Shepard, L.A. (1993). "Evaluating Test Validity." In L. Darling-Hammon (Ed.), Review of Research in Education, Vol. 19. Washington, DC: AERA.

Chapter 9

Evaluating Test Validity

LORRIE A. SHEPARD
University of Colorado, Boulder

Validity theory has evolved over time, but the pace of change accelerated dramatically in the 1980s beginning with Cronbach's (1980) "Validity on Parole: How Can We Go Straight?" and culminating in Messick's (1989) landmark chapter in the third edition of *Educational Measurement*. In addition to innumerable articles on specific validity issues, the decade produced major conceptual syntheses such as the 1985 test standards (American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, 1985), articles by Anastasi (1986) and Landy (1986), several more contributions by Cronbach (1989) and Messick (1980, 1981a, 1981b), and no less than three keynote addresses at the 1986 Educational Testing Service (ETS) conference, "Test Validity for the 1990s and Beyond" (Angoff, 1988; Cronbach, 1988; Messick, 1988).

The purposes of this chapter are to trace the evolution of the current consensus on test validity and to point toward the direction of the next consensus. How have the definition of validity and methods of investigation changed over time? What common understandings are held currently by measurement theorists in education and psychology about the evaluation of test validity? What theoretical controversies or discrepancies between theory and practice exist that, if resolved, would strengthen both theory and practice?

The chapter is organized into sections as follows: (a) rejection of the old trinitarian doctrine; (b) construct validity as the whole of validity theory; (c) Messick's unified theory—the integration of test use, values, and consequences; (d) reformulating Messick's theory—evaluation argument as the construct validation of test use; (e) validity cases; and (f) conclusion: implications for the 1990s standards. In the paragraphs that

I am indebted to Lee J. Cronbach for his challenging and thoughtful critique of an earlier draft of this chapter. I also thank Robert L. Linn, Kenneth R. Howe, and Hilda Borko for their many helpful suggestions.

follow I provide a brief overview foretelling arguments advanced in the chapter and explaining the purpose of each section.

In the first two sections, I summarize the history of different types of validity and document the consensus that has emerged supporting a unified theory. Whereas before there were separatist camps advocating content validity for achievement tests and predictive correlations for selection tests, now there is considerable agreement that there should not be distinct types of validity for different kinds of tests. All types of test use require the multiple sources of evidence necessary in construct validation.

Validity does not inhere in a test—an insight found in some writings of measurement pioneers early in the century. Validity must be established for each particular use of a test. As stated by Cronbach (1971), "One validates, not a test, but an interpretation of data arising from a specified procedure" (p. 447). Procedure may refer to a formal test or other data-gathering instrument and includes the conditions of examinee preparation, test administration, and so forth. The new consensus now recognizes the full implication of Cronbach's statement. Every test use involves inferences or interpretation; therefore, all validation requires the combination of logical argument and empirical evidence needed to support those inferences. Landy (1986) likened this process of validation to traditional hypothesis testing. Anastasi (1986) concurred that construct validity is the superordinate category subsuming both content validity and criterion-related validity requirements. Not only must the logic of test development be checked by analyses using real-life criteria, but the validity of criterion measures must be evaluated, and so forth.

Despite apparent unanimity, however, there is a great deal more in what Cronbach and Messick have suggested than is acknowledged or accepted by the field. In the third section of this chapter, I discuss Messick's (1989) validity model, which significantly extends the original conception of construct validation. Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). What does it mean to say that actions based on test scores must be supported by evidence? If a school readiness test claims to measure which children are ready for regular kindergarten and which would benefit by waiting a year, then validity requires more than a correlation between test scores and school performance. It must be shown specifically that low-scoring children who spend an extra year in preschool or developmental kindergarten are more successful when they enter regular kindergarten (are better adjusted, learn more, etc.) than they would have been without the extra year. This demand for evidence of "aptitude-by-treatment" interactions to support placement decisions was laid out in

Cronbach's 1971 chapter and reiterated by the 1985 standards. Yet a review of current practice reveals widespread use of school readiness tests without any such validity evidence. In fact, existing comparative studies show no academic benefit and some emotional harm from extra-year kindergarten placements (Shepard, 1989), but these "research" studies of program effects have not been treated as germane to the "test validity" question.

Examples of the gap between validity theory and measurement practice are numerous. Several explanations for this dissonance are possible. One is that old habits die hard. Because existing terminology has been imbued with new meanings (rather than inventing new terms to signify changed understandings), it is possible for students of measurement to persist in the old forms. Thus, one can open a 1991 issue of *Educational and Psychological Measurement* and find a bald statement affirming the consensus of 20 years ago that in applied settings "the validity of tests is established by one of two strategies—content- or criterion-related."

Even psychometricians who acknowledge the preeminence of construct validity and who cite Messick's (1989) authoritative definition, however, are guilty of offering validity evidence in practice that is simplistic and incomplete. For example, test manuals are now likely to mention a little of each type of validity evidence but without weighing the evidence in the context of an organizing theory. As noted by Cronbach (1989), the construct validity sections in test manuals "rarely report incisive checks into rival hypotheses, followed by an integrative argument. Rather, they rake together miscellaneous correlations" (p. 155). Explanations for this type of practice are possibly that the integrative nature of construct validation is not understood or that its demands are perceived to be too complex really to be implemented. Messick's chapter goes on for 100 dense pages conveying the magnitude of the construct validation enterprise. Furthermore, like every other treatise on the topic, it contains the caveat that construct validation is a never-ending process. Notwithstanding their intent, these necessary qualifications give practitioners permission to stop with incomplete and unevaluated data.

Given this landscape, in the fourth section I suggest a different model or schema for evaluating test validity, one that focuses more centrally on intended test use. The model I propose does not differ substantively from that offered by Messick (1989). However, I argue that measurement professionals must take yet another step to make Messick comprehensible. Would-be authors of the 1990s version of the standards should try to find a simpler model for prioritizing validity questions, one that clarifies which validity questions must be answered to defend a test use and which are academic refinements that go beyond the immediate, urgent questions. Messick offers an integrated but faceted conception of validity that starts

with a traditional investigation of test score meaning and then adds test relevance, values, and social consequences. Orchestrated this way, typical validity investigations never get to consequences. A different approach would be to ask directly, "What does a testing practice claim to *do*?" and to organize the gathering of evidence around this question. This approach borrows most closely from Cronbach's (1988, 1989) conception of validation as evaluation and Kane's (1992) argument-based approach.

A final explanation for the mismatch between validity theory and practice might be the lack of real examples. Although each major review contains dozens of examples illustrating each piece in the validity puzzle, portrayals of complete validity investigations with the necessary weighing of evidence are not offered. Cronbach (1989) points to Wylie's (1974, 1979) review of self-concept research and to the century-long accumulation of insights about intelligence, but these examples speak to the validation of constructs in a research context rather than the evaluation of a particular test use. Validity cases are presented in the fifth section of this chapter. I address the claims and counterclaims and the corresponding research evidence for these cases: (a) the use of the Scholastic Aptitude Test (SAT) to make college selection decisions, (b) the use of the General Aptitude Test Battery (GATB) to make employee referral and selection decisions, and (c) the use of tests to make educational placement decisions in special education and 2-year kindergartens. These cases serve to illustrate the nature of evaluation argument but also to clarify how demanding validity standards are likely to be if the conceptions of Messick and Cronbach are taken seriously. Often, the use of a test is defended on the grounds that some information (i.e., test scores) is better than none. But for tests with unknown or inadequate validity evidence, this defense can no longer be accepted routinely. In some cases, if a test has negative effects, it is just as important to ask whether it is better *not* to test.

REJECTION OF THE TRINITARIAN DOCTRINE

When validity standards were first codified in 1954 (American Psychological Association, 1954), four types of validity were identified corresponding to different aims of testing.

> Validity information indicates to the test user the degree to which the test is capable of achieving certain aims. Tests are used for several types of judgment, and for each type of judgment, a somewhat different type of validation is involved. (p. 13)

Content validity was required for tests describing an individual's performance on a defined universe of tasks. *Predictive validity* was called for when a test was used to predict future performance and necessitated collecting criterion data later than the test. *Concurrent validity*, a separate type of validity involving an external criterion, was more appropriate when a new test was proposed as a substitute for a less convenient measure that was already accepted (e.g., a multiple-choice history test in place of a difficult-to-score essay examination). Concurrent validity data might also serve as a shortcut approximation of longitudinal predictive data. *Construct validity* was needed when making inferences about unseen traits such as intelligence or anxiety. In a subsequent revision of the standards (American Psychological Association, 1966), the two types of validity pertaining to an outside criterion were reduced to one category, *criterion-related validity*.

Content, criterion-related, and construct validity were referred to as the holy trinity by Guion (1980) because eventually the reification of the three separate paths to validity took on the character of a religious orthodoxy. In this section I summarize briefly the history of the first two types, focusing on the strengths and limitations of both. What are the defining concepts of content validity analysis that remain an essential part of today's more comprehensive validity studies? Why are content analyses alone insufficient to support test inferences? And similarly for criterion-related validity, what from this framework has been appropriated into current theory and why is it inadequate alone? Construct validity requires its own historical treatment because the theory invoked today by the same name is markedly different from the 1954 version. The evolution of construct validity, which has come to encompass both the empirical and logical demands of criterion and content validity, is described in the next section.

Criterion-Related Validity Evidence

Von Mayrhauser (1992) provides a "prehistory" of early means conceived to check test accuracy before there was a psychometric community. His account entwines the development of statistical procedures to assess the convergence of several measures with the ascendency of belief in a unitary mental ability. Thus, in the first part of the century, psychologists used correlations to learn about their tests but focused on the convergence or "reliability" of measures as evidence of validity. Before World War I, Scott (1917) was unusual in his advocacy of checking against a practical, external criterion—averaged judgments of on-the-job performance. Initially, such attention to practical utility was eschewed as unscientific. However, after World War I and the success of the Army Alpha, which epitomized the triumph of practical criterion correlations, the tide shifted.

From 1920 to 1950, test-criterion correlations became the standard for judging test accuracy. The classical verbal definition of validity, "Does

a test measure what it purports to measure?" was answered by a single correlation. In the 1940s, this identity was so strong that the term *validity* came to be used synonymously to mean a predictive correlation coefficient (a practice that, unfortunately, continues). Guilford (1946), for example, said that "a test is valid for anything with which it correlates" (p. 429). Angoff (1988) politely called this perspective atheoretical, while Anastasi (1986) called it "blind empiricism."

