Ann. Rev. Psychol. 1986. 37:1-15

EVOLVING CONCEPTS OF TEST VALIDATION[1]

Anne Anastasi

Department of Psychology, Fordham University, Bronx, New York 10458

CONTENTS

In the early beginnings of standardized testing, validity was assessed by a diversity of procedures and was called by many names. The type of evidence adduced to demonstrate test validity varied with the purpose of the test, the theoretical orientation of the test author, and—all too often—with the ready availability of the data. Among the earliest empirical approaches to evaluating test items and selecting the most valid was the age-differentiation criterion employed by Binet and Simon (1908). On the assumption that the cognitive skills constituting intelligence increase with age through childhood, they chose tasks whose frequency of correct solution increased with age; they then assigned each task to the age level at which the percentage of children passing it fell within a specified range. This was also a major procedure followed in the construction of the Stanford-Binet and other individual intelligence tests of the period that assessed intelligence in terms of mental age.

Soon total test scores were being evaluated, not only against chronological age but also against judgments of individual achievement, such as teachers' ratings of pupils' performance or other evidence of the quality of behavior in

[1]This chapter is based in part on an invited address presented at the 1984 annual meeting of the American Educational Research Association in New Orleans. This is the seventh in a series of prefatory chapters written by eminent senior psychologists. For more information about this series, see the Preface to Volume 36.

0066-4308/86/0201-0001$02.00

daily life. Case history data and psychiatric diagnoses also served as criteria, especially for personality tests and tests designed for identifying mental retardation. With advances in statistical methodology, techniques of item analysis against total test scores or against external criterion measures came into use. Still later, factor analysis was introduced in test development; it was applied to items, to subtest scores, and to total test scores in combination with scores on other tests. Different investigators and test authors employed a confusing array of names for the validity they reported, ranging from face validity, validity by definition, intrinsic validity, and logical validity to empirical validity and factorial validity.

In 1954, in a major effort to introduce some order into the chaotic state of test construction procedures as a whole, the American Psychological Association (in collaboration with the American Educational Research Association and the National Council on Measurement in Education) published the *Technical Recommendations for Psychological Tests and Diagnostic Techniques*. This publication formally introduced the now familiar classification into content, predictive, concurrent, and construct validity. In subsequent editions of this document (*Standards* . . . 1974), predictive and concurrent validity were subsumed under criterion-related validity, and this tripartite division has survived to the present.

Although initially helping to clarify our thinking about validation procedures, the tripartite categorization of validity has had some adverse side effects on testing practice. Essentially it represents a crude and oversimplified grouping of many data-gathering procedures that contribute to an understanding of what a test measures. Yet there has been a tendency to lean too heavily on this neat, satisfying tripartite classification. The three labels have been reified and endowed with an existence of their own. They first came to be regarded as three distinct *types* of validity and later as three essential *aspects* or *components* of validity. Thus test constructors would feel obliged to tick them off in checklist fashion. It was felt that they should be covered somehow in three properly labeled validity sections in the technical manual, regardless of the nature or purpose of the particular test. Once this tripartite coverage was accomplished, there was the relaxed feeling that validation requirements had been met. This, of course, is a gross distortion of the role of validity in the test development process. It is noteworthy that in the 1985 edition of the *Standards for Educational and Psychological Testing,* some of the apparent rigidities of the earlier editions were eliminated and a more comprehensive and flexible approach to validation procedures was followed.

THE PLACE OF VALIDITY IN THE TEST CONSTRUCTION PROCESS

Let us turn to a basic question: How does one build a valid test? What are the ideal test-construction procedures? What is the general model of test develop-

ment that the test author endeavors to approximate within the constraints of practical demands and real-life limitations?

More and more we recognize that the development of a valid test requires multiple procedures, which are employed sequentially at different stages of test construction (Jackson 1970, 1973, Guion 1983). Validity is thus built into the test from the outset rather than being limited to the last stages of test development, as in traditional criterion-related validation. The validation process begins with the formulation of detailed trait or construct definitions, derived from psychological theory, prior research, or systematic observation and analyses of the relevant behavior domain. Test items are then prepared to fit the construct definitions. Empirical item analyses follow, with the selection of the most effective (i.e. 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 combinations of scores through statistical analyses against external, real-life criteria.

