Tests and Assessment

What Counselors Should Know About the Use and Interpretation of Psychological Tests

ANNE ANASTASI

The effective use of tests in counseling depends on the choice of appropriate tests for the particular individual and problem under consideration, as well as the proper interpretation of test scores and their integration within the total assessment process. The knowledge needed for these purposes is discussed with reference to, first, statistical and technical knowledge about tests, and second, substantive psychological knowledge about the behavior domain that the tests assess. Because of the rapid accumulation of knowledge in both areas, counselors need to keep abreast of current developments through continuing efforts to update their knowledge.

Counselors constitute a major group of test users. Moreover, because of the wide diversity of both counseling services and client populations served by different types of counselors, the tests used cover almost the entire range of available instruments. The relevant tests sample virtually the whole life span; they include measures of many aspects of both cognitive and affective behavior; and they represent all testing techniques, from projective and other clinical devices to self-administered inventories and computerized testing. I am starting with the assumption that most counselors already know about the basic requirements of proper test use, such as uniform administration and scoring procedures, appropriate use of norms, maintaining the security of test materials and observing copyright restrictions, and protecting the privacy of test takers and the confidentiality of the findings. Hence, I shall concentrate on a few key points that all too often receive inadequate attention from test users. Test misuses of this latter type are more subtle and less easily recognized than are the more obvious procedural errors, but they have potentially more serious consequences.

THE ROLE OF THE TEST USER

The test user, as distinguished from the test author or publisher, is anyone who has the responsibility for choosing tests, for monitoring test administration and scoring (by persons or computers), and for interpreting test scores and using them as one source of information in making practical decisions. Any one test user may be responsible for one or more of these functions—and often for all of them. If a test user as herein defined lacks adequate background for performing the required functions, he or she should have ready access to a properly qualified supervisor or consultant.

A conspicuous development in psychological testing during the 1980s and 1990s is the increasing recognition of the key role of the test user. Most popular criticisms of tests are clearly identifiable as criticisms of test use (or misuse), rather than criticisms of the tests themselves. Tests are essentially tools. Whether any tool is an instrument of good or harm depends on how the tool is used.

The growing concern with the test user is illustrated by the activities of broadly based national committees (see, e.g., Anastasi, in press; Eyde, Moreland, Robertson, Primoff, & Most, 1988, in press) and by the expanded coverage of test usage in the latest revision of the *Standards for Educational and Psychological Testing*[1] (1985). In this edition of the *Standards*, 11 of the 16 chapters are devoted to the *use* of tests in different professional applications and with special populations, as well as to general standards for proper test administration, scoring, and reporting, and for protecting the rights of test takers.

Why are tests misused? One reason is the all too human desire for shortcuts, quick solutions, and clear-cut answers to our questions. This common human weakness has been capitalized by soothsayers over the centuries, from phrenologists to astrologers and other self-styled expert advisers. People seeking guidance are often attracted by the facile promises of charlatans, in contrast to the slower, deliberate considerations and the carefully qualified suggestions of the scientifically trained professional. Similarly, if one or two short tests—whatever their technical limitations and defects—seem to offer a simple answer to questions about career choice, interpersonal difficulties, emotional problems, or learning deficiencies, many test takers will be temporarily satisfied.

At another level, some misuse of tests by a counselor or other test user may arise from time pressure or work overload, which renders shortcuts attractive. To some extent, too, there may be the tendency—deliberate or inadvertent—to shift decision-making

responsibility to an impersonal agent such as a test. For the responsible professional, awareness of these common human reactions is the best protection against them.

A second major reason for the misuse of tests is inadequate or outdated knowledge about testing. This I believe to be the most frequent cause of test misuse, a conclusion that is supported by surveys conducted in the United States and elsewhere (Eyde et al., 1988, in press; Tyler & Miller, 1986). It is also the source of misuse that can be most directly affected by more and better training programs for test users. Accordingly, it is with this source of test misuse that the rest of my article is concerned.