Empirical evidence of relations to external criteria is an integral part of today's definition of validity. Note here, however, that a "criterion" does not necessarily imply prediction of a future performance; it could refer to the generalization of performance to contemporaneous performances outside the testing context. Tests are always simplifications of what we intend to measure. Therefore, it is important to determine what distortions might be caused by a particular mode of assessment. To assess the fidelity of test scores in representing criterion skills and abilities, it is desirable to conduct more in-depth and elaborated studies, even though such data-collection efforts might not be practical as a regular part of the testing procedure. Thus, we might evaluate the validity of a written measure of reading comprehension for elementary school children by inviting story retellings or by oral interviews. Empirical verification is needed because, as has been shown repeatedly when such data are collected, the logic of test construction does not always ensure that a test measures as intended. Reading comprehension, for example, can become too great a component in a paper-and-pencil measure of mechanical aptitude, just as writing skill might confound open-ended assessment of reading ability.

As an aside, note that the meandering history of criterion-related validity has created more than one meaning for the term *criterion*. In some cases it implies the intended performance for which the test is only a proxy. In other cases, a criterion is used where a relationship is expected (as when IQ measures were correlated with teacher grades) but the test and criterion are not thought to be congruent or synonymous. Oddly, the measurement literature focused on distinctions in timing of the criterion measurement, whether concurrent or in the future, rather than addressing the substantive nature of the criterion. These issues are taken up under the construct validity framework. When considering fairness in employment testing, for example, validity conclusions will depend on what is "left out" of the prediction equation—is it random error or relevant but unmeasured predictors?

Criterion-related validity is especially important in practice when the decision following from the test is based expressly on the correspondence between test performance and expected criterion performance, as is true for both selection and placement decisions. If a test is used to select police officers for promotion or to place college freshmen in remedial English classes, a practically significant statistical relationship should be evident between test score and relevant criterion. In these examples, criteria might be performance as a supervisor or the adequacy of students' writing in normal course work prior to remedial intervention. (Note that these relationships are better reported in the form of regression equations than as correlations because regression coefficients are less vulnerable to fluctuations in sample variances.)

Empirical relations are necessary but not sufficient to establish the validity of test use. It is ironic that a field so attuned to the fallacy of mistaking correlation for causation in experimental contexts would be willing to accept correlations in the measurement sphere as immediate proof of test validity. Clearly, a test and criterion could correlate for the wrong reasons (e.g., if they shared the same bias). Guion (1974) offered the example of a spurious relation between arm length and assembly-line packing speed caused by the arbitrary distance examinees were seated from task materials. Even in the heyday of blind empiricism, some experts recognized the "criterion problem" (i.e., that it did not make sense to hang the validity of a test on an inaccurate or invalid criterion) (Gulliksen, 1950; Thorndike, 1949). There are also arguably instances when a measure of prerequisites used in a selection context would not show the desired statistical relation because of the lack of variability in admitted candidates. This would be true, for example, of a written driving test as a predictor of safe driving records or a measure of strength for firemen. After a minimum threshold is reached, test performance would not be expected to correlate with performance criteria. The relevance of these measures to job performance must therefore be evaluated in other ways.

Today's broader conception of validity requires not only that the relevance and integrity of criterion measures be evaluated, but that predictive claims themselves be defended. For example, Cronbach (1980) suggested that selection devices may be unfair despite predictive correlations if short-term training for low-scoring examinees can appreciably disrupt predictions. If a small investment in training can create a real boost in job performance and improve the performance of some candidates over others, it may not be fair (to the applicants) or wise (for the employer) to use test-based qualifications. Therefore, there is an increasing obligation to explain both why a test predicts and why that relationship should be relied upon in making decisions.

Content Validity Evidence

Content validity has its own checkered history, most often associated with the history of achievement testing but influenced by other schools of psychology contesting with the individual differences paradigm. Tyler's (1934) seminal work established both the methods of educational test de-

velopment and the framework for evaluating the substantive integrity of a test. For Tyler, achievement test construction should begin with the development of educational objectives that cover all important course aspirations. Unlike "tables of specifications" that were mere content outlines (e.g., Ruch, 1929), Tyler emphasized attention to the actual skills and knowledge that students would be able to demonstrate. Tyler was clear that items eventually written to test the objectives were samples of behavior. "A fundamental assumption in all testing is that a sampling of student reactions will give a measure of his reactions in a much larger number of situations" (Tyler, 1989, p. 23).

Content validity arguments have also been strongly influenced by the perspectives of operationalism and behaviorism (Bridgman, 1927; Skinner, 1945). Operational definitions of concepts used in scientific investigations are the rules specifying procedures by which each concept is to be observed and quantified. For operationalists, the validity of a measure was established directly by the quality of the substantive rationale for its development (i.e., by the precision of an explicit verbal definition and the reasonableness of the specific operations used to delimit population studied, setting, and means of data collection). However, in contrast to Tyler, operationalists intended no inference beyond the precise observation under stipulated conditions; hence, there was no need to verify the meaning of test results empirically. Measurement results (test scores) were valid tautologically.

In its extreme form, operationalism held that every instrument defined a different concept. One measure of IQ defined its version of intelligence, the next IQ test defined its version, and so forth. By definition, radical behaviorists denied that there could be any "surplus meaning" in a construct not captured by a specified measurement. Over time, advocates for this perspective have lost ground in their battle with construct validity because invariably test-based interpretations assume the generality of broader concept labels, such as "verbal ability" or "mathematics achievement," rather than limiting claims to a particular test given on a particular occasion. However, as recently as the 1970s, educational measurement specialists operating in this tradition argued that criterion-referenced tests were automatically valid because the test development process ensured content validity.

The behaviorists, like Tyler before them, offered a model for detailed and rational development of test content that countered blind empiricism. Although it is now unacceptable to rely on content validity as the sole basis of test validity, their model still provides the basic framework for "building validity into a test" (Anastasi, 1986) or for appraising the reasonableness of test content. First, conceptual analysis is used to define a construct or content domain. Then, test items or tasks are developed

to represent the intended domain. Construct and content domain are used interchangeably here because, as we shall see, the old distinction based on the notion that constructs refer to hidden psychological traits has been abandoned. Tenopyr (1977), for example, refers to typing ability as a construct. Thus, constructs might be knowledge of first-year college chemistry, competencies required of a beginning attorney to be measured on a bar exam, or achievement motivation. Note that domain and universe can also be used interchangeably. Content domains for constructs are specified logically by referring to theoretical understandings, by deciding on curricular goals in subject matter fields, or by means of job analysis.

The evaluation of content validity follows the same steps as test development. First, it must be established whether the content universe addressed by the test is appropriate. Then the adequacy of the sampling from the domain is considered. A test evaluation begins by reasking questions such as "What aspects of case law must candidates for the bar know by heart?" and "What should they know how to reference well?" When judging a lawyer's competence, what is the proper balance between book learning and ability to mount an oral argument? In the context of medical education, is it reasonable for a consortium of medical school deans to assert that some amount of quantitative knowledge is essential to success in medicine, to read graphs, follow quantitative arguments, and so forth? What level of mathematics is defensible? Should mathematics knowledge be assessed with as little demand on other content areas as possible or should problems be contextualized in relevant science content, and so forth?

The second step in content evaluation is to judge whether the tasks or items in the test adequately sample the intended universe. At one time it was imagined that a domain could be defined so exactly that items could be sampled literally at random from it or that item-generating rules could be written making it possible to translate intended domains into operational tests by a mechanical procedure. This never worked, however, with content domains any more complex than two-digit addition problems. Therefore, instead of mechanical verification, the representativeness of test content must be established by logical analysis. This analysis should address not just the obvious question of whether tasks match the domain specifications but of whether different formats or modes of assessment might alter content or construct meanings.

The method for establishing or evaluating the reasonableness of test content is usually expert judgment. For example, the principles for validation developed by industrial and organizational psychologists (Society for Industrial and Organizational Psychology, 1987) argue that content sampling for both selection procedures and criterion measures

follows from the professional judgment of the researcher. It may also involve the judgments of job experts and a job analysis that details critical tasks, important components of the work context, and behaviors required of the worker to perform the job. (pp. 19–20)

In the same vein, mathematics experts were convened to determine the content framework for the National Assessment of Educational Progress (NAEP); and a different group of mathematics experts participated in an evaluation study of the NAEP (National Academy of Education Panel, 1992) to investigate further the correspondence between the test framework and the curriculum standards established by the National Council of Teachers of Mathematics. An example of an evaluative question that can be answered only by making a reasoned judgment is "How much should the content of the NAEP be geared to current instructional practice and how much should it assess curricular goals that are aspirations for the future?"

Conceptual analysis is necessary not just to determine the outline of a content domain but to elaborate all of its internal elements: the subdomains of content such as geometry, statistics, and number concepts in mathematics; the relation of specific tasks to the intended constructs; and the processes thought to underlie test performance. The same type of analysis is needed to establish the rationale for a particular test use and for evaluating the reasonableness of a variety of external criteria. As becomes clearer in the context of construct validity, conceptual and substantive analyses are not carried out as a separate enterprise from empirical studies; rather, they guide and shape empirical research questions.

Content analysis alone is not sufficient to defend the validity of a test because there can always be unanticipated effects that disrupt the intended connection between test score and construct. For example, advanced placement tests are based on careful curriculum specifications. Yet empirical data in several subject areas show substantial gender effects. The multiple-choice portions of the tests are relatively easier for men and the essay portions are differentially easy for women. Can the validity of the current content configuration be defended? This finding has led to a series of follow-up studies examining the influence of writing ability on essay scores and the separate predictive correlations for each subpart of the tests (Breland, 1991; Bridgeman & Lewis, 1991). This whole complex of investigations exemplifies the type of evaluative research needed to establish a validity claim. Ultimately, evidence that a test may not be measuring as expected leads to a new round of substantive debate within a community of stakeholders who now have access to the insights gained from empirical investigations.

Messick (1975) has suggested that the term *content validity* not be used

because the conceptual analysis it invokes leads to a conclusion about the test, whereas conclusions about validity must always pertain to the meaning of scores and examinee responses. Yalow and Popham (1983) objected sharply, fearing that "efforts to withdraw the legitimacy of content representativeness as a form of validity may, in time, substantially reduce attention to the import of content coverage" (p. 11). They argue that content validity, which focuses "on the test itself," is a necessary precursor to drawing reasonable inferences based on test scores. In my view, the content-response dichotomy is merely a restatement of the logical-empirical categories of validity evidence. Once it is agreed that each is insufficient without the other and without a construct evaluation framework, it seems unnecessary to denigrate content validity without similarly demoting criterion-related validity. The objection should be to the use of either term as if it could stand alone. So far, I have hedged my bets by adopting the familiar terms to modify the word *evidence*. Later I develop the thesis that "content validity evidence" and "criterion-related evidence" are assessed in the context of an integrated construct validity evaluation.

