PSYCHOLOGICAL TESTING: BASIC CONCEPTS AND COMMON MISCONCEPTIONS

Anne Anastasi obtained an AB degree from Barnard College and a PhD from Columbia University. She taught at Barnard and then at Queens College of the City University of New York, where she was the first chairperson of the psychology department at the newly established college. Next she joined the Graduate Faculty of Arts and Sciences of Fordham University, where she subsequently served two terms as chairperson of the joint graduate and undergraduate psychology departments. She retired in 1979 with the title of Professor Emeritus.

A past president of the American Psychological Association (APA), Anastasi held many other offices, including the presidencies of the Eastern Psychological Association, the APA Divisions of General Psychology, and of Evaluation and Measurement. She is the recipient of the APA Distinguished Scientific Award for the Applications of Psychology, the Educational Testing Service Award for Distinguished Service to Measurement, the American Educational Research Association Award for Distinguished Contributions to Research in Education, the Edward Lee Thorndike Medal for Distinguished Psychological Contribution to Education awarded by the APA Division of Educational Psychology, and the American Psychological Foundation Gold Medal. Her publications include *Psychological Testing, Differential Psychology,* and *Fields of Applied Psychology,* as well as some 150 monographs and journal articles.

ANNE ANASTASI

PSYCHOLOGICAL TESTING: BASIC CONCEPTS AND COMMON MISCONCEPTIONS

As I thought about the purpose of the G. Stanley Hall Lectures and about the audience to which they are addressed, I decided to orient my presentation toward three major objectives. The first objective is to consider what test users, test takers, and undergraduates in general should know about contemporary psychological testing. This rather broad objective covers both an overview of basic concepts and an updating of information. The second is to examine some of the most prevalent popular misconceptions about tests that lead to misuses of tests and misinterpretation of test scores. The third is to illustrate how complex and sophisticated concepts and methodologies can be presented simply and in readily comprehensible terms. I shall illustrate my own efforts to meet the third objective with the treatment of statistical topics, ranging from the standard deviation to factor analysis and item response theory. The three objectives are not segregated into different sections of this paper, but are intermingled within my discussions of appropriate topics.

How to Evaluate a Psychological Test

The fundamental question that a nonspecialist needs to answer about tests is, "How can I evaluate or judge a test?" This question can, in

turn, be broken down into more specific questions. First, what kind of information can this test provide about the person who takes it? Second, how good is this test—that is, how well does it meet general standards that apply to all tests? Third, how well does the test fit the particular purpose for which it is to be used?

Much of the information needed to answer these questions can be found in a properly prepared technical manual, which should accompany any test that is ready for operational use. Some of the more responsible test publishers also provide simplified but accurate versions of the necessary information, in order to help technically untrained persons to gain some understanding of particular tests. An outstanding example is the booklet about the Scholastic Aptitude Test entitled *Taking the SAT*, distributed by the College Entrance Examination Board (1983a, 1983b, 1983c), together with the accompanying slide shows for use by counselors in orienting students. Informative brochures have likewise been developed for the Graduate Record Examination and for several other testing programs conducted by Educational Testing Service. A few major commercial publishers have also taken steps to disseminate explanatory materials among test takers and other concerned groups.

Despite the availability of these simplified materials that are designed for the general public, the *test user* cannot properly evaluate a test without having some familiarity with the major steps in test construction and some knowledge of the principal psychometric features of tests, especially as they pertain to norms, reliability, and validity. By test user, as contrasted to test constructor, I mean essentially anyone who has the responsibility either for choosing tests or for interpreting scores and using them as one source of information in reaching practical decisions. Many test users serve both of these functions. Test users include teachers, counselors, educational administrators, testing coordinators, personnel workers in industry or government, clinicians who use tests as aids in their practice, and many others in an increasing number and variety of real-life contexts. Anyone responsible for choosing tests or using test results needs some technical knowledge in order to understand the tests properly. If a test user, in this sense, lacks adequate background for this purpose, he or she needs access to a properly qualified supervisor or consultant. Many current criticisms of tests are actually directed not to the tests themselves, but rather to misuses of tests by unqualified users.

Instructors of undergraduate psychology can serve a dual role in helping to disseminate accurate information about testing and thereby combating misuses and misinterpretation of scores. First, they can contribute through what they transmit to the students in their own classes. Second, they can serve as resource persons to answer special questions that will surely be directed to them, not only by students in general, but also by colleagues, parents, and members

of the general public. Therefore, undergraduate psychology instructors, whatever their own fields of specialization, need to know somewhat more about the technical properties of tests than would be required for one or two class periods or even for an undergraduate course in psychological testing.

The instructor should be familiar with general sources of information about tests, such as the *Mental Measurements Yearbooks* (MMY). In addition to providing routine information (publisher, price, forms, age levels), these yearbooks include critical reviews by one or more test specialists and a complete bibliography of published references about the tests. The publication of the MMY is now handled by the Buros Institute of Mental Measurements, located at the University of Nebraska.[1] The ninth MMY is in preparation. Beginning in 1983, the test entries of the MMY, together with the critical reviews, are also being included in the online computer service for the Buros Institute database, which is offered through Bibliographic Retrieval Services, Inc. (BRS). Each month, updates are transmitted to this retrieval service as they become available. BRS coverage has been extended back to 1977, with some retrospective coverage for selected entries. This retrieval service can be obtained by communicating with the yearbook editor at the Buros Institute.

Another important reference is the test standards prepared by the American Psychological Association (APA) in collaboration with two other national associations concerned with testing, the American Educational Research Association and the National Council on Measurement in Education. The first edition was published in 1954; subsequent editions appeared in 1966 and 1974. A fourth edition is now in press. These standards provide a succinct but comprehensive summary of recommended practices in test construction, reflecting the current state of knowledge in the field. The later editions give increasing attention to proper test use and to the correct interpretation and application of test results.

Norms and the Interpretation of Test Scores

There are several psychometric features of tests that a test user should know about. First, there is the concept of norms. A common error among the general public, including undergraduates, is to confuse *percentile scores* with traditional *percentage scores*. The difference, of course, is that the former refer to people and the latter to items.

[1]Contact James V. Mitchell, Jr., Editor; Buros Institute of Mental Measurements; University of Nebraska—Lincoln; 135 Bancroft Hall; Lincoln, NE 68588. Telephone: (402) 472-1739.

The latter, moreover, are raw scores, in the sense that they do not carry a built-in interpretation. If a child gets 50 percent correct on an arithmetic test, one cannot judge whether this performance is good, bad, or indifferent. If the difficulty level of the test were altered, the percentage correct that resulted could vary from 0 to 100. By contrast, a 50th percentile score indicates that this child performs like the typical individual of comparable experiential background. This example illustrates why norms are needed in the first place; it is one way to introduce the concept of norms.

Another source of truly prodigious confusion is the IQ. Much of this confusion is associated with the psychological rather than the statistical interpretation of this score, to which I shall return later. In its purely psychometric aspects, the IQ began as a ratio between mental age and chronological age. Although it was used for several decades, this type of score was eventually abandoned because of its many technical deficiencies, including the fact that it did not yield comparable scores at different ages. Moreover, it was applicable only to traits that showed a regular increase with age, and it did not apply at all to adults except through a bizarre sort of logical extrapolation. Thus the ratio IQ was abandoned—well and good! It seems, however, that the term fulfilled a special popular need for magic and for simple answers. Hence was born the deviation IQ, which is nothing more than a standard score with an IQ label attached to it.