What specialized knowledge do test users need? For the proper interpretation and application of test results, they need some basic understanding of (a) statistical techniques of psychometrics and (b) relevant facts and principles of behavioral science. The first has been called "psychometric literacy" (Lambert, 1989, August). It is essential, but it is not enough. We need to distinguish between the technical properties of a test and the substantive interpretation of a test score. The latter requires some knowledge about the behavior domain assessed by the test and the conditions that affect behavior development.

STATISTICAL AND TECHNICAL KNOWLEDGE ABOUT TESTS

Nature of Individual Assessment

Counselors typically use tests as one source of information in individual assessment. The term *assessment* is being used increasingly to refer to the intensive study of an individual, leading to recommendations for action in solving a practical problem. The effectively functioning counselor engages in a continuing cycle of hypothesis formulation and hypothesis testing about the particular individual. Each item of information—whether it is an event recorded in the case history, a comment by the client, or a test score—suggests a hypothesis about the person, which will be either confirmed or refuted as other facts are gathered. Such hypotheses themselves indicate the direction of further lines of inquiry, such as choice of tests or follow-up questioning.

One should keep in mind that even highly reliable tests with well-established validity do not yield sufficiently precise results for individual assessment. Hence, counselors (as well as clinical psychologists) tend to be receptive to some instruments that, while psychometrically crude, may nevertheless provide a rich harvest of leads for further exploration. The ultimate responsibility for integrating the information and using it in individual assessment and decision making rests with the counselor. Such responsibility, however, entails certain assessment hazards that the counselor must guard against.

Common Assessment Hazards

These hazards lead to frequent misuses of tests. The first is the *hazard of the single score* on any particular test. This test misuse arises from ignoring the chance variation in the score shown by the standard error of measurement (SEM). It involves using a single number to represent the individual's test performance, instead of computing a score band at a specified confidence level. This practice fails to allow for the random sampling fluctuations among sets of items and over short periods of time that are measured by traditional measures of test reliability. Such a misuse of test results may be compounded when evaluating a pattern of performance on a multiscore battery, unless all relevant statistical pitfalls are considered (see Anastasi, 1985a).

The second is the *hazard of the single time period.* Apart from the previously mentioned random errors, systematic and progressive changes occur over longer time periods. These are the changes that are likely to render old scores in a person's file untrustworthy. Such changes also argue against labeling an individual with a numerical score, rather than restricting the score to that person's performance on a specified test at a specified time. These improper practices assume trait stability regardless of intervening experiences. Periodic reevaluation is needed when a child is considered for educational placement, an employee for promotion or transfer, or a client for counseling or psychotherapy. There is evidence that, in such instances, it is the criterion performance itself, rather than just the test score, that is likely to change progressively over time (e.g., Henry & Hulin, 1987).

The third is the *hazard of the single indicator.* Failure to consider the moderate size of the correlations between different indicators of a behavioral construct may lead to undue reliance on scores from a single test, without corroborative and qualifying data from other tests or from other sources of information about the person. Examples of common constructs assessed in counseling include intelligence, scholastic aptitude, learning ability, attitude toward math, motivation for school learning, and interpersonal relations. It is well to ask what proportion of the variance of the behavioral construct under consideration is covered by any one test. Putting the question in this form highlights the limitations of a single indicator, while recognizing its demonstrated contribution to assessment. An easy way to estimate this proportion under ordinary conditions is to square the correlation between test score and criterion measure.