Reification of Separate Validity Types

By the time the 1974 version (American Psychological Association, American Educational Research Association, and National Council on Measurement in Education, 1974) of the test standards was formulated, the interrelatedness of the three different aspects of validity was recognized, at least in theory: "These aspects of validity can be discussed independently, but only for convenience. They are interrelated operationally and logically; only rarely is one of them alone important in a particular situation" (p. 26). Separate and exclusive methods of validation became entrenched, however, because of camps that identified with either the empiricist or operationalist traditions and because of the legal force given to the distinctions in the uniform guidelines (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice, 1978). In education, many authors of criterion-referenced tests and minimum competency tests argued that content validity was sufficient to establish the meaning of test scores (see Hambleton, 1980, 1984, to the contrary). In the world of personnel psychology, both Guion (1980) and Landy (1986) decried the oversimplification of validity principles fostered by the legal (rather than scientific) status of the guidelines. Guion coined the metaphor of the "holy trinity," and Landy likened Title VII litigation to a primitive form of stamp collecting—whereby a test is pasted wholly into the content space, the construct space, or the criterion-related space. The remedy, proposed by Guion, Landy, Tenopyr, and others, along with Cronbach and Messick,

is to embrace a unified conception of validity under the construct validity framework.

CONSTRUCT VALIDITY AS THE WHOLE OF VALIDITY THEORY

Construct validity was formally introduced in the 1954 standards and extended in the now-classic paper written by two of the standards committee members, Cronbach and Meehl (1955). The formulation of construct validity was simply the application of scientific theory testing to confirm (or disconfirm) the interpretation of test scores. It drew together the requirements for both rational argument and empirical verification. According to the standards (American Psychological Association, 1954), construct validation involves first making predictions based on theory and then gathering data to confirm those predictions.

Cronbach and Meehl (1955) elaborated the model of theory testing by developing the concept of "nomological net." The construct to be measured is located in a conceptual space showing its hypothesized connections to other constructs and observed behaviors. These theoretical relationships are then tested empirically through correlational and experimental studies. The term *nomological,* meaning law-like, was used because scientific investigations were intended to identify lawful regularities. The following are key points, as summarized by Cronbach and Meehl (1955):

> 1. A construct is defined implicitly by a network of associations or propositions in which it occurs. . . .
>
> 2. Construct validation is possible only when some of the statements in the network lead to predicted relations among observables. While some observables may be regarded as "criteria," the construct validity of the criteria themselves is regarded as under investigation. . . .
>
> 4. Many types of evidence are relevant to construct validity, including content validity, interitem correlations, intertest correlations, test-"criterion" correlations, studies of stability over time, and stability under experimental intervention. High correlations and high stability may constitute either favorable or unfavorable evidence for the proposed interpretation, depending on the theory surrounding the construct. (pp. 299–300)

The early version of construct validity was both too demure and too ambitious compared with our present-day understandings. The 1954 standards and Cronbach and Meehl (1955) introduced construct validation as the weak sister to the existing view of validity, suggesting it as a substitute when a real criterion was not available. "Construct validity is ordinarily studied when the tester has no definitive criterion measure of the quality with which he is concerned, and must use indirect measures to validate the theory" (American Psychological Association, 1954, p. 14). In contrast, today we say that all test interpretations require construct validation.

Early theory was overly ambitious to the extent that it hoped to prove ultimately a hard-wired system of regularities. Cronbach and Meehl's (1955) construct validity model was steeped in the assumptions of logical positivism, which dominated the philosophy of science at the time but has since been repudiated. Today, it is less tenable for social scientists to assume either that human behavior is governed by laws, akin to the laws of physics, waiting to be uncovered or that constructs and observables can be meaningfully separated given that observation occurs through an interpretive lens. Cronbach himself (1989) has abandoned the positivist requirements of the 1950s conceptualization: "It was pretentious to dress up our immature science in positivist language" (p. 159). Nevertheless, by some other name the organizing and interpretive power of something like a nomological net is still central to the conduct of validity investigations. Perhaps it should be called a conceptual network or a validity framework. Despite the changes in philosophy, however, any validation effort still consists of stating hypotheses and challenging them by seeking evidence to the contrary.

In addition to the central requirement for conjoint logical and empirical analysis, the traditional literature on construct validity has several other key features that continue to inform present-day research. First, validity studies must address both the internal structure of the test and external relations of the test to other variables. Second, empirical evidence may include both correlational data and experimental interventions. Third, support of a theory requires both confirming evidence and the ruling out of plausible rival hypotheses. Finally, convergent and discriminant correlations are a convenient tool for pursuing typical construct validity questions, but they cannot be interpreted mechanically. These points are addressed briefly in the paragraphs that follow. For a more complete treatment, see Messick (1989).

Internal and External Components of Construct Validity

The conceptual framework that lays out our understanding of a construct includes both an internal model of interrelated dimensions or subdomains of a construct and an external model depicting its relationship to other constructs. Loevinger (1957) referred to three aspects of construct validity: the substantive component, the structural component, and the external component, the first two of which are subsumed by the internal model. Embretson (1983) called the two parts *construct representation* and *nomothetic span.*

The internal model should reflect all aspects of the theory that defines a construct, including its subdomains or subconstructs, the expected in-

terrelationships among dimensions of the construct, and the processes believed to underlie test performance. For example, Shavelson, Hubner, and Stanton (1976) outlined the internal features of "self-concept" that should guide a validity investigation. Self-concept is believed to be "organized, multifaceted, hierarchical, stable, developmental, and evaluative" (p. 411). The multifaceted and hierarchical hypotheses predict that measures of academic self-concept should be correlated with measures of social, emotional, and physical self-concept, but not as highly as they are with each other. Students could think that they are bad at doing math but still have a positive self-concept. General self-concept is expected to be more stable than are self-evaluations of more specific behaviors lower in the hierarchy, and so forth. The model delineated by Shavelson et al. implies correlations of different strengths that are amenable to empirical corroboration. The internal portion of construct validation includes gathering data about all of the traditional psychometric questions regarding item intercorrelations and the like, but also includes questions about the appropriate weighting of different components and the influence of format on what is tested. What weight should be given to geometry in a test of mathematics? Does choosing correct answers on a test rather than producing problem solutions imply the same or different representations of the construct?

Methods for evaluating internal structure have become more sophisticated in recent years. Loevinger's (1957) early conception of substantive validity included both the logic of content validation and the empirical corroboration of hypothesized item-item and item-criterion correlations. Structural validity referred to the correspondence between the test (interitem) structure and the structure of these same relationships outside the test—again assessed by means of correlations. More recent theorists have proposed that these internal relationships be evaluated using mathematical models to account for item responses. For example, Embretson (1983) showed how multicomponent latent trait modeling could be used to decompose item responses so as to examine different theoretical mechanisms that account for task success. These components might be cognitive processes, strategies, or knowledge stores. Similarly, Wiley (1991) has proposed a deep analysis of both the valid and invalid sources of variance in task performance using variance-covariance structures. Wiley's modeling permits an evaluation of the internal portion of the nomological net by distinguishing between intended abilities underlying test performance (such as reading comprehension) and ancillary skills (such as reading speed, motivation, simple recognition and recall, prior knowledge, vocabulary, or test-taking strategy).

The external portion of the conceptual framework models the relation of the intended construct to other constructs. For example, self-concept

is believed both to be the cause of and to be caused by successful experiences, such as doing well in school. Therefore, the conceptual network would show a positive relationship between school achievement and self-concept, but the model would also require a time dimension to test whether change in one indicator was followed by change in the other.

Traditionally, the nomological net has represented construct meaning. Do measures of intelligence correlate closely with measures of scholastic aptitude? Do so-called measures of aptitude and measures of achievement respond differently to exposure to course content? As I discuss in later sections of the chapter, however, there is no reason that the construct framework should not include, and even emphasize, those relationships most centrally implicated by an intended test use. If a general knowledge test is believed to be related to job performance, the traditional approach would be to surround the test with other indicators of general knowledge and then to include its correlation with a criterion measure. This approach addresses two validity questions separately: Does the test measure general knowledge? and Is it predictive of job performance? If we think instead that construct validity should evaluate test use, then for an employee selection test we would include a larger web of interconnections involving the concomitant influences of prior work experience, motivation, and level of education, as well as multiple criterion measures involving both independent indicators of employee success and additional outcomes such as job satisfaction and longevity. Notice that these relationships still bear on test meaning but also illuminate the contribution of the test in the decision context. In purely scientific work, it is appropriate to narrow one's questions, focusing on the meaning of a test and its adequacy as an indicator of the construct. However, in practical testing, hypothesized relationships with educational or job outcomes become a part of test score interpretation. Instead of a statistical correlation tacked on at the end, an expanded construct validity framework requires a theory of domain performance and a means for evaluating the link between predictor and criterion measures (see Messick, 1989).

Correlational and Experimental Evidence

Although correlational data have been the predominant mode for collecting validation evidence, experimental research should be undertaken when it is more appropriate to the hypothesized relations. As Landy (1986) has expressed, psychologists have always had a substantial repertoire of methods for investigating scientific hypotheses, and these methods are the same ones that should be brought to bear on validity questions. For example, suppose that a particular cognitive process is expected, on theoretical grounds, to account for performance on one type of test task. Then controlled studies showing the predicted change in results when task fea-

tures are manipulated or improved performance following strategy training provide empirical evidence to support both the theory and the measurement.

Experiments are also often the appropriate means to test practical relations underlying test use. If coachability is posed as a threat to the validity of the SAT, then an experimental study might be called for to determine, first, whether coached examinees do better than a randomly equivalent group that is not coached, and, second, whether score gains detract from or enhance the prediction of subsequent college performance. If an assessment device is intended to target remedial instruction (and thereby improve learning), then it would be best to evaluate the validity of the assessments by comparing how much students learn with and without the diagnostic assessment. Admittedly, this kind of study cannot distinguish, in a precise sense, between interventions that fail because of bad diagnosis and those that fail because of an ineffective treatment. Nevertheless, a test cannot claim to ensure a student of better instruction if it has never been shown to do so.

Plausible Rival Hypotheses

Good theory testing requires more than gathering supporting evidence. It requires exposing desired interpretations to counterexplanations and designing studies in such a way that competing interpretations can be evaluated fairly. Cronbach (1989) confesses that the early standards (American Psychological Association, 1954) had a "confirmationist" bias that still dogs much of current practice. Test authors tend to marshall evidence to satisfy requirements for reported validity data. They do not energetically plan studies that might discredit their products. Campbell (1957) introduced the notion of "plausible rival hypotheses" as a practical means to challenge research conclusions. Rival hypotheses derive from Popper's (1962) notion of falsification, which holds that a theory is not creditable until it has survived serious efforts to disconfirm it. Although Popper's efforts to create a rigorous system of falsification are undermined by the same problems that beset positivism, the idea of giving priority to fair trials of plausible competing explanations is still central to serious validation.