Because standard scores in various forms are today the most common type of test score, they need to be explained. This is also a good place to introduce the student to the standard deviation (*SD*); without an acquaintance with the *SD,* one cannot go very far in understanding modern psychology. Here let me confess to a strong bias on my part: I draw a sharp distinction between the mechanics of statistical computation and the understanding of statistical concepts. In my statistics courses, I taught both, but the concepts always preceded the computation. In other courses, I taught the concepts only and let the computation go unless it was needed for a special purpose. For this reason, I always presented the basic formulas, which might or might not be followed by the short-cut formulas with labor-saving devices. It has been my observation that when students are taught the short-cut formulas first, such as computing the *SD* or the Pearson *r* with raw scores and correction term in lieu of deviations, they usually learn to go through the necessary motions and arrive at an answer. But they rarely understand what they are doing, why they do it, or what the answer means.

In teaching about the *SD,* I like to begin with our efforts to describe the performance of a group. We start with the mean, which provides a single number that best characterizes the group as a whole. Then we ask what more we want to know about the group, especially if we want to compare it with another group. Pretty soon, the idea

of the extent of individual differences (that is, variability) emerges. After working our way through the range and its limitations, someone is bound to come up with the idea of finding the difference between each person's score and the mean. Averaging these differences should tell us something. But unfortunately, the plus and minus deviations add up to zero, because that is one of the properties of the mean. So how do we get rid of those minus numbers? It soon becomes clear that the only mathematically defensible way of doing so is to square each deviation and take the square root of their mean. I used to go through the same sort of process in arriving at the basic formula for the Pearson *r,* which incidentally, can be logically derived by recognizing the need for standard scores to express the two variables in comparable units.

So much for statistics. Before I leave the topic of norms, let me consider so-called *criterion-referenced tests* (see Anastasi, 1982, pp. 94–98, 419–420). I say "so-called" because the label is a bit confusing. In psychometrics, the *criterion* generally refers to some independent, external measure that the test is designed to predict, such as a direct measure of job performance used in evaluating an applicant selection test. The criterion-referenced tests, however, can be more accurately described as content-referenced. Typically, such tests use as their interpretive frame of reference a specified *content domain* rather than a specified population of persons. The focus is on what the individual can do and what he or she knows, rather than on how the individual compares with others. Several other terms have been proposed by different writers, such as *domain-referenced* and *objective-referenced.* However, criterion-referenced has remained the most popular term. I shall therefore use it in this discussion, even though it is not the most appropriate term.

When first introduced some twenty years ago (Glaser, 1963), criterion-referenced tests were regarded as an alternative to norm-referenced tests, which represented a fundamentally different approach to testing. Actually, criterion-referenced and normative interpretations can in fact be combined to provide a fuller evaluation of the individual's performance on certain types of tests. Several recently developed tests for assessing mastery of specific subject-matter areas do in fact combine both interpretations. This is illustrated by the Stanford Diagnostic Test in reading and in mathematics and by the Metropolitan Instructional Tests in reading, mathematics, and language (both series published by The Psychological Corporation).

Thus far, criterion-referenced testing has found its major applications in education, especially in individualized instructional systems in which testing is closely integrated with instruction. Administered at different stages, these tests are used to check on prerequisite skills, diagnose learning difficulties, and prescribe subsequent instructional procedures. Criterion-referenced tests have also been used in broad

surveys of educational accomplishment, such as the National Assessment of Educational Progress (Womer, 1970), and in meeting demands for educational accountability (Gronlund, 1974). Testing for the attainment of minimum requirements, as in qualifying for a driver's license or a pilot's license, also utilize essentially criterion-referenced testing. A related application is in testing for job proficiency where the mastery of a small number of clearly defined skills is to be assessed, as in many military occupational specialties (Maier & Hirshfeld, 1978; Swezey & Pearlstein, 1975).

In general, criterion-referenced tests are most suitable for assessing minimum competency in readily identifiable basic skills. A fundamental requirement for the construction of this type of test is a generally accepted domain of knowledge or skills to be assessed. The selected domain must then be subdivided into small units that are defined in highly specific behavioral terms, such as "multiplies three-digit by two-digit numbers" or "identifies the misspelled word in which the final *e* is retained when adding *-ing*." Unless the content domain is itself quite narrowly limited, it is difficult to include a representative content sample in the test. Without such careful specification and control of content, however, the results of criterion-referenced testing could degenerate into an idiosyncratic and uninterpretable jumble. The approach is limited chiefly to testing for basic skills in well-established subject-matter areas. While useful in providing supplementary information in such contexts, criterion-referenced testing does not represent a new approach to testing, and its application presents several technical as well as practical difficulties (Angoff, 1974; Ebel, 1962, 1972).

Reliability and Measurement Error

A second major concept relevant to test evaluation is that of reliability (Anastasi, 1982, chap. 5). In psychometric terminology, reliability means consistency of performance. Specifically, it refers to the consistency of the scores obtained by the same persons when reexamined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions. In its broadest sense, test reliability indicates the extent to which individual differences in test scores are attributable to "true" differences in the characteristics under consideration and the extent to which they are attributable to chance errors. In other words, measures of test reliability make it possible to estimate what proportion of the total variance of test scores is *error variance.* The crux of the matter, however, lies in the definition of error variance. Factors that might be considered error variance for one purpose would be included under true variance for another. For example, if one is interested in

measuring fluctuations of mood, then the day-by-day score changes on a test of cheerfulness–depression would be relevant to the purpose of the test and would thus be part of the true variance of the scores. If, on the other hand, the test is designed to assess more durable personality traits, the daily fluctuations would fall under the heading of error variance.

This distinction is illustrated by some current tests designed to yield separate measures of traits and states, such as the State–Trait Anxiety Inventory developed by Spielberger and his associates (Spielberger, Gorsuch, & Lushene, 1970). Another example is provided by Atkinson's use of an adaptation of the Thematic Apperception Test (TAT) cards in assessing achievement drive (Atkinson, 1981; Atkinson & Birch, 1978, pp. 370–374). Using a computer simulation, Atkinson demonstrated that it is possible to obtain a construct validity as high as .90, with a reliability across cards as low as .07. According to Atkinson's theory of motivation, the individual responds to the successive cards through a continuous stream of activity, which reflects the rise and fall in the relative strength of different behavior tendencies. When a behavior tendency is expressed in activity, its strength is thereby reduced. Consequently, fluctuations in the expression of a particular drive would be expected in the course of taking the test. If responses are aggregated across cards, however, one can identify individual differences in drive strength that correspond to stable traits over time. In this example, reliability over time, as well as validity, would be high, while internal-consistency reliability would be low.

There could, of course, be as many varieties of test reliability as there are conditions affecting test scores. But in standardized test administration, most irrelevant variance is minimized by controlling the testing environment, instructions, time limits, rapport, and other testing conditions. The kinds of reliability measured in actual practice, therefore, are few. The major sources of error variance arise from time sampling and content sampling. Retest reliability provides a measure of stability over time. If alternate forms are administered on different occasions, the error variance arises from both temporal fluctuations and differences between the specific items or content sampled in the two test forms. Split-half reliability, such as commonly found by analyzing scores on odd and even items, measures only reliability across content samples, or the consistency of performance on two sets of comparable items administered at the same time. This reliability is often labeled a measure of internal consistency of the test.

Another measure of reliability frequently reported in test manuals is that developed by Kuder and Richardson (1937). Again utilizing a single administration of a single test form, the K-R reliability is based on the consistency of responses to all items in the test. This *interitem consistency* is influenced by two sources of error variance: (a)

content sampling, as in alternate-form and split-half reliability; and (b) heterogeneity of the behavior domain sampled. The more homogeneous the domain, the higher the interitem consistency. For example, if one test includes only multiplication items, while another comprises addition, subtraction, multiplication, and division items, the former will probably exhibit more interitem consistency than the latter. It can be shown mathematically that the K-R reliability coefficient is actually the mean of all possible split-half coefficients resulting from different splits of a test (Cronbach, 1951).[2] The ordinary split-half coefficient, on the other hand, is based on a planned split designed to yield equivalent sets of items. Hence, unless the test items are highly homogeneous, the K-R coefficient will be lower than the split-half reliability. In fact, the difference between these two coefficients may serve as a rough index of the homogeneity of the test.