The fourth is the *hazard of illusory precision.* The availability of numerical scores from instruments designed chiefly as aids for the skilled practitioner (such as projective techniques) may create a misleading impression of quantification and objectivity. The interpretive pitfalls inherent in such instruments are not limited to misuse by inadequately qualified test users; they also occur in the rapidly proliferating computerized scoring systems that provide narrative interpretations of performance. These systems must demonstrate the reliability and validity of their score interpretations through the publication of adequate supporting data. Moreover, they can serve only as aids to the trained practitioner, not as a substitute for the practitioner. The special problems presented by commercially marketed computerized systems of narrative score interpretation have aroused widespread concern on the part of psychological practitioners, committees on testing, and national professional associations. Serious attention is being given to the formulation of workable guidelines for the effective use of such interpretive services (Butcher, 1987; *Guidelines*, 1986).

Evolving Approaches to Test Validation

Some recent developments in psychological testing reflect trends discernible in American psychology as a whole. Conspicuous among these trends is an increasing concern with theory and a movement away from the blind empiricism of earlier decades. This theoretical orientation is illustrated by the growing emphasis on constructs in the description of ability and personality, as well as the increasing use of construct validation in the development and evaluation of tests. The term *construct validity* was introduced into the psychometric vocabulary in the first edition

of the testing *Standards* (*Technical Recommendations*, 1954). The discussions of construct validation that followed—and that continue with undiminished vigor—have served to make the implications of its procedures more explicit and to provide a systematic rationale for their use. In psychometric terminology, a *construct* is a theoretical concept closely akin to 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 mathematical reasoning, scholastic aptitude, neuroticism, or anxiety.

The overemphasis on purely empirical procedures during the early decades of the 20th century arose in part as a revolt against the armchair theorizing that all too often served as the basis for some so-called psychological writings of that period. But empiricism need not be blind, nor does theory need to be subjective speculation. Psychologists gradually realized that theory *can* be derived from an analysis of accumulated research findings, and it *can* in turn lead to the formulation of empirically testable hypotheses. Tests published since the 1970s show increasing concern with theoretical rationales throughout the test development process. A specific example of the integration of empirical and theoretical approaches is provided by the assignment of items to subtests on the basis of logical as well as statistical homogeneity. In other words, an item is retained in a scale if it had been written to meet the specifications of the construct definition of the particular scale and was also shown to belong in that scale by the results of statistical item analysis.

It is being recognized more and more that the development of a valid test requires multiple procedures, which are used sequentially, at different stages in test construction (Anastasi, 1986a; 1988, chap. 6; Jackson, 1970, 1973; Messick, 1980, 1988, 1989, 1991; *Standards*, 1985, pp. 9–18). Thus, validity is built into the test from the outset, rather than being apparently limited to the last stages of test development, as in the traditional reporting of criterion-related validation in test manuals. The validation process begins with the formulation of trait or construct definitions, which are derived from psychological theory, from prior research, or from systematic observation and analysis of real-life behavior domains, such as job analyses. Test items are then prepared to fit the construct definitions. Empirical item analyses follow, with the selection of the most valid items from the initial item pools. Other appropriate internal analyses may then be carried out, including factor analyses of item clusters or subtests. The final stage includes validation and cross-validation of various scores (and interpretive score patterns) through statistical analyses against real-life criteria.

The traditional concepts of content validity and criterion-related validity can be more accurately designated as content relevance and content representativeness for the first, and predictive and diagnostic utility for the second, as proposed by Messick (1980, 1988, 1989, 1991). In certain practical situations, data obtained by these traditional procedures may be needed to answer specific questions. Nevertheless, even in such cases, information about constructs enriches the understanding of the test findings. For one thing, constructs are more generalizable than are particular tested variables, and some generalizability beyond the immediate testing context is nearly always implied in the use of test results. As for criterion-related validity, ideally both tests and criteria should be described in terms of empirically established constructs, and the correspondence between the two sets of constructs investigated. Moreover, the constructs identified in the criteria (as in educational or work performance) could also be examined in relation to the goals explicitly set for such activities within specified value systems[2].