This multistage process for building validity into a test is illustrated in varying degrees by several recently developed tests. Among them are the Comrey Personality Scales (1970) and the Millon Clinical Multiaxial Inventory (1983). It is most clearly exemplified by the Personality Research Form developed by Jackson (1974), who has contributed substantially to the dissemination of the multistage procedure (Jackson 1970, 1973). In the cognitive domain, the procedure is illustrated by the recently published Kaufman Assessment Battery for Children (1983; see also Anastasi 1984), although the reporting of validity in the interpretive manual still follows the traditional approach. There are separate sections labeled construct, concurrent, and predictive validity; and relevant information from other stages of test development, such as construct formulation and several kinds of item analysis, is scattered through other chapters.

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 our 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.

If we think of test validity in terms of understanding what a particular test measures, it should be apparent that virtually any empirical data obtained with the test represent a potential source of validity information. After a test is released for operational use, the interpretive meaning of its scores may continue to be sharpened, refined, and enriched through the gradual accumulation of

clinical observations and through special research projects. The former was well illustrated by the Stanford-Binet, the latter by the MMPI. Test validity is a living thing; it is not dead and embalmed when the test is released. Obviously, this does not mean that the test is not ready for use until all possible data bearing on its validity are in. Construct validation is indeed a never-ending process. However, that should not preclude using the test operationally to help solve practical problems and reach real-life decisions as soon as the available validity information has reached an acceptable level for a particular application. This level varies with the type of test and the way it will be used. Establishing this level requires informed professional judgment within the appropriate specialty of professional practice.

THE NATURE OF CONSTRUCTS IN TEST VALIDATION

By now it is undoubtedly apparent that I have been talking about what is traditionally known as construct validation. What about the other types, aspects, components, or modifying labels that have become generally associated with test validity? The answer is that what has come to be designated construct validity is actually a comprehensive approach that includes the other recognized validation procedures—and much more besides. This point has been made repeatedly: in the test standards (from the first, 1954 version to the latest), in textbooks, symposium papers, and journal articles. Yet the ambiguity persists. Probably the confusion results from the many usages of the term validity. In a 1980 paper, Messick argued convincingly that the term validity, insofar as it designates the interpretive meaningfulness of a test, should be reserved for construct validity. Other procedures with which the term validity has traditionally been associated, he maintained, should be designated by more specifically descriptive labels. Thus, content validity could be labeled 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 concurrent validation. These changes in terminology should help, but it may be some time before the old terms can be dislodged.

If we turn from labels to procedures, we can see 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. Certain procedures may be singled out for special emphasis in order to answer specific practical questions. But constructs are always involved, in both the questions and the answers, even though we may not be aware of it.

Let us consider the nature of the constructs employed in test development. Essentially they are theoretical concepts of varying degrees of abstraction and

generalizability which facilitate the understanding of empirical data. They are ultimately derived from empirically observed behavioral consistencies, and they are identified and defined through a network of observed interrelationships. In the description of individual behavior, such a construct corresponds closely to what is generally termed a trait. A simple example, with narrowly limited generalizability, is speed of walking. If we take repeated measurements of an individual's walking speed, we still obtain a whole distribution of speeds, depending upon the person's condition at the time, the context in which the walking occurs, and the purpose of the walking, among other circumstances. Nevertheless, it is likely that an analysis of such varied measures would reveal a substantial common factor that reliably differentiates one person from another in overall walking speed. This common factor would be a construct; it does not necessarily correspond to any single empirical measure.

Another relatively simple example is spelling ability. A test for this ability appears to be a likely candidate for content validation. But as in walking speed, we must guard against overgeneralizing from a behavior sample drawn from a limited domain. There may be several differentiable spelling abilities, and there is evidence that this is the case (e.g. Knoell & Harris 1952, Ahlström 1964). Such diverse spelling behaviors may be illustrated by the recognition of correctly and incorrectly spelled words, as in a multiple-choice or true-false test; frequency of misspellings in spontaneous writing; correctness of spelling when writing from dictation; and sensitivity to one's potential spelling errors, with the associated readiness to verify spelling by consulting sources. In designing a spelling test for a particular purpose, as for inclusion in a job selection battery, one can define the scope and boundaries of the construct that best fits the specific needs. This practice has been followed, to quote another example, in designing tests of functional literacy, or reading ability, for various industrial and military occupational specialties (Sticht 1975, Schoenfeldt et al 1976).