The K-R formula is applicable to tests whose items are scored as right or wrong, or according to some other pass-fail system. Some tests, however, may have multiple-scored items. On a personality inventory, for example, the respondent may receive a different numerical score on an item, depending on whether he or she checks "usually," "sometimes," "rarely," or "never." For such tests, a generalized measure is available, known as coefficient alpha (Cronbach, 1951; Novick & Lewis, 1967; for a clear computational layout, see Ebel, 1965, pp. 328–330). Although requiring a little more computational labor, coefficient alpha is interpreted in the same way as the Kuder-Richardson coefficient.

Another source of error variance that has received some attention is *scorer variance.* Most tests provide sufficiently objective and standardized scoring procedures to ensure empirical uniformity of scoring. Certain types of tests, however, leave a good deal to the judgment of the examiner or scorer. This is especially true of clinical instruments employed in intensive individual examinations, as well as projective tests of personality and some creativity tests. For such tests, scorer reliability may be found by having a sample of test papers independently scored by two examiners and correlating the two sets of scores. When several scorers are to be compared, two-way analysis of variance and intraclass correlation may be employed for the same purpose.

The reliability coefficient, however computed, is itself a measure of the percentage (or proportion) of error variance in the test score. This is one case in which a correlation coefficient corresponds directly to a percentage and can be interpreted as such. For example, a reli-

[2]This is strictly true only when the split-half coefficients are found by the Rulon formula, based on the variance of the differences between the two half-scores, not when the coefficients are found by correlation of halves and the Spearman-Brown formula (Novick & Lewis, 1967).

ability coefficient of .85 signifies that 85 percent of the variance in test scores depends on true variance in the trait measured and 15 percent depends on error variance (as operationally defined by the procedures followed). This interpretation frequently causes confusion, because students learned in general statistics that it is the square of a correlation coefficient that represents proportion of common variance. Actually, the proportion of true variance in test scores is the square of the correlation between scores on a single form of the test and true scores, free from chance errors. This correlation, known as the index of reliability,[3] is equal to the square root of the reliability coefficient. Hence, when the index of reliability is squared, the result is the reliability coefficient, which can therefore be interpreted directly as the percentage of true variance.

A particularly useful application of test reliability to the interpretation of an individual's test scores is provided by the standard error of measurement (SEM). So important is the SEM for this purpose that the College Board includes data on the SEM and an explanation of its use, not only in brochures distributed to high school and college counselors and the accompanying slide show, but also in the individual score reports sent to students (College Entrance Examination Board, 1983c, 1984a, 1984b). Given the reliability coefficient of a test (r_{tt}) and the standard deviation (*SD*) of its scores obtained on the same group, the SEM can be computed with a simple formula: $SEM = SD\sqrt{1-r_{tt}}$. When so computed, the SEM is expressed in the same units as the test scores. With it, the probable range of fluctuation of an individual's score that resulted from irrelevant, chance factors can be estimated.

More and more tests today are reporting scores not as a single number, but as a score band within which the individual's true score is likely to fall. These score bands are found by employing the familiar normal curve frequencies to establish specified confidence intervals. Thus a distance of ± 1 SEM from the obtained score yields a probability of roughly 2:1 (or 68:32) that the individual's true score falls within that range. Similarly, the 95 percent confidence interval can be found by taking ± 1.96 SEM; and ± 2.58 SEM gives the 99 percent confidence interval.

The use of such score bands is a safeguard against placing undue emphasis on a single numerical score. It is also useful in interpreting differences between scores. The SEM can be used to compute the statistical significance of score differences between persons on the same test or score differences within the individual on different tests. The latter is especially important in score pattern analysis, as on the

[3]Derivations of the index of reliability, based on two different sets of assumptions, can be found in Gulliksen (1950, chaps. 2 and 3). See also Guilford and Fruchter (1978, pp. 411–412).

Wechsler scales and on multiple aptitude batteries such as the Differential Aptitude Tests (DAT). The test manuals for such tests now regularly include the data that are needed to identify the smallest score difference corresponding to specified significance levels. Users are cautioned against drawing conclusions from differences that fall below these limits and that may therefore indicate no more than chance fluctuations.

Validity: What Does a Test Measure?

And now I come to test validity, which is undoubtedly the most basic and pervasive feature of any test (Anastasi, 1982, chaps. 6–7). The validity of a test concerns what the test measures and how well it does so. It indicates what can be inferred from test results, and it enables users to judge how well the test meets their particular assessment needs. Obviously, validity relates to all three questions that were cited earlier as the essential elements of test evaluation.

Validity in the Test Development Process

Many test manuals are still organized in accordance with the traditional view that validity is one among several technical properties of a test and that it belongs in the later stages of test development. Certainly that is the impression created as one reads most manuals. For several decades, there has also been a widespread belief that there are three distinct kinds or aspects of validity, namely, content, criterion-related, and construct validity (APA et al., 1974). When first introduced in the 1954 *Technical Recommendations* (APA et al.), this tripartite classification brought some order and uniformity into what was then a rather chaotic approach to test validity. Soon, however, there appeared a tendency to reify the three validities, to apply them too rigidly, and to lean on the labels for support. Initially this rigidity was manifested by the acceptance of content validation as applicable to achievement tests; of criterion-related validation as appropriate for personnel selection and classification and for other essentially predictive uses of tests; and of construct validation as relevant principally to theoretically oriented basic research. Construct validation was looked upon with awe and great respect, but usually kept at arm's length by practical test users. More recently, this tripartite rigidity has taken on another, more global form. There seems to be a compulsion, exemplified in several recent test manuals, to do something—anything—that could be classified under each of the three headings, regardless of the nature or purpose of the test. The results

are then reported in three separate, neatly labeled sections. Once the three validities have been ticked off in checklist fashion, there is a relaxed feeling that validation requirements have been met.

In contrast to such rigid approaches, it is now being gradually recognized that validity is built into a test through a wide variety of possible procedures—many more than three. These procedures are employed sequentially, at different stages of test construction (Guion, 1983; Jackson, 1970, 1973). The validation process begins with the definition of the constructs to be assessed. These definitions may be derived from psychological theory, from prior research, or from analyses of real-life behavior domains, as illustrated by functional job analyses or curricular surveys. In psychometric terminology, a construct is a theoretical concept of varying degrees of abstraction or generalizability. It corresponds closely to what is commonly termed a trait. Constructs may be simple and narrowly defined, such as speed of walking or spelling ability, or they may be complex and broadly generalizable, such as abstract reasoning or scholastic aptitude. The formulation of detailed specifications for subtests and test items is guided by available knowledge about the constructs to be included. Items are then written to meet these specifications. Empirical item analyses follow, with the selection of the most effective (i.e., most valid) items from the initial item pool. Other appropriate internal analyses may also be carried out, including factor analyses of item clusters or subtests. The final stage includes validating and cross-validating various scores through statistical analyses against external, real-life criteria.

It should be noted that almost any information gathered in the process of developing or using a test is relevant to its validity. It is relevant in the sense that it contributes to an understanding of what the test measures. Certainly, data on internal consistency and on retest reliability help to define the homogeneity of the construct and its temporal stability. Norms may well provide additional construct specification, especially if they include separate normative data for subgroups classified by age, sex, or other demographic variables that affect test performance. Remember that systematic age increment was a major criterion in the development of early intelligence tests, such as the Stanford-Binet.