Suppose a test is claimed to measure aptitude for subsequent learning. A competing claim might be that the test is a good predictor only when all students have been exposed to the same curriculum. What if the apparent relationship is due to traditional ways of sequencing instruction rather than a real hierarchy of prerequisite skills? This might occur, for example, if a biology test were used to make placements for high school physics classes or when recognizing letters is used as a screening device

for school entry. Some of the best current research on test validity has been initiated to address issues of test or item bias—are group differences due to construct-relevant or construct-irrelevant test difficulty? Such studies reflect an effort to take criticisms of tests seriously and to look for evidence that might support the critic. For example, a test intended as a measure of reading comprehension is shown to be confounded by test wiseness or background knowledge if reasonably good scores can be earned by examinees who answer the questions without seeing the passages. What if a measure of mathematical reasoning has a substantial speed effect (i.e., an appreciable number of examinees do not finish)? Then the validity of the test for certain uses would be questionable unless it can be shown that speed in solving this kind of problem is essential to job or college work. And it will not do merely to show that speed of responding is correlated with quality of response on an untimed test. The question is whether (with the effects of unequal variances taken into account) the speeded or unspeeded version of the test is the better measure of what workers or college students actually need to know in the context for which they are being selected.

The series of studies mentioned earlier created to investigate gender differences on advanced placement examinations is an example of competing explanations being generated and assessed. When multiple-choice items appear to favor men and essay questions favor women, the charge of test bias arises. But bias in which direction? Plausible hypotheses were thought of and pursued by a team of researchers at ETS. For example, Breland (1991) used alternative scoring procedures on the U.S. history and European history essay portions and collateral data sources to determine whether any of the following factors could account for the relative advantage of female examinees: English composition quality, historical content, responsiveness, factual errors, handwriting quality, neatness, and words written. English composition quality and SAT Verbal scores appeared to be the best candidates for explaining why women earned higher essay scores than men when the two groups were matched on multiple-choice history performance. More work is needed, however, to determine whether writing ability is a valid or invalid influence on scores. It may be impossible to organize one's thoughts, identify key points in an argument, and provide supporting detail without thorough knowledge of historical content. Women's advantage on essay tests is only a distortion if by some artifice of good writing examinees can appear to know more than they do (or if men are correspondingly hindered).

Later, in attempting to think more about validation using the perspective of evaluation, we can look to various audiences or constituencies affected by test use to generate rival hypotheses.

Convergent and Discriminant Validity Evidence

The convergent and discriminant validity framework does not add new insights beyond those implied in the foregoing sections. The conceptual net or construct framework must first include external correlations to other indicators of the same construct for confirmation of meaning. Plausible rival constructs as explanations for test performance must also be included in the external net. Campbell and Fiske (1959) formalized these requirements in the multitrait-multimethod matrix, which has become a familiar tool in construct validation research.

A multitrait-multimethod matrix "presents all of the intercorrelations resulting when each of several traits is measured by each of several methods" (Campbell & Fiske, 1959, p. 81). The matrix for a particular study must be developed logically, on the basis of the construct theory, and then evaluated empirically. The multitrait entries require the measurement of several related, but presumed to be distinct, traits. The multimethod entries require that each of these traits (i.e., constructs) be measured by different methods. Evaluation of the matrix results follows from the logic of the construct theory. Several independently measured indicators of the same construct should be more highly correlated than are measures of different constructs. If the same method leads to strong correlations across constructs, then method-specific variance confounds these measures. If the correlations between two constructs are as high as those within each construct, then the two constructs lack "discriminant" validity.

Halo effects are classic examples of method variance that undermines the validity of rating scales. Campbell and Fiske (1959) cited an instance of peer ratings of "effort" versus "intelligence." Because peer ratings of the two constructs correlated .66 and did not show a differential pattern of correlations with independent measures of the two characteristics, peer ratings could not be regarded as valid measures of the constructs. In the controversial arena of school readiness testing, the Gesell School Readiness Screening Test is claimed to measure "developmental maturity" or "behavioral age." However, Kaufman (1971) found that the Gesell test correlated just as highly with the Lorge-Thorndike Intelligence Tests (r = .61) as with a test of Piagetian developmental tasks (r = .64), suggesting that there might not be discriminant validity for the two constructs of developmental maturity and intelligence. Lack of discriminant validity is suggested further by a conceptual analysis showing identical or highly similar tasks on the Gesell and preschool intelligence tests (Shepard & Graue, in press).

There is a tendency for the application of the multitrait-multimethod approach to become routinized, just as thoughtlessness has afflicted other

brands of validity. Messick (1989) warned against cookbook implementation of the multitrait-multimethod matrix as if any assortment of traits will do. The results of such analyses will not provide compelling evidence of construct validity unless the most provocative issues challenging test use find a place in the validity investigation. Similarly, the interpretation of results cannot be reduced to a set of algorithmic decisions (e.g., the principal component of the convergent correlations is *not* the true construct).

MESSICK'S UNIFIED THEORY: THE INTEGRATION OF TEST USE, VALUES, AND CONSEQUENCES

For two decades Cronbach's (1971) chapter, "Test Validity," in the second edition of *Educational Measurement* was widely influential among students and specialists in educational measurement. Messick's (1989) chapter in the third edition replaces it as the most cited, authoritative reference on the topic. Taken as a whole, Messick's extensive treatise accomplishes two purposes: (a) It cements the consensus that construct validity is the one unifying conception of validity, and (b) it extends the boundaries of validity beyond test score meaning to include relevance and utility, value implications, and social consequences. Measurement specialists who have just caught up to the requirements for both content and criterion-related validity evidence as part of construct validation may think that Messick has raised the stakes yet again. However, Messick's demand that validity support both the inferences and actions based on test scores simply repeats Cronbach's (1971) consideration of validation efforts needed to support either a descriptive or decision-oriented interpretation of a test. Similarly, Messick's consideration of values and consequences merely makes explicit the need to consider hidden assumptions and implicit claims about what test use will accomplish.

Messick (1989) presented his unified but faceted validity framework via the fourfold table shown in Figure 1 (Table 2.1 in the original text).

> A unified validity framework . . . may be constructed by distinguishing two interconnected facets of the unitary validity concept. One facet is the source of justification of the testing, being based on appraisal of either evidence or consequence. The other facet is the function or outcome of the testing, being either interpretation or use. If the facet for source of justification (that is, either an evidential basis or a consequential basis) is crossed with the facet for function or outcome of the testing (that is, either test interpretation or test use), we obtain a four-fold classification as in Table 2.1 (Messick, 1980).
>
> As indicated in Table 2.1, the evidential basis of test interpretation is construct validity. The evidential basis of test use is also construct validity, but as buttressed by evidence for the relevance of the test to the specific applied purpose and for the utility of the test in the applied setting. The consequential basis of test interpretation is the appraisal of the value implications of the construct label, of the theory underlying test interpretation, and of the

FIGURE 1
Messick's Facets of Validity Framework

	TEST INTERPRETATION	TEST USE
EVIDENTIAL BASIS	Construct validity	Construct validity + Relevance/utility
CONSEQUENTIAL BASIS	Value implications	Social consequences

Note. From "Validity," by S. Messick, in *Educational Measurement* (3rd ed., p. 20) edited by R. L. Linn, 1989, New York: American Council on Education and Macmillan. Copyright 1989 by the American Council on Education and Macmillan. Reprinted by permission.

ideologies in which the theory is embedded. A central issue is whether or not the theoretical implications and the value implications of the test interpretation are commensurate, because value implications are not ancillary but, rather, integral to score meaning. Finally, the consequential basis of test use is the appraisal of both potential and actual social consequences of the applied testing. (p. 20)

Messick's (1989) treatment of the first cell, labeled construct validity, includes a more extensive discussion of the same conceptual and analytic issues presented in the preceding section of this chapter. Traditional construct validity investigations focus on evidence needed to support (and challenge) the theory underlying test score interpretation. Messick's addition of the consequential basis of test interpretation (in the lower-left cell of the figure) requires that we also explicitly address the value assumptions implied by the concept labels and theoretical framework selected to guide the validity investigation. For example, the same behavioral indicator has quite different interpretive meaning depending on whether it is labeled as a measure of "flexibility versus rigidity" or "confusion versus consistency" (Messick, 1989, p. 60), each of which implies a different schema for empirical evaluation.

Test uses are obviously derived from value positions that are amenable to political debate, as, for example, when meritocratic or egalitarian principles are the basis for allocating educational opportunities. It is less widely recognized, however, that much scientific research is also value directed, to a greater or lesser degree. Value assumptions shape how research questions are framed, what data are gathered, and how results are interpreted. It is these perspectives, which influence scientific inquiry, that should be acknowledged in the validity framework. The point is to make assumptions explicit so that competing interpretations have an equal opportunity to shape the investigation. Our research (Smith & Shepard, 1988) on kindergarten readiness provides a perfect illustration of how construct labels can smuggle in whole theories without test users being aware of the choices they have made. When a test is said to measure "developmental maturity," test users are encouraged to envision a trait that is strongly influenced by biology. Given the label, parents, teachers, and policymakers find it reasonable to accept the arguments from the authors of the Gesell School Readiness Screening Test that the best treatment for low-scoring children is to wait for biological maturity to unfold (Gesell Institute of Human Development, 1982). When Shepard and Graue (in press) note, however, that the Gesell test closely resembles IQ tests, then questions are raised about the influence of prior learning opportunities on this measure of supposedly unalterable developmental age. Furthermore, the decision to wait as the treatment of choice, rather than intervene with appropriate learning experiences, cries out for critical evaluation.

Having considered the evidential and value issues affecting test score meaning, Messick (1989) turned to the evidential basis of test use. This cell in the framework represents a set of issues that have become quite familiar as the concept of construct validity has evolved—so familiar, in fact, that in preceding sections of the chapter I have used examples indiscriminantly of empirical evidence that pertains to both test interpretation and test use. Here is where Messick, however, systematically addresses questions about construct meaning and relationships that support an applied test use. He calls for construct validation of outcome criteria themselves, including their relevance, representativeness, and multidimensionality. Mere statistical correlations are insufficient to justify test use without analyses to show that shared variance is construct relevant and not due to some kind of shared bias. Under this rubric, he also considers the weighing of two types of errors in selection contexts and the generalizability of predictive correlations across populations and settings. Examples of the many different substantive studies cited by Messick in this category include documentation of the eroding correlation between college selection tests and grades as students self-select into departments with different grading standards (Willingham, 1985), two-step validation studies needed to show that professional knowledge is related to client outcomes and that a certification examination measures that professional knowledge (Kane, 1986), and the unsatisfying search for evidence to support the effectiveness of the Test of Standard Written English as a placement device (Breland, 1977).