I have deliberately chosen examples of relatively narrow constructs because they can be more readily grasped. If we go to the other extreme of breadth, complexity, and generalizability, many of us would undoubtedly think of intelligence as a construct. I would rather not use that example, however, for at least two reasons. First, the term intelligence has acquired too many excess meanings that obfuscate its nature. Second, the construct measured by tests of intelligence requires some modifying adjectives and delimiting specifications. No test was actually designed to measure universal human intelligence. Some tests could be more accurately described as measures of academic intelligence, or scholastic aptitude, or that cluster of cognitive skills and knowledge demanded and positively reinforced in modern, technologically advanced societies. Even more precise construct definitions would certainly improve the interpretability of scores obtained with most so-called intelligence tests.

When a test author sets out to develop a new test, it is highly unlikely that he or she does so without some idea about the construct or constructs to be assessed, however vaguely defined. Nor did this practice originate with the formal introduction of construct validation into the psychometric lexicon. Binet devoted considerable time to a formulation of his concept of intelligence; the development of his ideas on the subject can be traced through published writings spanning many years (Wolf 1973). At the time when the Binet-Simon tests were prepared, Binet's conception of intelligence included such behavioral qualities as attention control, directed thinking, comprehension, judgment, and self-criticism (Binet 1909/1911). The influence of these constructs in guiding the preparation of the scale can be readily recognized in the test items, many of which have survived in the Stanford-Binet.

Once the testing movement had been fully launched, however, there was a tendency to veer away from theoretical rationale and construct formulation. The knowledge and hypotheses that undoubtedly still guided initial item writing were deemphasized and were rarely discussed in connection with test validity. The test manuals created a general impression of almost blind empiricism. There was heavy reliance on empirical item selection from large, preliminary item pools, followed by ex post facto evaluation of the total test through validation and cross-validation against external criteria. Substantial validity shrinkage was regularly expected in cross-validation because of the large contribution of chance factors to item selection. It is well known that such shrinkage will be largest when the initial item pool is large, the proportion of retained items is small, and the sample of persons is small. Under conditions of blind empiricism, test validity may drop to virtually zero in cross-validation—there are some dramatic demonstrations of this fact in the literature (Kurtz 1948, Cureton 1950). Shrinkage can be drastically reduced, however, when items are prepared to fit clearly formulated hypotheses derived from psychological theory or from previous investigations of criterion requirements (Primoff 1952). It is apparent that clear construct definition as a guide to item writing is not only logically defensible but also efficient.

Empiricism need not be blind. The overemphasis on purely empirical procedures during the early decades of this century arose in part as a revolt against the armchair theorizing that all too often served as the basis for so-called psychological writings of the period. But theory need not be subjective speculation. Theory *can* be derived from an analysis of accumulated research findings and can in turn lead to the formulation of empirically testable hypotheses. The shift toward stronger theoretical orientation discernible in American psychology since midcentury produced a noticeable spinoff in test construction. Tests published in the 1970s and 1980s show increasing concern with theoretical rationales throughout the test development process. A specific example of the integration of empirical and theoretical approaches to test construction is

provided by the assignment of items to subtests or scales on the basis of logical as well as statistical homogeneity. In other words, an item is retained in a scale if it was written to meet the specifications of the construct definition of the particular scale and *also* was shown to belong in that scale by the results of factor analysis or other statistical procedures of item analysis (Comrey 1970, Jackson 1974, Millon 1983).

Let us look more closely at the sources of the constructs employed in test development. How are these constructs formulated by the test author? This question actually pertains to criterion analysis, that is, an analysis of *what* the author wants the test to assess. Regardless of the purpose of the test, this is the criterion question.

For the most general types of tests, designed for wide-ranging uses, a major source of guiding constructs is psychological theory and the accumulated store of prior research findings. Among the most common sources actually used by test developers are personality theories, clinical observations, factor-analytic investigations of human abilities, and, more recently, information-processing studies from cognitive psychology.