Actually, the entire test development process contributes to construct validation, whatever the nature or purpose of the test. If one considers procedures rather than labels, it becomes clear that content analyses and correlations with external criteria fit into particular stages in the process of construct validation, that is, in the process of both determining and demonstrating what a test measures. In a 1980 article, Messick (1980b) argued convincingly that the term validity, insofar as it designates the interpretive meaningfulness of a test, should be reserved for construct validity. He maintained that other

procedures with which the term validity has traditionally been associated should be designated by more specifically descriptive names. Thus, content validity could be called content relevance and content coverage to refer to domain specifications and domain representativeness, respectively. Criterion-related validity could be labeled predictive utility and diagnostic utility to correspond to predictive and diagnostic uses. These changes in terminology would help, but it may be some time before the old terms can be dislodged. In the meantime, we should not be misled by rigid applications of the traditional terminology.

Item Analysis

In my rapid overview of test construction stages, I referred to two statistical procedures that call for a closer look in their own right. They are item analysis and factor analysis. Some acquaintance with the basic concepts and techniques of item analysis (Anastasi, 1982, chap. 8; see also pp. 301–304, 410–413) can help test users in their evaluation of published tests. Items can be examined qualitatively, on the basis of their content and form, and quantitatively, on the basis of their statistical properties. Qualitative analysis concerns chiefly content relevance and content coverage. The necessary information for this evaluation is often given in achievement test manuals, in tables that show the number of items falling into different content categories. Items can also be examined with reference to effective item-writing guidelines (Gronlund, 1977; 1981, chaps. 5–7; Hopkins & Stanley, 1981, chap. 9; Thorndike & Hagen, 1977, chap. 7).

Quantitative techniques of item analysis cover principally statistical assessment of the difficulty and the discriminative value of items. The selection of items through these techniques influences both the reliability and the validity of the finished test. For ability tests, *item difficulty* is defined in terms of the percentage of persons who answer the item correctly. In personality inventories, the percentage of persons who answer the item in the keyed direction serves the same function for statistical analysis. In either case, the percentages may be converted to normal curve sigma-distances. The item values are thereby expressed on a scale of approximately equal units, on the assumption that the trait is normally distributed. Items are chosen to fit the specified level and range of difficulty for the particular testing purpose. For most tests, which are designed to assess each person's performance level with maximum accuracy, the most suitable items are spread over a moderate difficulty range around the 50 percent level.

Item discrimination, the second major feature, refers essentially to the relation between performance on an item and standing on the

trait under consideration. To investigate this relation, the persons who pass a given item and those who fail it may be compared either on an external criterion or on the total test score. In the initial stages of test development, the total test (or subtest) score provides a first approximation to a measure of the relevant trait or construct. The relation is usually expressed as a biserial correlation for each item.

It is apparent that the item statistics I have been considering are restricted to the samples from which they were derived. For several testing purposes, however, what is needed is item information applicable across samples that differ in ability level. In educational achievement tests, for example, it is advantageous to be able to compare a child's score over several grades on a uniform scale. Another example is provided by large-scale testing programs, such as that of the College Board, that need many equivalent test forms to be administered at different times (Donlon, 1984). It would be unfair to individuals to evaluate their scores in terms of their particular sample, insofar as the performance level of samples tested at different times of the year or in different years varies significantly.

Until recently, the standard procedure employed to provide a comparable scale across samples was some variant of Thurstone's (1925, 1947) absolute scaling. What is required for such scaling is the inclusion of a set of common anchor items in the test forms administered to two samples. By computing the mean and *SD* of the difficulty values of these anchor items in each of the two samples, a conversion formula can be worked out for translating all the item values obtained in one group into those of the other. A different set of anchor items can be used to link different pairs of groups. Each new form is linked to one or two earlier forms, which in turn are linked to other forms, through a chain extending back to the group chosen as the fixed reference group. For the College Board SAT, this reference group was the sample of approximately 11,000 candidates tested in 1941 (Donlon, 1984). Scales built in this way—from a fixed reference group—are analogous to scales used in physical measurement, in at least one respect. In measuring distance, for example, the foot is a convenient and uniform unit; we do not know nor care whose foot was originally measured to define this standard.

With the increasing availability of high-speed computers, more precise mathematical procedures have been developed to provide sample-free measurement scales for psychological tests (Baker, 1977; Hambleton & Cook, 1977; Lord, 1980; Weiss & Davison, 1981). These procedures were originally grouped under the general title of latent trait models. The basic measure they use is the probability that a person of specified ability (the so-called latent trait) succeeds on an item of specified difficulty. There is no implication, however, that such latent traits or underlying abilities exist in any physical or physiological sense, nor that they cause behavior. They are statistical con-

structs derived mathematically from empirically observed relations among test responses. A rough, initial estimate of an examinee's ability is the total score obtained on the test. In order to avoid the false impression created by the term, *latent trait,* some of the leading exponents of these procedures have substituted the term *item response theory,* or IRT (Lord, 1980; Weiss & Davison, 1981); this designation is now gaining usage within psychology.

By whatever name they may be called, these procedures utilize three parameters: item discrimination, item difficulty, and a lower-asymptote or "guessing" parameter corresponding to the probability of a correct response occurring by chance. Some simplified procedures, such as the Rasch model (Andersen, 1983; Wright, 1977), use only one parameter, the difficulty level, on the assumption that item differences on the other two parameters are negligible. But this assumption has to be empirically verified for different tests. IRT is gradually being incorporated in large-scale testing programs. For example, beginning in 1982, this model has been adopted for equating scores on the new forms of the SAT, so as to express them on the continuing uniform scale[4] (Donlon, 1984).

One of the most important applications of IRT is to be found in *computerized adaptive testing* (Green, 1983a, 1983b; Lord, 1977; Urry, 1977; Weiss, 1976). Also described as individualized, tailored, and response-contingent testing, this procedure adjusts item coverage to the responses actually given by each examinee. As the individual responds to each item, the computer chooses the next item. If an item is passed, it is followed by a more difficult item; if it is failed, an easier item follows. This will be recognized as the basic testing procedure used in individually administered intelligence tests, such as the Binet. Adaptive testing achieves the same objective in less time, with far greater precision, and without one-to-one administration by a trained examiner.

After each successive item is presented, the computer calculates the examinee's cumulative ability score, together with the error of measurement of that score. Testing is terminated when the error of measurement reaches a preestablished acceptable level. Exploratory research on computerized adaptive testing has been in progress in various contexts. Its operational use is under consideration in several large-scale testing programs in both civilian government agencies and the military services. An example is provided by current efforts to develop a computerized adaptive version of the Armed Services Vocational Aptitude Battery (ASVAB).

[4]However, selection of items for inclusion in any one form still currently follows the earlier procedures (i.e., equated difficulty values in terms of fixed reference group and biserial correlation with total score).

Factor Analysis

Factor analysis has been in use for several decades and is familiar to most psychologists. Hence, I shall merely touch upon some procedural highlights and say something about what factors mean. The principal object of factor analysis is to simplify the description of data by reducing the number of necessary variables or dimensions. For example, beginning with the intercorrelations among 20 tests, it may be possible to show that two or three factors are sufficient to account for nearly all the common variance in the set. These are the types of factors identified in such factor-analytic systems as Thurstone's primary mental abilities and Guilford's structure-of-intellect model. If the data are obtained from a sufficiently heterogeneous population, Spearman's *g* factor may emerge as a second-order factor to account for the correlations found among the factors themselves.

All techniques of factor analysis begin with a complete table of intercorrelations among a set of tests (or other variables), known as a correlation matrix.[5] Every factor analysis ends with a factor matrix, that is, a table showing the weight or loading of each of the factors in each test. It is also customary to represent factors geometrically as reference axes, in terms of which each test can be plotted as a point on a graph. This can easily be done if one works with two factors at a time; the results from successive graphs can then be combined mathematically. It should be noted that the position of the reference axes is not fixed by the data. The original correlation table determines the position of the tests only in relation to each other. The same points can be plotted with the reference axes in any position. For this reason factor analysts frequently rotate the axes until they obtain the most satisfactory and easily interpretable pattern. This is a legitimate procedure, somewhat analogous to measuring longitude from, say, Chicago rather than Greenwich.