Counselors are concerned with information on test validation, as reported in test manuals and related supplementary publications, for at least two reasons. First, such information should help in selecting appropriate tests and evaluating their potential effectiveness for specific uses. Second, the correct interpretation of test scores requires knowledge about what the particular test measures, as indicated by the validation data. Interpretation of test results in terms of constructs, rather than specific measured variables, is likely to be especially relevant for counselors. Most counseling situations call for a broad understanding of the individual's assets and liabilities, in contrast to those testing situations that require a matching of individual skills and qualifications to narrowly defined task performance.

SUBSTANTIVE INTERPRETATION AND USE OF TEST RESULTS

What Test Results Do and Do Not Tell About a Person

In addition to basic knowledge about the statistical and technical properties of psychological tests illustrated in the preceding section, the counselor needs current knowledge about the behavior domain that the test is designed to assess. This requires familiarity with major substantive developments in the relevant areas of psychological science. A significant point for all test users to bear in mind is that test scores tell us *how well* individuals perform at the time of testing, not *why* they perform as they do. To find out why, we have to consider the test score within the person's *antecedent context*. We need to delve into the individual's reactional biography or learning history. In what environment did this person develop? What conditions and events were encountered, and how did the person respond to them?

From another angle, we need to examine the test score within the *anticipated context*. What is the setting—educational, occupational, societal—in which this person is expected to function, and for which he or she is being evaluated? What can we find out about the intellectual, emotional, and physical demands of that context? Several concepts encountered in the recent psychological literature, such as functional literacy and the assessment of competence, arise from this approach to test interpretation (Anastasi, 1988, pp. 424–428; Sticht, 1975; Sundberg, Snowden, & Reynolds, 1978). Can this person read at the level required for a job she is considering? Can this youth manage his own life in the community? Is this child ready to benefit from a particular educational program or some other planned intervention? Thus, we can see that the full understanding and proper interpretation of a test score has both a past and a future reference to specific real-life contexts.

It is now widely recognized in psychometrics that all cognitive tests measure *developed abilities*, which reflect the individual's learning history. This is equally true of tests traditionally labeled "aptitude tests" and those labeled "achievement tests." The two types of tests differ principally in the degree to which the requisite prior learning is specified and controlled (Anastasi, 1980; 1984; 1988, pp. 411–415).

The concept of developed abilities is also helpful in examining the widely debated question of *test bias*. The goal is for tests to be free from cultural bias against any group with which the tests are

used. This does not mean that there can be no group differences in test scores. Such differences may correctly reflect differences in antecedent development of the skills and knowledge covered by the particular test, which may also be required for the criterion performance that the test is designed to assess—in a course of study, a job, or other real-life context. Essentially, a test is free from bias and is equally fair to two groups if its scores have the same validity for both groups and do not underpredict the performance of either group. In terms of the familiar regression model, this refers to the avoidance of slope bias and intercept bias (see Anastasi, 1988, pp. 193–201 for fuller discussion and references).

Integration of Cognitive and Affective Data

Another relevant idea contributed by psychological research concerns the interrelations of different behavior domains (for further discussion and references, see Anastasi, 1985b; 1988, pp. 368–370). We commonly think of different tests as assessing either abilities or personality—the latter covering such areas as motivation, emotion, interests, attitudes, values. There is increasing evidence of mutual influence between these two major behavioral domains. Nor is this influence limited to immediate performance, as when a person tries harder or persists longer on a task that interests him or in an activity that ranks high in her value system. The influence is also evident in the development of traits over the life span. One way that motivation and other affective variables may contribute to the development of aptitudes is through the cumulative amount of time that the individual spends on a particular kind of activity relative to other, competing activities.