When tests are designed for use within special contexts, the relevant constructs are usually derived from content analyses of particular behavior domains. Such analyses have varied widely in their thoroughness, fullness, and precision. In educational contexts, the most characteristic tests are the so-called achievement tests, whose purpose is to assess the effects of academic learning and the individual's readiness for further learning of a similar nature. At the broadest level, the constructs for such tests are educational goals, translated into testable behavioral specifications. The sources are essentially consensual judgment data. Ideally, these data are systematically gathered under conditions that are clearly described and amenable to replication. At more specific levels, the criterion analyses are represented by systematic surveys of curricula, course syllabi, and textbooks, as well as judgment data obtained from recognized experts within subject-matter specialities.

In occupational testing, designed for personnel selection and classification, the criterion analysis is generally called a job analysis. To be effective, a job analysis should concentrate on those aspects of performance that differentiate most sharply between the better and the poorer workers. In many jobs, workers of different levels of proficiency may differ little in the way they carry out most parts of their jobs—only certain features of their jobs may bring out the major differences between successes and failures. In his classic book on *Aptitude Testing,* Clark Hull as early as 1928 stressed the importance of these differentiating aspects of job performance which he called "critical part-activities" (p. 286). Later this concept was reemphasized by John Flanagan (1949, 1954), under the name of "critical requirements." To implement the concept of critical requirements, Flanagan proposed the critical incident technique. This tech-

nique called for factual descriptions of specific instances of job behavior characteristic of either satisfactory or unsatisfactory workers. The focus on critical job requirements led, through various routes, to the development of the job element method for constructing tests and demonstrating their validity. Variants of this procedure have been applied to a wide diversity of jobs, in industry and in the public sector at the federal, state, and local levels (McCormick et al 1972, Primoff 1975, Menne et al 1976, Tordy et al 1976, Eyde et al 1981).

Essentially, job elements are the units describing critical work requirements. The job element statements are generated and rated by job incumbents and supervisors, chosen because they are thoroughly familiar with the job. Job elements refer to those specific job behaviors that differentiate most clearly between marginal and superior workers. Relying ultimately on the observations and judgment of experienced workers, the job element method provides techniques for systematically collecting and quantifying these judgments. Although various adaptations of the job element method differ in procedural details, all provide for the description of job activities in terms of specific behavioral requirements, from which test items can be directly formulated. The individual behavioral statements can, in turn, be grouped into broader categories or constructs, such as computational accuracy, spatial visualization, manual dexterity, or ability to work under pressure. There is a growing body of research aimed at the development of a general taxonomy of job performance in terms of relatively broad behavioral constructs (Fleishman 1975, Pearlman 1980). The job element method contributes to this goal and thereby facilitates the effective use of a test across many superficially dissimilar jobs.

TRAITS AND SITUATIONS

Any discussion of trait constructs must take into account the question of situational specificity. A long-standing controversy regarding the generalizability of traits versus the situational specificity of behavior reached a peak in the late 1960s and the 1970s. Several developments in the 1960s focused attention on narrowly defined "behaviors of interest" and away from broadly defined traits. In the cognitive domain, this focus is illustrated by individualized instructional programs and criterion-referenced testing and by the diagnosis and treatment of learning disabilities. In the noncognitive or personality domain, the strongest impetus toward behavioral specificity in testing came from social learning theory and the general orientation associated with behavior modification and behavior therapy (Bandura & Walters 1963; Bandura 1969; Goldfried & Kent 1972; Mischel 1968, 1969, 1973). All the advocates of behavioral specificity in both cognitive and noncognitive areas

directed their criticisms especially toward the early view of traits as fixed, unchanging, underlying causal entities. This kind of criticism had already been vigorously expressed in earlier writings by several psychologists and had been supported by appropriate psychometric research. In fact, few psychologists today espouse such extreme views of traits with their excess meanings and unwarranted implications.