Often the object of rotation is to approximate simple structure, that is, to describe each test with the minimum possible number of factors. Most factor patterns employ *orthogonal axes,* which are at right angles to each other. Occasionally, the test clusters are so situated that a better fit can be obtained with *oblique axes.* In such a case, the factors themselves will be correlated. It can be argued that meaningful categories for classifying individual differences need not be uncorrelated. For example, height and weight are highly correlated; yet they have proved to be useful categories in the measurement of physique. When the factors are correlated, the intercorrelations among the factors can themselves be factor analyzed to derive second-order factors. This process has been followed in several studies, with both aptitude and personality variables.

[5]A matrix is any rectangular arrangement of numbers into rows and columns.

There are several different methods for carrying out a factor analysis and for subsequent rotation of axes. The resulting factor matrices may look quite different. This has sometimes created the impression that the findings are arbitrary and artificial; occasionally, it has engendered controversies about the "true" solution. Actually, these factor matrices represent alternative and equally applicable ways of describing the same data. It has been shown that different factor solutions are mathematically interchangeable; they can be transformed one to another by computing the appropriate transformation matrix (Harman, 1976, pp. 338–341).

By whatever statistical procedures such factors are found, and however elaborate such procedures may be, we must bear in mind that the factors, like the test scores from which they were derived, are descriptive and not explanatory. The constructs identified through factor analysis do *not* represent underlying entities, causal factors, or fixed personal characteristics. There is an increasing accumulation of evidence showing the role of experiential background in the formation of factors (Anastasi, 1970, 1983a). It is not only the level of performance in different abilities, but also the way in which performance is organized into distinct traits, that is influenced by the individual's experiential history. Differences in factor patterns have been found to be associated with different cultures or subcultures, socioeconomic levels, and types of school curricula. Changes in factor patterns over time have also been observed. These include long-term changes, which may reflect the cumulative effects of everyday experience, as well as short-term changes resulting from practice and other experimentally controlled learning experiences. Research on animals has also yielded suggestive evidence regarding the experiential production of factors through the control of early experiences.

Validity Generalization

In the practical utilization of validity data, there are two important questions that have received increasing attention in recent years. The first pertains to validity generalization. Test manuals commonly report correlations between test scores and various practical criteria, in order to help the potential user in understanding what a test measures. Although a test user may not be directly concerned with the prediction of any of the specific criteria employed, by examining such criteria he or she is able to build up a concept of the behavior domain sampled by the test. If we follow this thinking a bit further, we can see that all test use and all interpretation of test scores imply construct validity. Because tests are rarely, if ever, used under conditions that are identical to those under which validity data were gathered, some

degree of generalizability is inevitably involved. Thus, the interpretive meaning of test scores and their practical utilization is always based on constructs, which may vary widely in breadth or generalizability with regard to behavior domains, situations, and populations.

When standardized aptitude tests were first correlated with performance on presumably similar jobs in industrial validation studies, the validity coefficients were found to vary widely (Ghiselli, 1959, 1966). Similar variability among validity coefficients was observed when the criteria were grades in various academic courses (Bennett, Seashore, & Wesman, 1984). Such findings led to widespread pessimism regarding the generalizability of test validity across different situations. Until the mid-1970s so-called situational specificity of psychological requirements was generally regarded as a serious limitation in the usefulness of standardized tests in personnel selection[6] (Guion, 1976). In a sophisticated statistical analysis of the problem, however, Schmidt, Hunter, and their associates (Schmidt & Hunter, 1977; Schmidt, Hunter, & Pearlman, 1981) demonstrated that much of the variance among obtained validity coefficients may be a statistical artifact resulting from small sample size, criterion unreliability, and restriction of range in employee samples.

The industrial samples available for test validation are generally too small to yield a stable estimate of the correlation between predictor and criterion. For the same reason, the obtained coefficients may be too low to reach statistical significance in the sample investigated and may thus fail to provide evidence of the test's validity. It has been estimated that about half of the validation samples used in industrial studies include no more than 40 or 50 cases (Schmidt, Hunter, & Urry, 1976). This is also true of the samples often employed in educational settings to compute validity coefficients against grades in particular courses or specialized training programs (Bennett, Seashore, & Wesman, 1984). With such small samples, criterion-related validation is likely to yield inconclusive and uninterpretable results within any single study.

By applying some newly developed techniques to data from many samples drawn from a large number of occupational specialties, Schmidt, Hunter, and their co-workers were able to show that the validity of tests of verbal, numerical, and abstract reasoning aptitudes can be generalized far more widely across occupations than had heretofore been recognized (Pearlman, Schmidt, & Hunter, 1980; Schmidt, Gast-Rosenberg, & Hunter, 1980; Schmidt, Hunter, Pearlman, & Shane, 1979). The variance of validity coefficients typically found in earlier industrial studies proved to be no greater than

[6]Situational specificity has played a different and significant role in both the theory and practice of personality testing (see Anastasi, 1983b; Mischel, 1977, 1979; Mischel & Peake, 1982).

would be expected by chance. This was true even when the particular job functions appeared to be quite dissimilar across jobs. Evidently, the successful performance of a wide variety of occupational tasks depends to a significant degree on a common core of cognitive skills. It would seem that this cluster of cognitive skills and knowledge is broadly predictive of performance in both academic and occupational activities demanded in advanced technological societies.

Differential Validity

The second question of special relevance to the test user concerns differential validity. From a practical standpoint, we should bear in mind that the term *differential validity* is currently used in two different senses, one referring to different criteria, the other to different populations. When the term is used in the first sense, the goal of good test usage is to maximize differential validity; when it is used in the second sense, the goal is to minimize it.

Classification Decisions

Differential validity against separate criteria is a major consideration when tests are used for classification purposes, as contrasted with selection. In selection decisions, each individual is either accepted or rejected, as in admitting students to college or hiring job applicants. In classification decisions, no one is rejected or eliminated from the program. Rather, all are assigned to appropriate treatment, so as to maximize the effectiveness of outcomes. One example of classification decisions is the assignment of individuals from an available personnel pool to training programs for occupational specialties. Another example is the counseling of students regarding field of concentration and career choice.

Ideally, a classification battery should include tests that yield very different validity coefficients with different criteria. The object of such a battery is to predict the *differences* between each person's performance in two or more jobs or other criterion situations. In the use of batteries for occupational classification, we need to identify the major constructs covered by the tests, on the one hand, and those covered by the job functions, on the other. The procedures used for this purpose can be illustrated by factor analysis of the tests and by job analysis expressed in terms of critical behavioral requirements. Validity generalization can then be investigated within functional job families, that is, groups of jobs that share major behavioral constructs, regardless of superficial task differences.

Such dual analyses of tests and jobs have been applied with promising results in recent research on the validity of the General Aptitude Test Battery (GATB) for some 12,000 jobs described in the *Dictionary of Occupational Titles* of the U.S. Employment Service (U.S. Department of Labor, 1983a, 1983b, 1983c). For this analysis, the jobs were classified into five functional job families. Factor analyses of the test battery yielded three broad group factors, identified as cognitive, perceptual, and psychomotor abilities. A meta-analysis of data from over 500 U.S. Employment Service (USES) validation studies was then conducted with the newly developed validity generalization techniques. This procedure yielded estimated validities of the appropriate composites for all jobs within each job family.

A more narrowly focused demonstration of the dual identification of behavioral constructs in tests and criteria also utilized USES data (Gutenberg, Arvey, Osburn, & Jeanneret, 1983). In this study, the job analysis dimensions pertaining to decision making and information processing correlated positively with the validities of the cognitive GATB tests (general, verbal, and numerical aptitudes); and they correlated negatively with the validities of psychomotor tests (finger and manual dexterity). In other words, the more a job called for decision making and information processing, the higher was the correlation of job performance with the cognitive tests and the lower was its correlation with the psychomotor tests. There is evidence that these results, as well as those of the previously cited, more broadly oriented studies, reflect an underlying dimension of *job complexity*. This dimension seems to be a major determinant of the differential validity of tests for predicting job performance (U.S. Department of Labor, 1983c).