In the last cell of his table, Messick (1989) considers the consequences of test use. Although this set of issues may seem new to many measurement specialists, it was implied by traditional conceptions of validity. In the first edition of *Educational Measurement*, Cureton (1951) stated that

"the essential question of test validity is how well a test does the job it was employed to do" (p. 621). The question of intended effects from testing was part of the validity picture as early as Scott (1917). Messick adds the consideration of unintended effects, but I submit that pursuing unintended effects is a logical extension of Campbell's (1960) inclusion of rival hypotheses when framing validity evaluations. The kinds of studies Messick has in mind to study consequences follow from the intended purposes of the tests. Does a credit-by-examination program improve social mobility and student access to higher education? Does the inclusion of a writing sample in admissions tests increase attention to the teaching of writing? Do data from the NAEP inform national education policy? Adverse consequences, usually unintended, must also be "weighed in appraising the functional worth of the testing" (Messick, 1989, p. 85). Examples of negative side effects that undermine test validity include adverse impact, "especially if irrelevant sources of test and criterion variance are likely culprits" (p. 85), and possible distortion of teaching and learning that might follow from heavy reliance on multiple-choice achievement tests.

Messick's (1989) framework identifies the full set of questions implied by a unified theory of validity. It is consistent with Cronbach's (1980) perspective:

> We might once have identified validation with a single question, What does the instrument measure? That question will not have an objective, universal answer. A yet more judgmental question now takes on equal importance: And why should *that* be measured?

Old conceptions of validity were analogous to truth in labeling standards. A more apt metaphor for current validity requirements is the Federal Drug Administration's standards for research on a new drug. The scientist is responsible for evaluating both the theoretically anticipated effects and side effects of any product before it is rendered "safe and effective" and released for wide-scale use.

REFORMULATING MESSICK'S THEORY: EVALUATION ARGUMENT AS THE CONSTRUCT VALIDITY OF TEST USE

In this section I quarrel with Messick, but not because I disagree with the substance of his arguments. Messick has aptly reported the current evolved state of validity theory, which has grown not by abstruse theoretical inventions but in response to pragmatic flaws in earlier conceptualizations. My disquiet is caused by the faceted presentation of his fourfold table that, I fear, invites a new segmentation of validity requirements. My concerns are as follows: (a) The faceted presentation allows the impression that values are distinct from a scientific evaluation of test score meaning. (b) By locating construct validity in the first cell and then reinvoking it in subsequent cells, it is not clear whether the term names the whole or the part. I argue for the larger, more encompassing view of construct validity. (c) The complexity of Messick's analysis does not help to identify which validity questions are essential to support a test use.

The positivistic, rigid fact-value distinction is no longer defensible in contemporary philosophy of science. The separate rows in Messick's table, however, make it appear that one would first resolve "scientific" questions of test score meaning and then proceed to consider value issues. Of course, this is not Messick's intention. The examples and arguments he advances are all instances of how value perspectives influence construct hypotheses and counterinterpretations that must be entertained as part of the initial delineation of a nomological network. Messick (1989) acknowledges that "scientific observations are theory-laden and theories are value-laden" (p. 62). He quotes Howe (1985) to the effect that social science is "doubly value-laden." Our epistemic principles—standards of predictive accuracy, internal coherence, and the like, used to accept or reject hypotheses—represent value choices, and the concepts we study are evaluative of human behavior. Nevertheless, when Messick lapses into discourse that separates "value implications" from "substantive or trait implications" (p. 63), he plays into the hands of researchers who deny that their construct definition or predictive equation follows from value choices. This should not be read to mean that scientific facts are indistinguishable from value judgments. Although scientific inquiry is distinct from politics and moral philosophy at the extremes, the concern here is with value perspectives that are entwined with scientific investigations. For example, as discussed later in the case of the GATB, an allegiance to individual merit principles can lead one to give greater weight to the test than to criterion performance when choosing a selection model.

My concern also pertains to the sequential segmentation of validity. This arrangement gives researchers tacit permission to leave out the very issues that Messick has highlighted because the categories of use and consequence appear to be tacked on to "scientific" validity, which remains sequestered in the first cell. Messick suggests that his conceptual framework be translated as a "progressive matrix." Construct validity is intended to appear in every cell with something more added each time. For example,

> the lower right cell—in recognition of the fact that the weighing of social consequences both presumes and contributes to evidence of score meaning, relevance, utility, and values—would now include construct validity, relevance, and utility as well as social and value consequences. (Messick, 1989, p. 21)

Messick has implicitly equated construct validity with a narrow definition

of score meaning, whereas I would equate it with the full set of demands implied by all four cells, which all involve score meaning. Intended effects entertained in the last cell are integrally a part of test meaning in applied contexts.

To be fair, Messick (1989) agrees that test interpretation and test use are closely intertwined, as are validity evidence and value consequences. He adopted a faceted framework "because of a conviction that the commonality afforded by construct validity will in the long run prove more powerful and integrative than any operative distinctions among the facets" (p. 21). Messick also concurs that "construct validity may ultimately be taken as the whole of validity in the final analysis" (p. 21). Thus, we are in dispute only about how these issues should be communicated, by theorists to measurement practitioners and ultimately by the field to external audiences.

These points about the boundaries of validity and what it should be called are controversial and are likely to be a focus of debate in extending the current consensus. Most theorists agree that validation includes the whole of Messick's framework, not only the first box. But can all of the implied questions be subsumed under construct validation without degrading its scientific meaning? Wiley (1991) takes a conservative position, focusing on the psychological processes intended to be measured rather than test use. Wiley (1991) acknowledges, however, that interpretation of two correlated traits influencing test scores, such as reading comprehension and vocabulary knowledge, would vary depending on test use. Moss (1992) addressed directly the problem of overburdening the concept of validity. Originally, in their article on bias in test use, Cole and Moss (1989) had reserved the term *validity* for only the interpretive component of their framework in contrast to the extra validity component, which focused on the consequences of test use. Moss (1992) noted, however, that "since then, we have expanded our definition of validity to include the consequential component, in part, because we were concerned that excluding consideration of consequences from the definition of validity risks diminishing its importance" (p. 235).

In my view, validity investigations cannot resolve questions that are purely value choices (e.g., should all high school students be given an academic curriculum versus being tracked into vocational and college preparation programs?). However, to the extent that contending constituencies make competing claims about what a test measures, about the nature of its relations to subsequent performance in school or on the job, or about the effects of testing, these value-laden questions are integral to a validity evaluation. For example, the question as to whether students are helped or hurt as a result of a test-based remedial placement is amenable to scientific investigation.

There is also a pragmatic concern that Messick's sequential approach may misdirect the conceptualization of theoretical frameworks intended to guide validity evaluations. A progressive model assumes that the first cell's construct framework for a measure of learning disabilities would be the same whether the test was used for research on the heritability of dyslexia or to place children in school resource rooms. For the applied purpose, one would merely add to the theory-based relationships. Imagine a web or schema depicting all of the meaning-confirming relationships and then add many more arrows to include predictive correlations, treatment effects, unintended side effects such as missing regular class instruction, and so forth. Perhaps this approach would be acceptable if researchers had infinite resources to test exhaustively all possible theoretical and practical relationships. For example, confirmation of heritability patterns does add minutely to the validity of the learning disabilities construct for school policy purposes (although such cases represent a tiny fraction of the school-defined catchall category). However, given the number of studies that can reasonably be undertaken, I argue that measurement specialists need a more straightforward means to prioritize validity questions. If a test is proposed for a particular use, a specific network of interrelations should be drawn focused on the proposed use. The kinds of experimental studies, for example, that bear on test meaning should be designed specifically to evaluate the effectiveness of a proposed test use.

Finally, the complexity of Messick's model and chapter creates the same difficulty as nearly every other treatise on construct validity before his. Each emphasizes that construct validation is a never-ending process, because there are so many hypotheses to be tested across so many settings and populations and because the generalizability of findings decays over time. While the never-concluding nature of construct validation is a truism, the sense that the task is insurmountable allows practitioners to think that a little bit of evidence of whatever type will suffice. Current standards do little to help prioritize validity questions. Validity standards are not organized in a coherent conceptual framework. Therefore, they do not help answer the question "How much evidence is enough?" nor do they clarify that the stringency of evidential demands should vary as a function of potential consequences.[1]

My proposal, alluded to previously, is that validity evaluations be organized in response to the question "What does the testing practice claim to do?" Additional questions are implied: What are the arguments for and against the intended aims of the test? and What does the test do in the system other than what it claims, for good or bad? All of Messick's issues should be sorted through at once, with consequences as equal contenders alongside domain representativeness as candidates for what *must*

be assessed in order to defend test use. If a test is proposed only for research purposes, then it is possible that the validity framework would invoke only the first column of Messick's table (i.e., the construct meaning and value implications of test interpretations). For example, a researcher might be interested in devising a model and concomitant measures of language development for children ages 2–5. The measures might be used to test for common stages of development across cultures and language communities. From a research perspective, an important validity question would be whether a test score could be shown to have similar properties in cultural contexts with very different norms for verbal fluency. However, as soon as such a language measure is proposed for use in screening children for school entry or for identifying potential learning handicaps, the locus of the validation inquiry shifts. If the purpose is handicap identification, the most important questions pertain to the predictive accuracy of the measure in the lower range of the distribution, the consequences of interventions based on test scores, and the relationship between socioeconomic status and construct-irrelevant variance in the test. Findings from the research-oriented studies are also pertinent to construct validation of a test use, but they cannot be substituted for these central applied questions.

This approach borrows most closely from Cronbach's (1988, 1989) conception of validation as evaluation argument. Many experts consider validation of test use to be a process of evaluation (Guion, 1980; Messick, 1980), so Cronbach (1988) turns to insights gained from program evaluation regarding the nature of argument and weighing of evidence, the posing of contending validity questions, and the responsibility to represent the various audiences affected by a program. "Validation speaks to a diverse and potentially critical audience; therefore, the argument must link concepts, evidence, social and personal consequences, and values" (p. 4). In *Designing Evaluations of Educational and Social Programs,* Cronbach (1982) emphasized that evaluators do not have the luxury of setting aside issues as basic researchers do when addressing constrained questions in a limitless series. The evaluator must "illuminate the whole program in a comparatively short period of time" (p. 7). Because this is impossible to do exhaustively, evaluation design involves identifying the most relevant questions and deciding what emphasis should be given to each.

In the context of test evaluation, Cronbach (1988, 1989) reminds us that construct validation cannot produce definitive conclusions and cannot ever be finished. He suggests, however, how test evaluators might set priorities. After the collection of relevant questions, priority should be assigned to potential lines of inquiry depending on prior uncertainty, information yield, cost, and leverage. Leverage refers to the import of study information for achieving consensus in the relevant audience "(consensus regarding appropriate use of the test, or consensus that it should not be used)" (Cronbach, 1989, p. 165). "After weighing these criteria, the evaluator will probably choose a few questions for intensive research, with other questions covered incidentally by inexpensive side-studies, or not at all. This prioritizing steers the test evaluator away from *Dragnet* empiricism" (Cronbach, 1989, p. 165).