On the other side of the controversy, the initial emphasis on extreme behavioral specificity, with its accompanying rejection of trait constructs, resulted at least in part from certain methodological constraints. These included predominantly low reliability of measures and failure to aggregate across observations so as to cancel out specific variance (Green 1978; Epstein 1979, 1980; Rushton et al 1983). There is now a growing consensus between the adherents of the opposing views (Mischel 1977, 1979; Anastasi 1983). We are coming to recognize more and more that in order to identify broad traits, we have to assess individuals across situations and aggregate the results. To meet different assessment needs, behavioral observations can be aggregated in different ways and with appropriate degrees of generality or specificity (Mischel & Peake 1982). The focus may be on intraindividual consistencies or on situational categories of varying degrees of breadth.

Both the theoretical discussions and the research on person-by-situation interaction have undoubtedly enriched our understanding of the many conditions that determine individual behavior. They have also contributed to the development of sophisticated research designs such as the application of multimode factor analysis (Tucker 1964, 1966; Levin 1965; Kjerulff & Wiggins 1976). By this technique, one can identify major factors in situations, in response styles, and in persons. In addition, there is a core matrix which integrates the three modes and permits their joint interpretation. For example, in an investigation of graduate student styles for coping with stressful situations (Kjerulff & Wiggins 1976), students who rated themselves as less professionally competent tended to feel anger at themselves for academic failures and anger at others for interpersonal difficulties; they were extremely anxious when facing academic problems, but not at all anxious in stressful situations for which there is no clear source of blame, such as losing subjects in an experiment.

When the heat of the controversy over traits and situations had dissipated, it was clear that situational variance is more conspicuous in analyses of personality traits than in analyses of abilities. For example, a person may be quite sociable and outgoing at the office, but shy and reserved at social gatherings. Or a student who cheats on examinations may be scrupulously honest in handling money. An extensive body of empirical evidence has been assembled by social learning theorists (Mischel 1968, Peterson 1968) showing that individuals

exhibit considerable situational specificity in several nonintellective dimensions, such as aggression, social conformity, dependency, rigidity, honesty, and attitudes toward authority.

Part of the explanation for the higher cross-situational consistency of cognitive than of affective functions may be found in the greater uniformity and standardization of the individual's reactional biography in the cognitive domain (Anastasi 1970). Schooling is a major influence in the standardization of cognitive experience. The formal school curriculum, for example, fosters the development of broadly applicable cognitive skills in the verbal and numerical areas. Personality development, in contrast, occurs under far less uniform conditions. Moreover, in the personality domain, the same response may elicit social consequences that are positively reinforcing in one type of situation and negatively reinforcing in another. The individual may thus learn to respond in quite different ways in different contexts.

From the standpoint of personality test development, it should be noted that one can also identify situationally linked traits. This way of categorizing behavior is illustrated by the familiar test anxiety inventories (Spielberger et al 1976, Sarason 1980, Spielberger 1980, Tryon 1980). Such inventories cover essentially a trait construct that is restricted to a specified class of situations, those covering tests and examinations. Individuals high in this trait tend to perceive evaluative situations as personally threatening. The test instructions may be modified to define the anxiety-provoking situations even more specifically by directing examinees to respond, for example, with reference to mathematics tests or essay tests. Constructs such as test anxiety can be identified by aggregating observations within the situationally defined behavior domain, thereby cancelling out error variance as well as specificity that is irrelevant to the construct definition. The behavioral consistencies identified through such aggregation may well prove to be of considerable interest both theoretically and practically

VALIDITY GENERALIZATION

The concept of situational specificity has played a somewhat different role in research on the validity of ability tests for personnel assessment. 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 school courses (Bennett et al 1984). Such findings led to widespread pessimism regarding the generalizability of test validity across different situations. Until the mid-1970s, "situational specificity" of psychological requirements was generally regarded as a serious limitation in the usefulness of standardized tests in personnel

selection (Guion 1976). In a sophisticated statistical analysis of the problem, however, Frank Schmidt, John Hunter, and their associates (Schmidt & Hunter 1977, Schmidt et al 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 et al 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 et al 1984). With such small samples, criterion-related validation is likely to yield inconclusive and uninterpretable results within any single study.

Applying their newly developed techniques to data from many samples drawn from a large number of occupational specialties, Schmidt, Hunter, and their coworkers 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 (Schmidt et al 1979, 1980, Pearlman et al 1980). The variance of validity coefficients typically found in earlier industrial studies proved to be no greater than 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.