The effect of sheer time-on-task is enhanced by attention control. What one attends to, how deeply attention is focused, and how long attention is sustained contribute to one's cognitive growth. The selectivity of attention leads to selective learning—and this selection will differ among persons exposed to the same immediate situation. Such selective learning, in turn, may influence the relative development of different aptitudes and thereby contribute to the formation of different trait patterns. Essentially, the several aspects of attention control serve to intensify the effect of time devoted to relevant activities, and hence increase its influence on aptitude development. How much time and attention does the individual spend on a particular kind of activity, such as studying a certain subject or carrying out certain job-related functions? Does this student devote more time to studying for the math class or for the English lit. class? Does this employee devote more attention to cultivating favorable interpersonal relations with fellow workers and subordinates or to figuring out an improved method for performing the work?

The relation between personality and intellect is reciprocal. Not only do personality characteristics affect intellectual development, but intellectual level also affects personality development. The success an individual attains in the development and use of his or her aptitudes is bound to influence that person's emotional adjustment, interpersonal relations, and self-concept. In the self-concept, we can see most clearly the mutual influence of aptitudes and personality traits. The child's achievement in school, on the playground, and in other situations helps to shape her or his self-concept; and this concept at any given stage influences his or her subsequent performance. In this respect, the self-concept operates as a sort of private self-fulfilling prophecy. From the standpoint of test score interpretation, what all this means is that the prediction of a person's subsequent development can be substantially improved by combining information about motivation and interests with information about aptitudes.

Evolving Concepts of Intelligence

As a final example of the contributions of current psychological knowledge to effective test use, I have chosen the changing concepts of intelligence. This choice seems appropriate for two reasons: (a) the assessment of intelligence plays an important part in many types of counseling; and (b) among test users in general, there is still considerable misunderstanding about the nature of intelligence and what so-called intelligence tests measure. For several decades after the rise and popularization of intelligence tests, the term *intelligence* was burdened by excess meanings that accounted for common misinterpretations of test scores and misuses of tests. This situation led to the banning of intelligence tests in several school districts and to the elimination of the word *intelligence* from the titles of most recently developed or revised tests.

More recently, there has been a revival of interest in more sophisticated redefinitions of intelligence and a growing recognition of the contributions that appropriate measures of intelligence can make to the solution of practical problems. Intelligence is not a single, unitary ability, but rather a composite of several functions. The term denotes that combination of abilities required for survival and advancement within a particular culture. It follows that the specific abilities included in this composite, as well as their relative weights, vary with time and place. In different cultures and at different historical periods within the same culture, the qualifications for successful achievement differ. The changing composition of intelligence can also be seen within the life span of the individual, from infancy to late adulthood. One's relative ability tends to increase with age in those functions whose value is emphasized by one's experiential context; and it tends to decrease in those functions whose value is deemphasized.

Most well-known intelligence tests designed for school-age children or adults measure largely verbal abilities; to a lesser extent, they also cover abilities to deal with numerical and other abstract symbols. These are the abilities that predominate in school learning. Therefore, such intelligence tests can be regarded as measures of scholastic aptitude or school learning. Performance on these tests is both a reflection of prior educational achievement and a predictor of subsequent educational progress. Because the functions taught in school are of basic importance in modern, technologically advanced cultures, the score on a test of academic intelligence is also a partial predictor of performance in many occupations and other spheres of daily life. Much of our information about what intelligence tests measure comes from practical studies of the utility of tests in predicting educational and occupational achievement.

At a more theoretical level, basic research on the nature of intelligence has been proceeding apace. One approach is through the statistical procedures of factor analysis (Anastasi, 1988, pp. 374–390). The controversy over Spearman's *g* versus the group factors or separate aptitudes proposed by Thurstone and others flourished in the 1920s and 1930s. Recently, this controversy has been revived and has received considerable attention in the popular media.

In trying to work our way through this tangle of conflicting claims, we should bear in mind at least two points. First, the general factor identified in any one battery has often been loosely described as Spearman's *g*, suggesting a comprehensive general ability that underlies all intellectual activity. Actually, it represents only a general factor common to the tests in that battery. To conclude from such an analysis that a given test is heavily loaded with Spearman's *g* is misleading. It would be more meaningful to say that the general factor identified in that battery is heavily loaded with what that test measures, and this can be specified by examining the content of that test (e.g., verbal comprehension, mechanical aptitude, or whatever). This is what is normally done in naming any factor identified in a factor analysis—one looks at the test or tests in which the factor is heavily loaded and names the factor accordingly. The same procedure should be followed in naming a factor common to the whole battery.