Test Bias

Differential validity with regard to populations (rather than criteria) is a major concept in discussions of test bias. One question asked in this connection is whether the test may be valid for the majority group and not valid for a minority group. This is sometimes called single-group validity. In accordance with this usage the term *differential validity* is reserved for situations where both groups yield statistically significant validity coefficients, but one is significantly higher than the other. Empirical research with ability tests administered to black and white samples in the United States has failed to support either of these hypotheses about group differences in test validity. A comprehensive meta-analysis covering 39 industrial studies demonstrated that the discrepancies in validity coefficients between blacks and whites did not exceed chance expectancy (Hunter, Schmidt, & Hunter, 1979). It could be argued that, because of inadequate sample

sizes and other methodological limitations, these results are merely inconclusive. It is noteworthy, however, that no evidence of differential validity was found in well-designed, large-scale studies of industrial samples (Campbell, Crooks, Mahoney, & Rock, 1973) or of army personnel (Maier & Fuchs, 1973). In general, the methodologically sounder studies proved to be those less likely to find differential validity. Similar results have been obtained in numerous investigations of black and white college students (Breland, 1979). Validity coefficients of the SAT and other college admission tests for black students were generally as high as those obtained for white students, or higher. At a very different educational level, the same results were obtained with large samples of black and white first-grade schoolchildren (Mitchell, 1967). When two educational readiness tests were correlated with end-of-year grades, the validities of total scores and of subtests were closely similar for the two ethnic groups, although tending to run somewhat higher for the blacks.

Even when a test is equally valid for two groups, however, it is possible that criterion performance is underpredicted for one group and overpredicted for the other. This point can be more readily understood if we visualize the relation between test and criterion by means of the familiar scatter diagram or bivariate distribution.[7] If test scores are represented along the horizontal axis (X) and criterion measures along the vertical axis (Y), each individual can be plotted by a point showing his or her standing in both variables. The straight line fitted to these points is the regression line for the particular group. If both variables have been expressed as basic standard scores, the slope of the regression line is exactly equal to the Pearson r between test and criterion. Hence, any difference between the validity coefficients obtained for two groups is known as *slope bias.*

The regression lines for two groups may, however, have the same slope—they may be parallel lines—and yet they may intersect the Y-axis at different points. These points are the Y-intercepts of the two lines; hence such group differences are designated *intercept bias.* Parenthetically, the terms *test bias* and *test fairness* are often used to refer specifically to intercept bias. By whatever name, intercept bias means that the identical test score would correspond to different criterion scores if predicted from the separate regression lines of the two groups; or conversely, different test scores would predict the same criterion performance when obtained by a person in one or the other group. Suppose, for example, that the majority excels on the test, but majority and minority perform equally well on the criterion. Selecting all applicants in terms of a test cutoff established for the majority

[7]Technical discussions of this *regression model,* in relation to other models of test bias, can be found in Cleary (1968); Gross and Su (1975); Gulliksen and Wilks (1950); Humphreys (1952); Hunter, Schmidt, and Rauschenberger (1984); and Petersen and Novick (1976). See also Anastasi (1982, pp. 183–191).

group would thus discriminate unfairly against the minority. Under these conditions, use of the majority regression line for both groups *underpredicts* the criterion performance of minority group members. This situation is most likely to occur when a large proportion of test variance is irrelevant to criterion performance and measures functions in which the majority excels the minority. Systematic comparison of major behavioral constructs in both test and job performance provides a safeguard against choosing such a test.

If, however, the two groups differ in a third variable that correlates positively with both test and criterion, then the test will *overpredict* the performance of minority group members (Linn & Werts, 1971; Reilly, 1973). Under these conditions, the use of the same cutoff score for both groups favors the minority. The findings of empirical studies do in fact support this expectation. Well-controlled studies with tests in current use have found either no significant differences or, more often, a tendency to overpredict the criterion performance of minority groups and hence to favor the members of minority groups in selection decisions. Such results have been obtained in the prediction of college grades (Breland, 1979), law school grades (Linn, 1975), performance in military training programs (Maier & Fuchs, 1973; Shore & Marion, 1972), and a variety of industrial criteria (Campbell et al., 1973; Gael, Grant, & Ritchie, 1975a, 1975b; Grant & Bray, 1970; Hunter, Schmidt, & Rauschenberger, 1984).

Psychological Interpretation of Test Scores

It is apparent that the test user should be familiar with current developments in the statistical concepts and methodology employed in test construction. Such knowledge is essential in understanding the information provided in test manuals and related publications. And this information is the basis for both choosing an appropriate test and correctly interpreting the test scores. At the same time, the test user needs some knowledge of current developments in psychology. Common misuses and misinterpretations of tests often arise from misconceptions, not about the statistics of testing, but about the behavior the tests are designed to measure (Anastasi, 1967). I shall cite a few outstanding examples from ability testing, where current confusions and controversies are most conspicuous.

Aptitude and Achievement Tests

Let us consider the traditional distinction between aptitude and achievement tests (Anastasi, 1984). Aptitudes are typically defined

more precisely than is intelligence, and they refer to more narrowly limited cognitive domains. Nevertheless, like intelligence, they have traditionally been contrasted with achievement in testing terminology. This contrast dates from the early days of testing, when it was widely assumed that achievement tests measured the effects of learning, whereas intelligence and aptitude tests measured so-called innate capacity, or potentiality, independently of learning. This approach to testing in turn reflected a simplistic conception of the operation of heredity and environment that prevailed in the 1920s and 1930s (Cravens, 1978; see also Anastasi, 1979).

These early beliefs led over the years to strange uses of tests, including attempts to compute some index, such as a ratio or a difference-score, that allegedly separated the influence of heredity from that of environment in the individual's performance. Thence arose such byproducts as the now defunct achievement quotient, as well as the still extant classification into underachievers and overachievers. Despite repeated efforts by testing specialists and psychological researchers to dispel the early misconceptions about aptitude and achievement tests, these misconceptions have proved highly viable among the general public and especially among some test users and test critics. One possible explanation for the survival of these beliefs is to be found in the popular desire for magic—the desire for easy answers, quick solutions, and shortcuts.

Actually, all cognitive tests, whatever they may be called, show what the individual is able to do at the time; they do not explain why the individual performs as he or she does. Both aptitude and achievement tests can be best described as tests of *developed abilities,* a term that is appearing increasingly often in current testing literature. Aptitude and achievement tests can be ordered along a continuum of developed abilities. Those near the center are so similar as to be nearly indistinguishable. As we approach the extreme positions, we can identify two major differences.

The first difference between aptitude and achievement tests pertains to *test use.* Traditional achievement tests are designed and used primarily to assess current status; traditional aptitude tests are designed and used to predict future performance. What can the individual learn—how far and how fast can he or she progress—when put through a particular course of study, educational program, industrial apprenticeship, or other systematic learning experience? At this point, the reader may be thinking that traditional achievement tests, too, can often serve as effective predictors of future learning. True! An achievement test in arithmetic is a good predictor of subsequent performance in an algebra class. And it must also be remembered that all tests show only what the individual can do at the time. How, then, can aptitude tests predict future progress? They can do so only by assessing the prerequisite skills and knowledge needed in order to advance toward the desired performance goal.

The second difference is to be found in the degree of *experiential specificity* underlying the construction of aptitude and achievement tests. Achievement tests are typically designed to reflect closely what the individual has learned within a clearly and narrowly defined knowledge domain; they are closely tied to a specified set of prior learning experiences. Obvious examples are a test in solid geometry, or medieval history, or motor vehicle operation. At the opposite extreme are tests like the Stanford-Binet, which specifies little about the experiential pool beyond growing up in the twentieth-century American culture.