When tests are used for *classification decisions,* whereby individuals are to be matched with the requirements of different types of jobs or different instructional programs, we need to investigate the boundaries of validity generalization for particular tests or combinations of tests. We need to identify the major constructs covered by the tests on the one hand and 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, consisting 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,b). For purposes of 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. A meta-analysis of data from over 500 U.S Employment Service validation studies was then conducted with the newly developed validity generalization techniques. This procedure yielded estimated validities of the appropriate aptitude composites for all jobs within each job family.

A more narrowly focused demonstration of the dual identification of behavioral constructs in tests and criteria was also based on analyses of U.S. Employment Service data (Gutenberg et al 1983). The investigators applied a behaviorally oriented job analysis inventory, the Position Analysis Questionnaire (PAQ), to 111 jobs for which validity data were available in the USES files The object of the research was to investigate the possible moderating effect of certain behavioral demands of a job on the predictive validity of different tests. Three job-analysis dimensions pertaining to decision making and information processing were found to correlate positively with the validities of the cognitive GATB tests (general, verbal, and numerical aptitudes) and negatively with the validities of psychomotor tests (finger and manual dexterity tests). 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. These findings are consistent with the aptitude constructs identified in the previously cited USES research on validity generalization. They also support the desirability of identifying behavioral constructs in both job functions and test performance when investigating the predictive effectiveness of tests.

SUMMARY

The concept of test validation has been undergoing continuing development, clarification, and refinement. Although test authors always begin with some notion, however vague, about the constructs they want to measure, there was an early period of atheoretical empiricism in test development. By the 1970s, the increasing emphasis on theory in American psychology was reflected in test development, with an increasing interest in construct validation. In effect, all validation procedures contribute to construct validation and can be subsumed under it. So-called content validation and criterion-related validation can be more appropriately regarded as stages in the construct validation of all tests. There is a growing recognition that validation extends across the entire test construction process; it encompasses multiple procedures employed sequentially at appropriate stages. Validity is built into a test at the time of initial construct

definition and the formulation of item-writing specifications; the hypotheses that guide the early developmental stages are tested sequentially through internal and external statistical analyses of empirical data. Depending upon the purpose of the test, trait constructs may be defined with different degrees of narrowness or breadth and may be linked to specified situational domains. The identification of constructs in both test performance and criterion behavior increases the efficiency of the test construction process and leads to the production of tests that are more valid theoretically, as well as more useful in meeting practical needs. An example of such effects is to be found in reduced item wastage and minimal validity shrinkage in cross-validation. Another example is the broadening of validity generalization through the identification of matching constructs in test performance and criterion behavior.

This approach is not limited to the test developer. Test users, too, can profitably use construct definition in specifying their particular testing needs (as in behaviorally oriented job analyses), and they can choose tests that have been shown to assess the relevant constructs. The same constructs should provide a basis for interpreting test scores. Finally, if it is feasible for the test user to obtain confirmatory follow-up data on the predictive effectiveness of a given test for a particular use, it would be more meaningful to correlate test scores with the relevant and practically significant criterion constructs than with a composite and amorphous assessment of overall criterion performance for each individual.

Literature Cited

Ahlstrom, K G 1964 *Studies in spelling Analysis of three different aspects of spelling ability*, Rep No. 20 Uppsala, Sweden Inst Educ , Uppsala Univ

Anastasi, A. 1970 On the formation of psychological traits *Am Psychol.* 25.899–910

Anastasi, A 1983 Traits, states, and situations. A comprehensive view. In *Principals of Modern Psychological Measurement A Festschrift for Frederic M Lord*, ed H Wainer, S Messick, pp 345–56 Hillsdale, NJ. Erlbaum

Anastasi, A. 1984 The K-ABC in historical and contemporary perspective *J Spec Educ.* 18.358–66

Bandura, A. 1969. *Principles of Behavior Modification* New York Holt, Rinehart & Winston

Bandura, A., Walters, R. H 1963 *Social Learning and Personality Development* New York Holt, Rinehart & Winston

Bennett, G. K., Seashore, H. G., Wesman, A G. 1984 *Differential Aptitude Tests Technical Supplement.* Cleveland Psychol Corp

Binet, A 1911 *Les idées modernes sur les enfants* Paris: Flammarion (original work published 1909)