Kane (1992) has extended Cronbach's (1988) recommendation, borrowing from the literature on practical reasoning (Toulmin, Rieke, & Janik, 1979) and evaluation argument (House, 1980). Kane conceptualizes validation as the evaluation of interpretive argument. (Note that when Kane discusses a test-score interpretation, he includes test uses as interpretations.)

> To *validate a test-score interpretation* is to support the plausibility of the corresponding interpretive argument with appropriate evidence. The argument-based approach to validation adopts the interpretive argument as the framework for collecting and presenting validity evidence and seeks to provide convincing evidence for its inferences and assumptions, especially its most questionable assumptions. One (a) decides on the statements and decisions to be based on the test scores, (b) specifies the inferences and assumptions leading from the test scores to these statements and decisions, (c) identifies potential competing interpretations, and (d) seeks evidence supporting the inferences and assumptions in the proposed interpretive argument and refuting potential counterarguments. (Kane, 1992, p. 527)

According to Kane (1992), the criteria for evaluating interpretive arguments are the same as those for evaluating any practical argument: (a) The argument must be clearly stated so that we know what is being claimed; (b) the argument must be coherent in the sense that conclusions follow reasonably from assumptions; and (c) assumptions should be plausible or supported by evidence, which includes investigating plausible counterarguments. A specific example constructed by Kane (1992) illustrates how an interpretive argument framework helps to focus a validity investigation specifically on the intended test use—in this case, the use of an algebra placement test to assign college students to either calculus or a remedial algebra course. Even without details of how each assumption is to be tested, the list of assumptions composing the argument shows us how the argument focuses what should be studied.

> Assumption 1: Certain algebraic skills are prerequisites for the calculus course in the sense that these skills are used extensively in the calculus course. (p. 531)
>
> Assumption 2: The content domain of the placement test matches the target domain of algebraic skills used in the calculus course. (p. 531)
>
> Assumption 3: Scores on the test are generalizable across samples of items, scores, occasions. (p. 531)

Assumption 4: There are no sources of systematic error that would bias the interpretation of the test scores as measures of skill in algebra. (p. 531)

Assumption 5: An appropriate measure of success in the calculus course is available. (p. 531)

Assumption 6: The remedial course is effective in teaching the algebraic skills used in the calculus course. (p. 532)

Assumption 7: Students with a high level of skill in algebra would not substantially improve these skills in the remedial course and therefore would not substantially improve their chances of success in the calculus course. (p. 532)

The argument-based approach to validity, exemplified by this outline, does not add to or subtract from, in any important way, what is implied by the domain of construct validation. It does, however, organize our thinking about important questions and identify priorities. In this specific case, for example, it helps us see that multiple studies, all supporting Assumptions 1–3, cannot compensate for lack of evidence supporting Assumptions 5 and 6.

VALIDITY CASES

If we take the demands of construct validation seriously and resolve to use the evaluation argument approach, what would applied validation studies look like? Kane (1992) has offered one hypothetical example. In this section of the chapter, I consider four current testing applications. These examples are intended to illustrate, with real cases, how a set of essential validity questions might be outlined.

The Scholastic Aptitude Test (SAT) Used to Make College Selection Decisions

Any validity evaluation must start by identifying not only the test but its intended use. In the case of the SAT, the test was originally designed for college selection and is still used in that way, although the decision context has changed considerably. The stated purpose of the SAT is to identify students who will be successful in college. First and foremost, then, the test must show predictive correlations with performance in college. The College Board provides a validity study service to help colleges analyze the predictive accuracy of the SAT for their respective institutions. Averaged over 685 institutions, the mean correlation of composite SAT with freshman grades was .42, with a range (10th to 90th percentile of correlations) from .27 to .57. Corresponding figures for the high school record correlated with freshman grades were a mean correlation of .48 with a range of .31 to .64 (Donlon, 1984). A research literature also exists to support the assumption that SAT and high school grades continue to

maintain the same degree of prediction if other criteria not so conveniently available, such as cumulative college grade point average (GPA) or post-freshman GPA, are used (Wilson, 1983).

Correlations do not explain why a relationship exists, however. The specific content of the SAT is defended on the grounds that generic measures of verbal and mathematical skills and reasoning, developed during the years of schooling, are indicative of or prerequisite to students' ability to do academic work in college. According to the technical manual, scholastic aptitude does not refer to an innate ability; the test is aimed at a concept of aptitude as "general readiness" for college studies (Donlon, 1984).

A myriad of studies exist showing, by content analysis, correlations with other measures (including high school grades), factor analysis, and so forth, that the two parts of the SAT are reasonably good measures of what they claim to measure. Interview studies have been used to examine the cognitive processes underlying item responses; the level of vocabulary and reading passage difficulty has been connected empirically to high school and college text material; and so forth. I do not mean to trivialize this body of work by giving it little attention here. In fact, one of the difficulties in thinking about any alternative to the SAT is the problem of replacing a known with an unknown and the Herculean task of beginning to amass new validity data commensurate with that supporting the current test. (Note that even the alternative of doing without the SAT cannot be evaluated accurately without collecting new data because, for example, the integrity of high school grades and their relation to college performance could change appreciably in an environment without the SAT.)

However, just because the SAT "taps" constructs that can reasonably be labeled as developed verbal and mathematical reasoning abilities, this does not mean that test scores are perfect instantiations of those traits, nor does it mean that there is a perfect correspondence between what is measured by the test and academic skills needed to do well in college. (*Taps* is a well-chosen verb that has been used historically to talk about construct validity. Appropriately, it connotes that a test "touches" or "gets at" the intended construct but does not represent it fully or exhaust its meaning.) Different ways of conceptualizing test content might look as good by all of the above logical and empirical analyses but have different implications for other consequences of testing such as adverse impact, sensitivity to coaching, or effects on the high school curriculum. Therefore, questions about test content—such as the decision to measure generic skills rather than curriculum-specific content or to use antonym items or not—hinge on studies of testing effects that go beyond predictive correlations and convergent and internal validity evidence. Indeed, these considerations have strongly influenced planned revisions in the SAT.

As Cronbach (1989) suggested, a good evaluator generates questions to be investigated by listening both to affirmative claims and to program critics. The most salient complaints against the SAT are that it is biased against minority groups and women, that scores can be raised by expensive coaching, and that studying for the SAT distorts the high school curriculum by emphasizing vocabulary drill and test-taking skills. Here I focus on the question of adverse impact—mean scores for women and disadvantaged minority groups are lower than for White males.

In response to the bias concern, researchers at ETS and elsewhere have conducted countless studies examining differential predictive validity and relative difficulty of test items for different groups. A gross summary of this research is that the SAT predicts better (i.e., has a steeper regression slope) for women than for men and better for Whites than for African Americans or Hispanics (Linn, 1982). The consequences of using total group regression equations are somewhat counterintuitive. Actual college grades are higher for women than predicted from their SAT scores. (Similar findings hold true for American College Testing [ACT] assessment scores.) In contrast, African Americans, on average, earn lower grades in college than expected from their SAT scores. Thus, the claim of bias in prediction is documented for women but not for minority groups. Internal studies of differential item functioning have shown few conclusive patterns, except that statistical flags occur more often for verbal than for mathematical items and more often for short, decontextualized verbal items than for reading comprehension, the latter finding possibly contributing to the decision to remove antonym items and increase reading comprehension items in the revised SAT.

Evidence showing that the test measures similarly for all groups does not answer adequately all of the questions about the validity of testing effects. Suppose several factors such as prior academic knowledge, motivation, and study skills account for success in college. As we see likewise in the next section dealing with employment testing, the decision to rely on the test alone or to give the test undue weight is unfair because it disadvantages applicants who are relatively low on the test but high on the other factors.

Crouse and Trusheim (1988), critics of the SAT, argued that the SAT should be eliminated or replaced with curriculum-specific achievement tests because of its adverse impact on African-American and low-income applicants and because it adds little to the accuracy of college selection decisions after high school grades. The SAT adds only about .06 to .08 to the multiple correlation with college grades. Claims about redundancy or the biasing effects of test use cannot be accurately evaluated using multiple regression equations, however, because these linear models do not reflect how test scores are actually used.

Studies reported by the National Research Council (NRC) Committee on Ability Testing (Wigdor & Garner, 1982) and by the College Board (Willingham & Breland, 1982) show that most colleges use grades and scores to make only a rough sorting of students into categories of likely admit, possibly admit, and likely reject. Other factors such as special musical or athletic talents, minority group status, alumni son or daughter, and geographical distribution then account for substantial departures from the initial sorting. For example, in the nine selective colleges studied by Willingham and Breland (1982), minority group status improved an applicant's chances of being accepted by 31% compared with the rate predicted from grades and SAT scores alone. Based on informal knowledge of selection processes, it is more plausible that admissions officers use scores and grades to select the most qualified minority candidates rather than using a strict equation to pit majority against minority candidates.

At one level, examination of these selection practices might provoke a debate between different philosophical positions. Should decisions be guided by meritocratic or other theories of social justice (e.g., Rawls, 1971)? At a more technical level, such decisions can be defended "scientifically" given that academic predictors are both incomplete and fallible predictors of success (multiple correlation of .55; Donlon, 1984). Furthermore, in highly selective colleges all of the admitted applicants have strong academic qualifications in keeping with the original purpose of test-based decisions. Therefore, one value perspective holds that colleges can reasonably select among qualified applicants using criteria aimed at other goals such as increasing the diversity of perspectives represented among their students. This value choice cannot be resolved within the validity framework but should be made explicit and examined for consequences as part of a validity investigation. Do admissions officers go too far in favoring qualified, but not the highest scoring, minority candidates? As a matter of fairness, should a similar selection model be used for qualified majority candidates? For example, Hofstee (1983) described a "compromise" selection model used in the Netherlands, such that candidates in different score categories entered a lottery with different probabilities of being selected. Candidates in the highest-score category had a high probability of being selected, but some candidates were also drawn by lot from qualified but lower scoring categories. There are many more questions to be answered here, but the point is that validity evaluations must address test use as it actually occurs and should consider the full set of valued outcomes.

If test scores are used to select the strongest candidates within majority and minority groups rather than to make comparative judgments between groups, they will not have an adverse impact despite mean group differences. However, this reasoning does not alleviate concerns about biasing

effects of test use in situations where scores are used in strict top-down fashion, such as the awarding of National Merit Scholarships (which have quotas by state but not by sex and racial groups) or qualifying rules of the National Collegiate Athletic Association. For example, the accusation that use of the SAT alone discriminates against women in the allocation of scholarships gains support when additional data on relevant variables (e.g., high school studiousness and attitudes about mathematics) account for the test's underestimation of women's college performance (Stricker, Rock, & Burton, 1991). In other words, important qualifications that contribute to women's success in college are missed by the test. (Note, however, that patterns of gender differences on the SAT are consistent across studies but tend to be relatively small in magnitude.)