The second point pertains to *why* factor analysis is conducted. Factor analysis is no longer regarded as a means of searching for *the* primary, fixed, universal units of behavior, but rather as a method of organizing empirical data into useful categories through an analysis of behavioral consistencies. Like the test scores from which they were derived, factors are descriptive, not explanatory; they do *not* represent underlying causal entities. Interest has shifted to the conditions in the individual's learning history that lead to the formation of factors or traits. What brings about the particular behavioral relationships that lead to identifiable and differentiable ability constructs, such as verbal comprehension or numerical reasoning? (Anastasi, 1970, 1983, 1986b.)

Once we recognize the descriptive nature of factors, we see that the description could occur at different levels. More and more, we are coming to think in terms of a hierarchical model of factors or abilities: at the top is a general factor; at the next level are broad group factors, similar to some of Thurstone's primary mental abilities; these major group factors subdivide into narrower group factors at one or more levels; the factors specific to each measure or indicator are at the bottom level. Different theories focus on one or another level of this comprehensive hierarchical model. No one level, however, need be regarded as of primary importance. Rather, each test constructor or test user should select the level most appropriate for her or his purpose.

Another approach for investigating the nature of human intelligence is that of *cognitive psychology*[3]. This is a more recent and rapidly spreading development in psychology as a whole. From the standpoint of testing, the principal contribution of cognitive psychology is its concern with what the individual does when performing an intellectual task. Cognitive research concentrates on the *processes* rather than the *products* of thinking. In contrast, test performance typically assesses the products, as reported in test scores. Although interest in processes is not new in the history of psychometrics, cognitive psychologists have carried the techniques of process analysis to new heights of refinement and sophistication. Knowledge about the processes an individual uses in solving problems or performing intellectual tasks is especially useful in diagnostic testing, because it can help to pinpoint the sources of an individual's difficulties. It is also highly relevant to the designing of training programs and other interventions to fit individual needs. Recognizing that what intelligence tests measure is *not* a fixed or unchanging entity within the individual, psychologists in several countries have been exploring training procedures for improving intelligence[4].

SUMMARY

In the proper and effective application of psychological tests, the counselor (among other test users) plays a prominent role. A major reason for the misuse of tests is inadequate or outdated knowledge about both the statistical aspects of testing technology and the psychological findings regarding the behavior assessed by the tests. Because advances in both kinds of knowledge are progressing rapidly, it is essential for test users to keep abreast of relevant developments in both areas. These needs of test users are receiving increasing attention through such channels as professional journals and other widely available publications, association meetings, refresher courses, and workshops. It is well to remember that many test users may have completed their formal course training as much as a decade back, in courses taught by instructors whose own training may have been even further outdated. Just as old scores in the test taker's file need to be updated, so do old courses in the test user's educational history.

[1]It is noteworthy that, in this edition, the last word in the title was changed from "tests" to "testing."

[2]An unusually detailed and thoughtful analysis of validity from diverse angles can be found in Messick (1989).

[3]See, e.g., Embretson (1983, 1986), Hunt (1985), Simon (1976), Sternberg (1981, 1984). For brief overview and additional references, see Anastasi (1988, pp. 159–161).

[4]For brief overview and references, see Anastasi (1988, pp. 364–367).

REFERENCES

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

Anastasi, A. (1980). Abilities and the measurement of achievement. In W. B. Schrader (Ed.), *Measuring achievement: Progress over a decade* (pp. 1–10). San Francisco: Jossey-Bass.

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

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.

