995). Validation of visual analogue scales in anxiety. *International Journal of Methods in Psychiatric Research, 5*, 1-9.

Boyce, B. R., Meadow, C. T., & Kraft, D. (1994). *Measurement in information science*. New York: Academic Press.

Cacioppo, J. T., & Tassinary, L. G. (1990). Inferring psychological significance from physiological signals. *American Psychologist, 45,* 16-28.

Calvert, E. J., & Waterfall, R. C. (1982). A comparison of conventional and automated administration of Raven's Standard Progressive Matrices. *International Journal of Man-Machine Studies, 17*, 305-310.

Casey, J. E., Ferguson, G. g., Kimura, D., & Hachinski, V. C. (1989). Neuropsychological improvement versus practice effect following unilateral carotid endarterectomy in

patients without stroke. *Journal of Clinical & Experimental Neuropsychology, 1*1, 461-470.

Collins, L. M., & Cliff, N. (1990). Using the longitudinal Guttman simplex as a basis for measuring growth. *Psychological Bulletin, 108*, 128-134.

Comrey, a. L., Bacher, T. E., & Glaser, F. M. (1973). *A source book for mental health measures*. Los Angeles, CA: Human Interaction Research Institute.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80*, 565-579.

Cronin, C. (1995). Construct validation of the Strong Interest Inventory Adventure Scale using the Sensation Seeking Scale among female college students. *Measurement & Evaluation in Counseling & Development, 28*, 3-8.

Davis, C., & Cowles, M. (1989). Automated psychological testing: Methods of administration, need for approval, and measures of anxiety. *Educational & Psychological Measurement, 49*, 311-320.

Dihoff, R. E., Hetznecker, W., Brosvic, G. M., & Carpenter, J. N. (1993). Ordinal measurement of autistic behavior: A preliminary report. *Bulletin of the Psychonomic Society, 31*, 287-290.

Exner, Jr., J. E. (1978). *The Rorschach: A comprehensive system. Vol 2*. Wiley: New York.

Exner, Jr., J. E. (1986). *The Rorschach: A comprehensive system: Vol. 1. Basic foundations* (2nd ed.). New York: John Wiley & Sons, Inc.

Fowler, F. J. (1992). How unclear terms affect survey data. *Public Opinion Quarterly, 56*, 218-231.

Gay, G. H. (1990). *Standardized tests: Irregularities in the administering of tests affect test results*. Dissertation Abstracts International, 51(03), 0828A. (University Microfilms No. AAI9020156).

Golembiewski, R. T. (1989). The alpha, beta, gamma change typology. *Group and Organization Studies, 14*, 150-154.

Guastello, S. J., & Rieke, M. L. (1990). The Barnum effect and validity of computer-based test interpretations: The Human Resource Development Report. *Psychological Assessment*, *2*, 186-190.

Haywood, H. C., Brown, A. L., & Wingenfeld, S. (1990). Dynamic approaches to psychoeducational assessment. *School Psychology Review, 19*, 411-422.

Henslin, J. M. (1993). *Sociology*. Boston: Allyn and Bacon.

Hood, A. B., & Johnson, R. W. (1991). *Assessment in counseling: A guide to the use of psychological assessment procedures*. Alexandria, VA: American Counseling Association.

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. K., Nance, D. W., & Gerber, S. K. (1979). Internal invalidity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. *Applied Psychological Measurement, 3*, 1-23.

Jung, C. (1923). *Basic writings*. New York: Modern Library.

Kaplan, D. M., Smith, T., & Coons, J. (1995). A validity study of the subjective unit of discomfort (SUD) score. *Measurement & Evaluation in Counseling & Development, 27*, 195-199.

Kazdin, A. E. (1992). *Research design in clinical psychology* (2nd ed.). Boston: Allyn & Bacon.

Lanyon, R. I., & Goodstein, L. D. (1982). *Personality assessment* (2nd ed.). New York: Wiley.

Levy, P., & Goldstein, H. (1984). *Tests in education: A book of critical reviews*. New York: Academic Press.

Luzzo, D. A. (1995). Gender differences in college students' career maturity and perceived barriers in career development. *Journal of Counseling & Development, 73*, 319-322.

Lyytinen, P. (1995). Cross-situational variation on children's pretend play. *Early Child Development & Care, 105*, 33-41.

Matarazzo, J. D. (1992). Psychological testing and assessment in the 21st century. *American Psychologist, 47,* 1007-1018.

Maxwell, J. A. (1992). Understanding and validity in qualitative research. *Harvard Educational Review, 62,* 279-300.

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests* (College Board Report No. 88-8). New York: College Entrance Examination Board.

Meier, S. (1988b). An exploratory study of a computer-assisted alcohol education program. *Computers in Human Services, 3*, 111-121.

Merton, R. K. (1956). The social-cultural environment and anomie. In H. L. Witmer and R. Kotinsky (Eds.), *New perspectives for research on juvenile delinquency* (pp. 24-50). Washington, DC: U. S. Department of Health, Education, and Welfare.

Merton, R. K. (1968). *Social theory and social structure*. New York: Free Press.

Murphy, K. R., & Davidshofer, C. O. (1994). *Psychological testing* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Musser, L. M., & Markus, A. J. (1994). The Children's Attitudes Toward the Environment Scale. *Journal of Environmental Education, 25*, 22-26.

Nichols, R. C. (1980). Individual differences in intelligence. In J. F. Adams (Ed.), *Understanding adolescence* (4th ed., pp. 164-206). Boston: Allyn & Bacon).

Ornberg, B., & Zalewski, C. (1994). Assessment of adolescents with the Rorschach: A critical review. *Assessment, 1*, 209-217.

Parker, K. C., Hanson, R. K., & Hunsley, J. (1988). MMPI, Rorschach and WAIS: A meta-analytic comparison of reliability, stability, and validity. *Psychological Bulletin, 103*, 367-373.

Patterson, W. M., Dohn, H. H., Bird, J., & Patterson, G. A. (1983). Evaluation of suicidal patients: The SAD PERSONS scale. *Psychosomatics, 24*, 343-349.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.

Reiss, S. (1994). Psychopathology in mental retardation. In N. Bouras (Ed.), *Mental health in mental retardation* (pp. 67-78). Cambridge: Cambridge University Press.

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (1991). *Measures of personality and social psychological attitudes*. New York: Academic Press.

Ryan, J. J., Prifitera, A., & Powers, L. (1983). Scoring reliability on the WAIS-R. *Journal of Consulting & Clinical Psychology, 51*, 149-150.

Schwartz, J. M., Stoessel, P. W., Baxter, Jr., L. R., Martin, K. M., & Phelps, M. E. (1996). Systematic changes in cerebral glucose metabolic rate after successful behavior modification treatment of obsessive-compulsive disorder. *Archives of General Psychiatry, 53*, 109-116.

Sprangers, M., & Hoogstraten, J. (1987). Response-style effects, response-shift bias, and a bogus-pipeline. *Psychological Reports, 61*, 579-585.

Swiercinsky, D. P. (Ed.). (1985). *Testing adults*. Kansas City, MO: Test Corporation of America.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

Weaver, S. J. (Ed.) (1984). *Testing children*. Kansas City, MO: Test Corporation of America.

Wetter, M. W., & Deitsch, S. E. (1996). Faking specific disorders and temporal response consistency on the MMPI-2. *Psychological Assessment, 8*, 39-47.

Wilder, J. (1957). The law of initial values in neurology and psychiatry. Journal of Nervous & *Mental Disease, 125*, 73-86.

Wilder, J. (1967). *Stimulus and response: The law of initial value*. Bristol: J. Wright.