These kinds of evaluative investigations motivated by claims and counterclaims reflect the principles of construct validation in its fullest sense. Evaluative arguments inform policy decisions about the use of test results—as when the College Board urges member institutions to avoid using cutoff scores. Arguments of this type should also be the basis for evaluating alternatives to the SAT. Some critics argue that high school grades are sufficient, pointing to colleges like Bowdoin that have eliminated the use of the SAT. A thoughtful evaluation of what it would mean to eliminate selection tests should consider effects under different contingencies, especially the degree of selectivity of the particular college and how widespread the practice is expected to become. Crouse and Trusheim (1988) reported data gathered by Schaffner at Bowdoin showing that students who withheld their SAT scores earned lower grades in college than did students reporting SAT scores. However, SAT withholders also had lower high school ranks, suggesting that the college was simply willing to lower its admissions standards. This finding is consistent with the conclusion of the NRC Committee on Ability Testing that test scores play an important role in admissions decisions for only the most selective institutions. In contrast, Elliott and Strenta (1990) analyzed data for enrolled students at Dartmouth, a highly selective college, and found that high school rank plus SAT scores not only improved the prediction of freshman GPA over a selection rule using only high school rank but substantially increased the number of academic commendations and reduced the number of academic probations. If all colleges were to eliminate the use of tests like the SAT and ACT, then effects on high school grades would have to be studied. In recent years, for example, the University of California system began recalculating applicant GPAs, giving extra weight to honors courses because admissions committees believed that students were avoiding challenging courses to protect their GPAs. Problems of this type might be expected to increase if high school grades become the only basis for judging academic preparation.

Crouse and Trusheim (1988) recommended that new achievement tests be used rather than the SAT, expecting that such a change would reduce adverse impact and stimulate students to study a more challenging curriculum in high school. Both of these claims, of course, become the basis for the next round of validity investigations. Early findings from Great Britain (Nuttall & Goldstein, 1990) show that new performance assessments increase the gap between male and female students and between majority and minority groups. Although replacing generic academic aptitude tests with curriculum-specific tests is not the same as the change from multiple-choice to performance examinations, the latter type of test in each case is more influenced by the quality of schooling. If members of minority groups have less access to high-quality instruction demonstrated to improve performance on syllabus-driven examinations, then such tests could actually have greater adverse impact than present tests. Of course, Crouse and Trusheim (1988) did not propose that achievement tests replace the SAT without critical examination. In fact, they recommended that a blue-ribbon panel be convened to study the SAT and admissions testing. Ultimately, as in any good evaluation, contending program choices will have to be judged on multiple criteria: accuracy in predicting college success (itself measured by multiple criteria), differential performance and validity by race and sex, and indirect influence on what students study in high school.

The General Aptitude Test Battery (GATB) Used to Make Employee Referral and Selection Decisions

When Cronbach (1989) gave advice to prospective test evaluators, he doubted whether anyone would be charged with "precisely the job thus defined" (p. 164). In my view, however, validity evaluation, broadly defined, is the role that has been taken up by three different National Academy of Sciences committees charged with evaluating testing applications. The three committees, impaneled by NRC (which is the principal operating agency of the academy), include the Committee on Ability Testing mentioned in the previous discussion of the SAT, the Panel on Selection and Placement of Students in Programs for the Mentally Retarded considered in the next section, and the Committee on the General Aptitude Test Battery considered here.

These examples are important because they address the validity of tests for an applied use. They stand in contrast to integrative reviews, such as Wylie's (1974, 1979) treatment of self-concept, that address the adequacy of measurement and the meaning of the construct in a research context. By showcasing these examples, I do not mean to suggest that validity evaluations should be the exclusive province of authoritative, blue-ribbon panels. For example, issues surrounding kindergarten readiness testing,

considered as part of the last case, have been enjoined in traditional research journals. National Academy of Sciences studies are useful, however, because they illustrate the importance of multiperspective inquiry; they also demonstrate the sheer magnitude of the evaluation task, which can nonetheless produce insightful (if not definitive) reports in less than a lifetime.

For 40 years, the U.S. Employment Service (USES) has used the GATB to counsel job seekers and to make referrals to employers. In recent years the Department of Labor, supported by research conducted by Hunter and Schmidt (1982; Schmidt & Hunter, 1977; U.S. Department of Labor, 1983a, 1983b), has sought to extend the use of a "validity generalization" (VG) GATB referral system. (Validity generalization is based on a meta-analysis of predictive correlations from studies of 500 jobs and is used to "generalize" the validity of the GATB to all 12,000 jobs in the U.S. economy.) One feature of the new GATB referral system was "within-group scoring" whereby percentile-rank scores were computed separately for three groups of examinees: African-American, Hispanic, and other. Given the peculiarities of legal constraints following from the uniform guidelines (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice, 1978), which require employers to show the validity of tests only if they produce an adverse impact, this procedure protected the referral system from legal challenge. The Department of Justice, however, charged that the system was unconstitutional on grounds of reverse discrimination. The Department of Labor requested the assistance of the National Academy of Sciences, which convened a committee of experts "to conduct a thoroughgoing *evaluation* [italics added] of the plan to use the GATB as the primary tool for deciding which applicants to refer to employers" (p. viii).

What follows is a brief summary of the book-length GATB committee report (Hartigan & Wigdor, 1989). It serves as an example of evaluation argument because it addresses both traditional psychometric standards and the most salient questions implied by the intended test use, including validity generalization and race-conscious scoring. In response to the organizing evaluation questions, the committee not only collected the available evidence but conducted research of its own.

Is the psychometric quality of the GATB adequate? It has respectable evidence of reliability and predictive validity and is comparable to other test batteries such as the Armed Services Vocational Aptitude Battery. Convergent validity studies with other test batteries support the interpretation of the cognitive subtests, but evidence is not so strong for the perceptual and psychomotor subtests. The committee cited three areas of technical deficiency, however, that would have to be remedied before

wide-scale use of the test could be considered. First, the GATB is highly speeded, which threatens the meaning of scores where speed is not a logical component of the construct and makes the test highly vulnerable to coaching (e.g., filling in the last third of the test at random would raise one's score substantially). Second, its norms are based on convenience samples rather than nationally representative and up-to-date samples. Third, because there are only four operational forms of the GATB, it would be highly vulnerable to test security violations and practice effects. Note that in practical settings the "validity" of a test is called into question if score results are susceptible to cheating or to coaching effects that do not produce a commensurate gain in criterion performance.

How well does the GATB predict job success? Does the predictive validity of the GATB generalize to most or all jobs? Averaged over 750 studies, "correlations of GATB-based predictors with supervisor ratings, after correction for sampling error, are in the range of .2 to .4" (Hartigan & Wigdor, 1989, p. 5). It is important to note that studies conducted before 1972 produced average correlations (corrected for both sampling error and criterion unreliability) of .35. Since that time, the average is only .25. The committee did not agree with all of the assumptions that had been used to make statistical corrections resulting in much higher predictive values reported in USES technical publications (U.S. Department of Labor, 1983a). The committee accepted the general thesis of validity generalization in the sense that "validity studies can be generalized to many jobs not actually studied, but we urge a cautious approach of generalizing validities only to appropriately similar jobs" (Hartigan & Wigdor, 1989, p. 8). However, the committee did not conclude that validity generalization could extend literally to all 12,000 jobs in the economy. It raised questions on both scientific and policy grounds about the use of a "single fallible test" (p. 8) as the central means for referring workers to jobs throughout USES.

Does the GATB predict less well for minority applicants? Is there scientific justification for adjusting minority test scores? Based on 78 studies, the committee found that the correlation between the GATB and supervisor ratings was .12 for African Americans, as compared with .19 for nonminorities. If the combined-groups regression line were used, however, test scores would slightly overpredict for African Americans rather than underpredict. The committee urged caution in concluding that the test was unbiased given that supervisor ratings were the sole criterion and that such ratings have not been shown to be unbiased themselves.

The GATB committee reviewed the extensive psychometric literature on fairness in selection. Given an imperfect relation between test and criterion, none of the proposed models provide an unambiguous technical definition of fairness. Because of the focus on test scores from which to

make decisions, the classic psychometric model has defined unbiased selection as equal job performance for a given test score. As outlined above, the GATB is not biased against minorities by this definition because it overestimates the job performance of minority test takers. The GATB committee, however, focused on the question of bias defined by the probability of being selected given equal job performance.

Given the less than perfect correlation between test and criterion, it can be shown that, for equal job performance, minority applicants have a smaller chance of being selected than do majority applicants.

> Majority workers do comparatively better on the test than they do on the job, and so benefit from errors of false acceptance. Minority workers at a given level of job performance have much less chance of being selected than majority workers at the same level of job performance, and thus are burdened with higher false-rejection rates. (p. 7)

This phenomenon is not the result of bias in the sense of measuring the test construct differently. It is straightforwardly the result of fallible measurement and unequal group means. Nonetheless, the committee concluded that "the disproportionate impact of selection error provides scientific grounds for the adjustment of minority scores so that able minority workers have approximately the same chances of referral as able majority workers" (p. 7). The committee recommended either the continued use of within-group scoring for referral, with reporting to employers of both within-group and total-group percentile scores, or the adoption of a performance-based scoring system that would vary the amount of score adjustment depending on the degree of predictive accuracy. The more predictive the test, the smaller the adjustment would be.

The GATB evaluation illustrates that the conduct of construct validation invokes all four cells of Messick's (1989) framework. The study addressed the convergent correlations supporting subtest interpretations and the value implications of choosing among selection models, as well as the issues of social justice ignored by statistical models, the strength of predictive correlations, the uncertainties introduced by possible criterion bias, and the consequences of using a test with only a .3 correlation to control referral to jobs in 1,800 employment offices. The committee listened to contending forces to identify key issues for investigation, especially regarding utility and consequences. As an example of validity argument, however, the committee's efforts were only partially successful. The Department of Labor was persuaded to curtail the extension of the VG-GATB system until technical problems with the test were remedied; however, Congress, in the Civil Rights Act of 1991 (Public Law 102–166), prohibited the use of score adjustments.

Tests Used for Educational Placement Decisions: Special Education and 2-Year Kindergartens

Each applied case makes it clear that the expanded conception of validity requires attention to both test score meaning and testing effects. Nowhere is the attention to effects more pronounced than in evaluation of test-based placement decisions. Kane (1992) has already provided an outline of the set of assumptions to be examined before the validity of a placement test can be defended. The most critical requirement is that placement tests show differential validity in the sense of aptitude-treatment interaction. Groups must be better off in their respective treatments than they would have been without the test-based placements. Two real cases involve the placement of children in classes for the mentally retarded and retention of children in 2-year kindergartens.