Binet, A , Simon, Th. 1908 Le développement de l'intelligence chez les enfants *Année Psychol* 14 1–94

Comrey, A L 1970. *Comrey Personality Scales* San Diego Educ. Ind. Test Serv

Cureton, E. E. 1950. Validity, reliability, and baloney *Educ Psychol Meas.* 10 94–96

Epstein, S 1979 The stability of behavior. I On predicting most of the people much of the time. *J Pers Soc Psychol.* 37:1097–1121

Epstein, S 1980. The stability of behavior II Implications for psychological research *Am Psychol* 35.790–806

Eyde, L. D , Primoff, E. S., Hardt, R H 1981. *A job element examination for state troopers* (PRR-81-3) Washington, DC Personnel Res. Dev Cent., U.S. Off. Personnel Manage Natl Tech Inf Serv , PB 81 198772

Flanagan, J. C 1949 Critical requirements. A new approach to employee evaluation *Personnel Psychol* 2.419–25

Flanagan, J C. 1954 The critical incident technique. *Psychol Bull* 51.327–58

Fleishman, E. A. 1975 Toward a taxonomy of human performance *Am. Psychol* 30.1127–49

Ghiselli, E. E 1959. The generalization of validity *Personnel Psychol.* 12 397–402

Ghiselli, E E 1966 *The Validity of Occupational Aptitude Tests* New York. Wiley

Goldfried, M R , Kent, R N 1972 Traditional versus behavioral personality assessment· A comparison of methodological and theoretical assumptions *Psychol Bull* 77 409–20

Green, B. F Jr. 1978 In defense of measurement *Am Psychol* 33·664–70

Guion, R M 1976 Recruiting, selection, and job placement In *Handbook of Industrial and Organizational Psychology,* ed M D Dunnette, pp 777–828 Chicago. Rand McNally

Guion, R M 1983 Disunity in the trinitarian concept of validity. In *Clearing Away the Cobwebs A Closer Look at Content Validity* Symp meet Am Educ Res Assoc , Montreal, P Sandifer, Chair

Gutenberg, R L , Arvey, R D , Osburn, H G , Jeanneret, P. R 1983 Moderating effects of decision-making/information-processing job dimensions on test validities *J Appl Psychol* 68 602–8

Hull, C L 1928. *Aptitude Testing* Yonkers, NY. World Book

Jackson, D N 1970 A sequential system for personality scale development In *Current Topics in Clinical and Community Psychology,* ed. C D Spielberger, 2 61–96 New York Academic

Jackson, D N 1973. Structured personality assessment In *Handbook of General Psychology,* ed B. B Wolman, pp 775–92 Englewood Cliffs, NJ. Prentice-Hall

Jackson, D N 1974 *Personality Research Form Manual* Port Huron, Mich Res Psychol Press

Kaufman, A S , Kaufman, N L 1983 *Kaufman Assessment Battery for Children Interpretive Manual* Circle Pines, Minn Am Guidance Serv

Kjerulff, K , Wiggins, N H 1976 Graduate student styles for coping with stressful situations *J Educ Psychol* 68 247–54

Knoell, D M , Harris, C. W 1952 A factor analysis of spelling ability *J Educ Res* 46·95–111

Kurtz, A K 1948 A research test of the Rorschach test *Personnel Psychol* 1 41–51

Levin, J 1965 Three mode factor analysis *Psychol Bull* 64 442–52

McCormick, E J , Jeanneret, P R., Mecham, R C 1972 A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ) *J Appl Psychol* 56 347–68

Menne, J. W , McCarthy, W , Menne, J 1976. A systems approach to the content validation of employee selection procedures *Public Personnel Manage.* 5.387–96

Messick, S 1980 Test validity and the ethics of assessment. *Am Psychol.* 35:1012–27

Millon, T. 1983. *Millon Clinical Multiaxial Inventory. Manual.* Minneapolis NCS Interpretive Scoring Syst. 3rd ed.