Intelligence Tests

This brings me to the subject of intelligence tests, the most widely misunderstood of all aptitude tests. The scores from these tests, whether reported as IQs or by some more innocuous term, have been so commonly misinterpreted that group intelligence tests were banned from schools in several parts of the country. Individual tests like the Binet and Wechsler scales were usually retained, on the assumption that they were used by trained clinical psychologists and hence were less subject to misinterpretation—an assumption that may not always hold. Another attempted solution was the elimination of the terms *intelligence* and *IQ* from nearly all recently published or revised tests. These precautions may have helped, but they do not go very far in the face of persistent demands for shortcuts and quick answers.

If properly interpreted, however, such tests can serve important functions in many practical contexts. Let us consider what traditional intelligence tests actually measure (Anastasi, 1983c). First, like all cognitive tests, they can show only what the individual knows and can do at the time. Intelligence tests are descriptive, not explanatory. They do not reveal the causes of individual differences in performance. To investigate such causes, we need additional data from other sources such as the individual's experiential history. Second, the tests do not measure all of human intelligence. There are many kinds of intelligence. Each culture demands, fosters, and rewards a different set of abilities, which constitute intelligence within that culture. Research in cross-cultural psychology provides a rich store of examples to illustrate this fact (Berry, 1972; Goodnow, 1976; Neisser, 1976, 1979). A frequent cultural difference pertains to the emphasis placed on generalization and abstract thinking, and the extent to which behavior is linked to specific contexts.

It is apparent that the term *intelligence* is too broad to designate available intelligence tests. They can be more accurately described as measures of academic intelligence or scholastic aptitude. They measure a kind of intelligent behavior that is both developed by formal

schooling and required for progress within the academic system. To help define the construct measured by these tests, there is a vast accumulation of data derived from both clinical observations and validation studies against academic and occupational criteria. The findings indicate that the particular combination of cognitive skills and knowledge sampled by these tests plays a significant part in much of what goes on in modern, technologically advanced societies. In the interpretation of intelligence test scores, the concept of a segment of intelligence, albeit a broadly applicable and widely demanded segment, is replacing that of a general, universal human intelligence.

Coaching and Test Validity

A question that has aroused considerable practical interest is that of coaching and test performance (Anastasi, 1981). Of particular concern is the possible effect of intensive coaching on such tests as the College Board SAT. On the one side, there are the rather extreme claims made by some commercial coaching schools. On the other, there is some well-designed research on this question, some of it conducted under the auspices of the College Board (Messick, 1980a; Messick & Jungeblut, 1981). In general, such research indicates that intensive drill on items similar to those on the SAT is unlikely to produce score gains appreciably larger than those found when students are retested with the SAT after a year of regular high school instruction. Any generalization about coaching as a whole, however, is likely to be misleading because of differences in the nature of the tests, the prior experiences of the samples examined, and the kind and duration of training provided in the coaching program. There are also methodological flaws that make the findings of several studies uninterpretable.

The basic question is not how far test scores can be improved by special training, but how such improvement relates to intellectual behavior in real-life contexts. To answer this question, we must differentiate among three approaches to improving test performance and consider how they affect the predictive validity of the test. The first is coaching, narrowly defined as intensive, short-term, massed drill, or "cramming," on items similar to those in the test. Insofar as such coaching raises test scores, it is likely to do so without corresponding improvement in criterion behavior. Hence, it thereby reduces test validity. It should be added that well-constructed tests employ item types shown to be least susceptible to such drill (Donlon, 1984; Evans & Pike, 1973).

A second approach, illustrated by the College Board (1983b) booklet, *Taking the SAT,* is designed to provide test-taking orientation and thereby minimize individual differences in prior test-taking ex-

perience. These differences represent conditions that affect test scores as such, without necessarily being reflected in the broader behavior domain to be assessed. Hence these test-orientation procedures should make the test a more valid instrument by reducing the influence of test-specific variance. Finally, training in broadly applicable cognitive skills, if effective, should improve the trainee's ability to cope with subsequent intellectual tasks. This improvement will and should be manifested in test performance. Insofar as both test scores and criterion performance are improved, this kind of training leaves test validity unchanged, while enhancing the individual's chances of attaining desired goals. Such broadly oriented training is receiving increasing attention today (Anastasi, 1983c). It reflects the growing recognition that the nature and extent of intellectual development depend on one's learning history.

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1954). *Technical recommendations for psychological tests and diagnostic techniques.* Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests.* Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (in press). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Anastasi, A. (1967). Psychology, psychologists, and psychological testing. *American Psychologist, 22,* 297–306.

Anastasi, A. (1970). On the formation of psychological traits. *American Psychologist, 25,* 899–910.

Anastasi, A. (1979). A historian's view of the nature–nurture controversy [Review of *The triumph of evolution: American scientists and the heredity–environment controversy, 1900–1941*]. *Contemporary Psychology, 24,* 622–623.

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist, 36,* 1086–1093.

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.

Anastasi, A. (1983a). Evolving trait concepts. *American Psychologist, 38,* 175–184.

Anastasi, A. (1983b). Traits, states, and situations: A comprehensive view. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measure-*

ment: A festschrift for Frederic M. Lord (pp. 345–356). Hillsdale. NJ: Erlbaum.

Anastasi, A. (1983c). What do intelligence tests measure? In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing: Intelligence, performance standards, test anxiety, and latent traits* (pp. 5–28). San Francisco: Jossey-Bass.

Anastasi, A. (1984). Aptitude and achievement tests: The curious case of the indestructible strawperson. In B. S. Plake (Ed.), *Social and technical issues in testing: Implications for test construction and usage* (pp. 129–140). Hillsdale, NJ: Erlbaum.

Andersen, E. B. (1983). Analyzing data using the Rasch model. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing: Intelligence, performance standards, test anxiety, and latent traits* (pp. 193–223). San Francisco: Jossey-Bass.

Angoff, W. H. (1974). Criterion-referencing, norm-referencing, and the SAT. *College Board Review, 92,* 3–5,21.

Atkinson, J. W. (1981). Studying personality in the context of an advanced motivational psychology. *American Psychologist, 36,* 117–128.

Atkinson, J. W., & Birch, D. (1978). *An introduction to motivation* (2nd ed.). New York: Van Nostrand Reinhold.

Baker, F. B. (1977). Advances in item analysis. *Review of Educational Research, 47,* 151–178.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1984). *Differential aptitude tests: Technical supplement.* Cleveland, OH: Psychological Corporation.

Berry, J. W. (1972). Radical cultural relativism and the concept of intelligence. In L. J. Cronbach & P. J. D. Drenth (Eds.), *Mental tests and cultural adaptations* (pp. 77–88). The Hague: Mouton.

Breland, H. M. (1979). *Population validity and college entrance measures* (College Board Research Monograph No. 8). New York: College Entrance Examination Board.

Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. (1973). *An investigation of sources of bias in the prediction of job performance.* Princeton, NJ: Educational Testing Service.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

College Entrance Examination Board (1983a). *The SAT: About taking the Scholastic Aptitude Test.* New York: Author. (Slide show to accompany 1983b)

College Entrance Examination Board (1983b). *Taking the SAT: A guide to the Scholastic Aptitude Test and the Test of Standard Written English.* New York: Author.

College Entrance Examination Board (1983c). *200–800: What does it all mean? How to interpret SAT and achievement test scores.* New York: Author. (Slide show to accompany 1984b)

College Entrance Examination Board (1984a). *ATP guide for high schools and colleges 1984–85.* New York: Author.

College Entrance Examination Board (1984b). *Your score report 1984–85.* New York: Author.

Cravens, H. (1978). *The triumph of evolution: American scientists and the heredity–environment controversy, 1900–1941.* Philadelphia: University of Pennsylvania Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Donlon, T. F. (Ed.) (1984). *The technical handbook for the College Board Scholastic Aptitude Test and achievement tests.* New York: College Entrance Examination Board.