Anastasi, A. (1985a). Interpreting scores from multiscore batteries. *Journal of Counseling and Development, 64,* 84–86.

Anastasi, A. (1985b). Reciprocal relations between cognitive and affective development: With implications for sex differences. In T. B. Sonderegger (Ed.), *Psychology and gender* (Nebraska Symposium on Motivation, Vol. 32, pp. 1–35). Lincoln: University of Nebraska Press.

Anastasi, A. (1986a). Evolving concepts of test validation. *Annual Review of Psychology, 37,* 1–15.

Anastasi, A. (1986b). Experiential structuring of psychological traits. *Developmental Review, 6,* 181–202.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Anastasi, A. (in press). The test user qualifications project: An evaluation. *American Psychologist.*

Butcher, J. N. (Ed.). (1987). *Computerized psychological assessment: A practitioner's guide.* New York: Basic Books.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S. E. (1986). Intelligence and its measurement: Extending contemporary theory to existing tests. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 3, pp. 355–368). Hillsdale, NJ: Erlbaum.

Eyde, L. D., Moreland, K. L., Robertson, G. J., Primoff, E. S., & Most, R. B. (1988). Test user qualifications: A data-based approach to promoting good test use. *Issues in Scientific Psychology* (Report of the Test User Qualifications Working Group of the Joint Committee on Testing Practices). Washington, DC: American Psychological Association.

Eyde, L. D., Moreland, K. L., Robertson, G. J., Primoff, E. S., & Most, R. B.

(in press). Test user qualifications: Overview of a data-based project on promoting good test use. *American Psychologist.*

Guidelines for computer-based tests and interpretations. (1986). Washington, DC: American Psychological Association.

Henry, R. A., & Hulin, C. L. (1987). Stability of skilled performance across time: Some generalizations and limitations on utilities. *Journal of Applied Psychology, 72,* 457–462.

Hunt, E. (1985). Verbal ability. In R. J. Sternberg (Ed.), *Human abilities: An information-processing approach* (pp. 31–58). New York: Freeman.

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). Orlando, FL: 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.

Lambert, N. M. (1989, August). *The crisis in measurement literacy in psychology and education.* Invited address presented at the annual convention of the American Psychological Association, New Orleans, LA.

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

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1991). Validity of test interpretation and use. In M. C. Alkin (Ed.), *Encyclopedia of educational research* (6th ed.). New York: Macmillan.

Simon, H. A. (1976). Identifying basic abilities underlying intelligent performance of complex tasks. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 65–98). Hillsdale, NJ: Erlbaum.

Standards for educational and psychological testing. (1985). Washington, DC: American Psychological Association.

Sternberg, R. J. (1981). Testing and cognitive psychology. *American Psychologist, 36,* 1001–1011.

Sternberg, R. J. (1984). What cognitive psychology can (and cannot) do for test development. In B. S. Plake (Ed.), *Social and technical issues in testing: Implications for test construction and usage* (pp. 39–60). Hillsdale, NJ: Erlbaum.

Sticht, T. G. (Ed.). (1975). *Reading for working: A functional literacy anthology.* Alexandria, VA: Human Resources Research Organization.

Sundberg, N. D., Snowden, L. R., & Reynolds, W. M. (1978). Toward assessment of personal competence and incompetence in life situations. *Annual Review of Psychology, 29,* 179–221.

Technical recommendations for psychological tests and diagnostic techniques. (1954). Washington, DC: American Psychological Association.

Tyler, B., & Miller, K. (1986). The use of tests by psychologists: Report on a survey of BPS members. *Bulletin of the British Psychological Society, 39,* 405–410.

Anne Anastasi *is professor emeritus of psychology, Graduate School of Arts and Sciences, Fordham University, New York. Correspondence regarding this article should be sent to Anne Anastasi, Psychology Department, Graduate School of Arts and Sciences, Fordham University, Bronx, NY 10458.*