Mischel, W 1968 *Personality and Assessment* New York: Wiley

Mischel, W 1969 Continuity and change in personality *Am Psychol* 24 1012–18

Mischel, W. 1973 Toward a cognitive social learning reconceptualization of personality *Psychol Rev* 80 252–83

Mischel, W 1977 On the future of personality measurement *Am. Psychol* 32:246–54

Mischel, W. 1979 On the interface of cognition and personality Beyond the person-situation debate *Am Psychol* 34:740–54

Mischel, W , Peake, P K 1982. Beyond déjà vu in the search for cross-situational consistency *Psychol Rev* 89:730–55

Pearlman, K 1980 Job families. A review and discussion of their implications for personnel selection *Psychol Bull.* 87:1–28

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 *J Appl Psychol* 65 373–406

Peterson, D. 1968 *The Clinical Study of Social Behavior* New York Appleton-Century-Crofts

Primoff, E S 1952 Job analysis tests to rescue trade testing from make-believe and shrinkage *Am Psychol* 7 386 (Abstr)

Primoff, E S 1975 *How to prepare and conduct job element examinations* Personnel Res Dev Cent , Tech Study 75–1 Washington, DC GPO

Rushton, J P , Brainerd, C J , Pressley, M 1983 Behavioral development and construct validity The principle of aggregation. *Psychol Bull* 94 18–38

Sarason, I G , ed 1980 *Test Anxiety Theory, Research, and Applications* Hillsdale, NJ Erlbaum

Schmidt, F L , Gast-Rosenberg, L , Hunter, J E 1980 Validity generalization results for computer programmers *J Appl Psychol* 65 643–61

Schmidt, F L , Hunter, J. E 1977 Development of a general solution to the problem of validity generalization *J Appl Psychol* 62 529–40

Schmidt, F L , Hunter, J E , Pearlman, K 1981 Task differences as moderators of aptitude test validity in selection. A red herring *J Appl Psychol.* 66 166–85

Schmidt, F L , Hunter, J E , Pearlman, K , Shane, G S 1979 Further tests of the Schmidt-Hunter Bayesian validity generalization model *Personnel Psychol* 32:257–81

Schmidt, F. L., Hunter, J. E., Urry, V W. 1976. Statistical power in criterion-related validation studies. *J. Appl. Psychol* 61:473–85

Schoenfeldt, L. F., Schoenfeldt, B. B , Acker, S. R , Perlson, M. R. 1976 Content validity revisited: The development of a content-oriented test of industrial reading. *J. Appl. Psychol.* 61 581–88

Spielberger, C. D 1980. *Test Anxiety Inventory: Preliminary Professional Manual* Palo Alto, Calif: Consult. Psychol. Press

Spielberger, C. D., Anton, W E , Bedell, J 1976 The nature and treatment of test anxiety In *Emotions and Anxiety. New Concepts, Methods, and Applications*, ed. M. Zuckerman, C. D. Spielberger, pp. 317–45. New York. LEA/Wiley

Standards for Educational and Psychological Testing. 1985. Washington, DC: Am Psychol Assoc.

Standards for Educational and Psychological Tests. 1974. Washington, DC: Am. Psychol. Assoc.

Stucht, T C., ed. 1975. *Reading for Working A Functional Literacy Anthology* Alexandria, Va. Human Resources Res Organ.

Technical Recommendations for Psychological Tests and Diagnostic Techniques 1954 Washington, DC Am Psychol Assoc

Tordy, G R., Eyde, L. D., Primoff, E. S , Hardt, R. H. 1976. *Job Analysis of the Position of New York State Trooper· An Application of the Job Element Method*. Albany. New York State Police

Tryon, G S 1980 The measurement and treatment of test anxiety. *Rev Educ. Res.* 50:343–72

Tucker, L R. 1964. The extension of factor analysis to three-dimensional matrices In *Contributions to Mathematical Psychology*, ed. N. Frederiksen, pp. 109–27. New York. Holt, Rinehart & Winston

Tucker, L R 1966 Experiments in multimode factor analysis. In *Testing Problems in Perspective*, ed. A. Anastasi, pp 369–79. Washington, DC: Am. Council Educ.

U S Department of Labor, Employment and Training Administration 1983a. *Overview of validity generalization*. USES Test Res. Rep No 43 Washington, DC: GPO

U.S. Department of Labor, Employment and Training Administration 1983b. *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: GPO

Wolf, T. H. 1973. *Alfred Binet*. Chicago: Univ Chicago Press