Ebel, R. L. (1962). Content standard test scores. *Educational and Psychological Measurement, 22,* 15–25.

Ebel, R. L. (1965). *Measuring educational achievement.* Englewood Cliffs, NJ: Prentice-Hall.

Ebel, R. L. (1972). Some limitations of criterion-referenced measurement. In G. H. Bracht, K. D. Hopkins, & J. C. Stanley (Eds.), *Perspectives in educational and psychological measurement* (pp. 144–149). Englewood Cliffs, NJ: Prentice-Hall.

Evans, F. R., & Pike, L. W. (1973). The effects of instruction for three mathematics item formats. *Journal of Educational Measurement, 10,* 257–272.

Gael, S., Grant, D. L., & Ritchie, R. J. (1975a). Employment test validation for minority and nonminority clerks with work sample criteria. *Journal of Applied Psychology, 60,* 420–426.

Gael, S., Grant, D. L., & Ritchie, R. J. (1975b). Employment test validation for minority and nonminority telephone operators. *Journal of Applied Psychology, 60,* 411–419.

Ghiselli, E. E. (1959). The generalization of validity. *Personnel Psychology, 12,* 397–402.

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests.* New York: Wiley.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18,* 519–522.

Goodnow, J. J. (1976). The nature of intelligent behavior: Questions raised by cross-cultural studies. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 169–188). Hillsdale, NJ: Erlbaum.

Grant, D. L., & Bray, D. W. (1970). Validation of employment tests for telephone company installation and repair occupations. *Journal of Applied Psychology, 54,* 7–14.

Green, B. F. (1983a). Adaptive testing by computer. In R. B. Ekstrom (Ed.), *Measurement, technology, and individuality in education* (pp. 5–12). San Francisco: Jossey-Bass.

Green, B. F. (1983b). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A festschrift for Frederic M. Lord* (pp. 69–80). Hillsdale, NJ: Erlbaum.

Gronlund, N. E. (1974). *Determining accountability for classroom instruction.* New York: Macmillan.

Gronlund, N. E. (1977). *Constructing achievement tests* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Gronlund, N. E. (1981). *Measurement and evaluation in teaching* (4th ed.). New York: Macmillan.

Gross, A. L., & Su, W. H. (1975). Defining a "fair" or "unbiased" selection model: A question of utilities. *Journal of Applied Psychology, 60,* 345–351.

Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education* (5th ed.). New York: McGraw-Hill.

Guion, R. M. (1976). Recruiting, selection, and job placement. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 777–828). Chicago: Rand McNally.

Guion, R. M. (1983, April). Disunity in the trinitarian concept of validity. In P. Sandifer (Chair), *Clearing away the cobwebs: A closer look at content validity.* Symposium conducted at the meeting of the American Educational Research Association, Montreal.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Gulliksen, H., & Wilks, S. S. (1950). Regression tests for several samples. *Psychometrika, 15,* 91–114.

Gutenberg, R. L., Arvey, R. D., Osburn, H. G., & Jeanneret, R. R. (1983). Moderating effects of decision-making/information-processing job dimensions on test validities. *Journal of Applied Psychology, 68,* 602–608.

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement, 14,* 75–96.

Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.

Hopkins, K. D., & Stanley, J. C. (1981). *Educational and psychological measurement and evaluation* (6th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Humphreys, L. G. (1952). Individual differences. *Annual Review of Psychology, 3,* 131–150.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86,* 721–735.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. (1984). Methodological, statistical, and ethical issues in the study of bias in psychological tests. In C. E. Reynolds (Ed.), *Perspectives on bias in mental testing* (pp. 41–99). New York: Plenum.

Jackson, D. N. (1970). A sequential system for personality scale development. In C. D. Spielberger (Ed.), *Current topics in clinical and community psychology* (vol. 2, pp. 61–96). New York: Academic Press.

Jackson, D. N. (1973). Structured personality assessment. In B. B. Wolman (Ed.), *Handbook of general psychology* (pp. 775–792). Englewood Cliffs, NJ: Prentice-Hall.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2,* 151–160.

Linn, R. L. (1975). Test bias and the prediction of grades in law school. *Journal of Legal Education, 27,* 293–323.

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8,* 1–4.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1,* 95–100.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Maier, M. H., & Fuchs, E. F. (1973). *Effectiveness of selection and classification testing* (Res. Rep. 1179). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Maier, M. H., & Hirshfeld, S. F. (1978). *Criterion-referenced job proficiency testing: A large scale application* (Res. Rep. 1193). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Messick, S. (1980a). *The effectiveness of coaching for the SAT: Review and reanalysis of research from the fifties to the FTC.* Princeton, NJ: Educational Testing Service.

Messick, S. (1980b). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012–1027.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89,* 191–216.

Mischel, W. (1977). On the future of personality measurement. *American Psychologist, 32,* 246–254.

Mischel, W. (1979). On the interface of cognition and personality: Beyond the person–situation debate. *American Psychologist, 34,* 740–754.

Mischel, W., & Peake, P. K. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review, 89,* 730–755.

Mitchell, B. C. (1967). Predictive validity of the Metropolitan Readiness Tests and the Murphy-Durrell Reading Readiness Analysis for white and Negro pupils. *Educational and Psychological Measurement, 27,* 1047–1054.

Neisser, U. (1976). General, academic, and artificial intelligence. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 135–144). Hillsdale, NJ: Erlbaum.

Neisser, U. (1979). The concept of intelligence. *Intelligence, 3,* 217–227.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32,* 1–13.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65,* 373–406.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13,* 3–29.

Reilly, R. R. (1973). A note on minority group test bias studies. *Psychological Bulletin, 80,* 130–132.

Schmidt, F. L., Gast-Rosenberg, L., & Hunter, J. E. (1980). Validity generalization results for computer programmers. *Journal of Applied Psychology, 65,* 643–661.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. *Journal of Applied Psychology, 66,* 166–185.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization model. *Personnel Psychology, 32,* 257–281.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61,* 473–485.

Shore, C. W., & Marion, R. (1972). *Suitability of using common selection test standards for Negro and white airmen* (AFHRL-TR-72-53). Lackland Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *STAI manual for the State–Trait Anxiety Inventory.* Palo Alto, CA: Consulting Psychologists Press.

Swezey, R. W., & Pearlstein, R. B. (1975). *Guidebook for developing criterion-referenced tests.* Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Thorndike, R. L., & Hagen, E. (1977). *Measurement and evaluation in psychology and education* (4th ed.). New York: Wiley.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16,* 433–451.

Thurstone, L. L. (1947). The calibration of test items. *American Psychologist, 2,* 103–104.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14,* 181–196.

U.S. Department of Labor, Employment and Training Administration (1983a). *The dimensionality of the General Aptitude Test Battery (GATB) and the dominance of general factors over specific factors in the prediction of job performance* (USES Test Res. Rep. No. 44). Washington, DC: U.S. Government Printing Office.

U.S. Department of Labor, Employment and Training Administration (1983b). *Overview of validity generalization* (USES Test Res. Rep. No. 43). Washington, DC: U.S. Government Printing Office.

U.S. Department of Labor, Employment and Training Administration (1983c). *Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery* (USES Test Res. Rep. No. 45). Washington, DC: U.S. Government Printing Office.

Weiss, D. J. (1976). *Computerized ability testing, 1972–1975* (Final Report of Project NR 150-343). Minneapolis: Psychometric Methods Program, Department of Psychology, University of Minnesota.

Weiss, D. J., & Davison, M. L. (1981). Test theory and methods. *Annual Review of Psychology, 32,* 629–658.

Womer, F. B. (1970). *What is National Assessment?* Ann Arbor, MI: National Assessment of Educational Progress.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97–116.