The National Academy of Sciences Panel on Selection and Placement of Students in Programs for the Mentally Retarded was convened to address the problem of overrepresentation of minority children and boys in special education classes. Competing "theories" or "constructions" motivated the inquiry. On the one hand, special education resources are presumed to be a benefit, and handicapped students are entitled to such services. On the other hand, the complaint against disproportionate placements reflected a concern that special education placements are stigmatizing and consign students to low-quality instruction. Thus, questions about the validity of placement decisions were entwined with questions about program effects.

In its comprehensive study (Heller, Holtzman, & Messick, 1982), the panel gave limited attention to the history and validity of IQ tests, which continue to play a major role in the identification of children as educable mentally retarded. Greater attention was given to the research evidence on the social and academic effects of special education self-contained placements. Although empirical findings were far from monolithic and unambiguous, the lack of clear evidence of benefit caused the panel to place the burden of proof on those who argued for placement in a segregated setting. Given the potential for social stigma and ineffective instruction, the panel recommended a two-stage assessment model. The first stage was to be a thoroughgoing examination of the child's learning environment to rule out poor instruction as the cause of referral to special education.

> Only after deficiencies in the learning environment have been ruled out, by documenting that the child fails to learn under reasonable alternative instructional approaches, should the child be exposed to the risks of stigma and misclassification inherent in referral and individual assessment. (Messick, 1984)

In the second stage, a comprehensive battery of assessments should be administered, including biomedical, behavioral, and academic skills tests as well as measures of cognitive functioning.

Although the panel reformulated its charge so that IQ tests were not at the center of the investigation, it should be clear how a validity evaluation of intelligence tests for the purpose of special education placement would be framed differently from one intended to evaluate the validity of intelligence tests used by a researcher for a very different purpose, such as assessing the long-term effects of metacognitive strategies training. In the view of the panel, IQ tests have limited practical utility for matching children's needs to instructional interventions. As noted by Messick (1984), it is not sufficient that the tests can predict who will perform poorly in the regular classroom.

> To justify separate placement based on IQ, it would be necessary to demonstrate that children with low IQs benefit from and require a different curriculum or type of instruction from that implementable in regular classes without adverse effects on the other students. (p. 6)

The issues of school readiness screening and 2-year kindergartens parallel those of special education placement. During the 1980s, the practice of retaining children in kindergarten increased dramatically. Shepard and Smith (1988) saw the increasing use of 2-year kindergarten programs as a response to escalating academic demands in kindergarten and first grade. The creation of 2-year programs was fostered by a set of beliefs widely shared among educators and parents: (a) that repeating kindergarten is an entirely benign intervention, (b) that both cognitive readiness and maturity are biological traits that should not be hurried, and (c) that the gift of an extra year will allow children to excel without being pushed (Smith & Shepard, 1988). These beliefs have not been critically examined and are not supported by research evidence.

A validity evaluation of readiness measures used in this context must address both traditional psychometric features of reliability and validity and differential predictive validity. The main findings from these investigations have been cited previously in this review. Tests such as the Gesell School Readiness Screening Test do not have adequate reliability or predictive accuracy to support their use in making decisions that seriously affect children's school careers. Controlled studies do not show any academic benefit for children in developmental kindergartens or pre-first grades compared with those measured to be unready who went directly on to first grade (Shepard, 1989). Contrary to the belief that children in kindergarten are too young to notice, a majority of parents reported that their children experienced some short-term or long-term trauma associated with retention (Shepard & Smith, 1989).

School readiness testing for the purpose of sending "unready" children home or assigning them to 2-year programs is an example of a huge testing enterprise largely unaffected by professional testing standards. Usually, the testing is done by classroom teachers who are unaware of the technical limitations of the tests. In a large percentage of school districts, measures originally devised for handicapped screening and classroom planning have been adopted instead to make school entry and kindergarten retention decisions (Gnezda & Bolig, 1988). Although readiness testing became widespread without much scrutiny, the set of issues implied by a construct validity framework has begun to be addressed in the literature on educational measurement and early childhood education, leading to policy reports such as the one published by the National Forum on the Future of Children and Families (Gnezda, Garduque, & Schultz, 1991).

CONCLUSION: IMPLICATIONS FOR THE 1990S TEST STANDARDS

A pointed way to summarize is to consider the implications of both current test theory and practical validity cases for the development of the next version of professional testing standards. Each decade since the 1950s has seen a revision of the standards. What should the 1990s version look like?

The consensus, already emergent before the 1985 standards, has been solidified. Construct validation is the one unifying and overarching framework for conceptualizing validity evaluations. Logical analysis of test content and empirical confirmation of hypothesized relationships are both essential to defending the validity of test interpretations; however, neither is sufficient alone.

The basic principles of construct validation were laid out by Cronbach and Meehl (1955). The concept has grown as the field has come to understand the nature of hypotheses, claims, and counterclaims that must be investigated to support applied test uses. Construct validation is still guided by a conceptual framework. (This framework was once called the nomological net; however, it is no longer assumed that underlying relationships have the power of enduring laws.) The conceptual framework portrays the theoretical relationships believed to connect test responses to a domain of performance and desired ends implied by the intended test use. In all but rarified research contexts, test uses have intended consequences that are an essential part of the validity framework. Given that theory testing must also include empirical evaluation of the most compelling rival hypotheses, construct validation entails a search for both alternative meanings and unintended consequences as well.

In the early years of testing, validity addressed the question "Does the test measure what it purports to measure?" A single correlation coeffi-

cient was often accepted as a sufficient answer. An appropriate metaphor might have been truth in labeling (i.e., does the test have the ingredients or meaning it says?). Today, a more appropriate question to guide validity investigations is "Does the test do what it claims to do?" A more contemporary analogy is the Federal Drug Administration's standards for testing a new drug: "Do the empirically demonstrated effects weighed against side effects warrant use of the test (drug)?"

We are witnessing a sea change in the field of psychological and educational testing. In 1980, only a few wise theorists (Cronbach, 1980; Guion, 1980; Messick, 1980) articulated these ideas. Now they have begun to have a pervasive effect on the field, as evidenced, for example, by the 1990 Personnel Testing Council Conference, "Construct Validity: Issues and Opportunities" (*Human Performance,* 1992). Still, there is a discouragingly high rate of test misuse, including wide-scale marketing of tests without validation for a particular use.

The 1985 standards, despite the affirmation that "validity is a unitary concept" (p. 9), are fragmented and enable test developers to pick and choose the standards they will consider and how rigorously they will meet them. For example, the standard that classification tests should show differential prediction for treatment categories (1.23, p. 18) is denoted of "secondary" importance. Yet if a placement test fails on this criterion, has it not failed in its central purpose? The very first validity standard requires that a rationale "should be provided to support the particular mix of evidence presented for the intended uses" (p. 13). The comment statement that follows the first standard, however, excuses a regression to old validity practices: "Whether one or more kinds of validity evidence are appropriate is a function of the particular question being asked and of the context and extent of previous evidence" (p. 13). Professional standards are necessarily the product of a political process. Although each set of standards is an advance over what came before, working committees and their critics tend to codify *what is* rather than go beyond present practice. Given the consensus regarding construct validation, we can expect less equivocal language in the next version of the standards regarding the separate paths to validity. It remains to be seen, however, whether a new consensus can be established that incorporates all of Messick's matrix.

To advance practice, authors of the new standards should try to create a coherent framework for prioritizing validity questions. If construct validation is seen as an exhaustive process that can be accomplished only over a 50-year period, test developers may be inclined to think that any validity information is good enough in the short run. Cronbach's (1989) conception of validation as evaluation argument helps to focus empirical investigations on the most critical issues. Convergent validity correlations

are useful but insufficient if the test is accused of bias and the challenge goes unexamined. Volumes of psychometric research will not help if the treatment to which low-scoring students are assigned is ineffective, or less effective than normal instruction. Kane's (1992) example of an interpretive argument required only seven statements of assumptions to frame an entire evaluation. Presented this starkly, it should be clear whether critical assumptions are supportable. If a test lacks validation evidence for a particular purpose, then its use is highly questionable, particularly if critical individual decisions will be based on test results.

Lastly, standards writers will need to think carefully about the assignment of responsibility for conducting validity evaluations. Although it was an innovation to create a separate set of standards for test users, there is a danger that test makers will defer to users to evaluate intended applications. This separation of responsibility would allow test makers to study only the "scientific" meaning of the test interpretation—the left side of Messick's (1989) framework—leaving it to the user to evaluate the test for its intended purpose.

Often, users of tests are not qualified, or lack the necessary resources, to conduct validity investigations. A classic example was the use of subparts of an existing early grades achievement test to retain low-scoring children in kindergarten in Georgia. The state legislature had mandated that kindergartners be retained on the basis of a test. No evaluation was done to establish the validity of the test for this purpose. Each group that might have been responsible assigned responsibility elsewhere. "Test developers assign responsibility to test directors, test directors believe that their jobs require them to acquiesce to legally empowered decision makers, and politicians at the top believe that they bought a valid test from a reputable firm" (Shepard, 1990, p. 41). Although test developers cannot anticipate bizarre misuses of their tests, there are clear intended uses for any test disseminated beyond a research study. For example, developers of commercial tests might reasonably be expected to have gathered validity evidence for the uses they recommend in their advertisements.

Because conceptual frameworks cannot be exhaustive, the validity framework appropriate to a test use will have a different focus than one for the same measure used to operationalize a construct in a research setting. Developers of practical tests, which might include commercial publishers, the College Board, a state department of education, or a corporate personnel office, should be able to specify the evaluation argument for the test use they intend and gather the necessary evidence, paying close attention to competing interpretations and potential side effects. In some cases, they may warn that local validity studies are also needed. At the same time, users who import a test from one context to the next

should be prepared to reexamine the validity argument and pursue new investigations as warranted by a shift in the purpose of testing.

NOTE

[1] The 1985 standards do identify some validity requirements as "conditional," meaning that whether they are considered primary or secondary varies as a function of the test use and potential consequences. However, standards labeled as secondary, including several deemed essential by the unified conception of construct validity, carry no such proviso.

REFERENCES

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*, 201–238.

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: Author.

American Psychological Association, American Educational Research Association, and 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, and National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

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

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 19–32). Hillsdale, NJ: Erlbaum.

Breland, H. M. (1977). *A study of college English placement and the Test of Standard Written English* (RDR-76-77, No. 4). Princeton, NJ: Educational Testing Service.

Breland, H. M. (1991, April). *The Advanced Placement Test item format study*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Bridgeman, B., & Lewis, C. (1991, April). *The predictive validity of advanced placement essay and multiple-choice scores*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Bridgman, P. W. (1927). *The logic of modern physics*. New York: Macmillan.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297–312.

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist, 15*, 546–553.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201–219). Washington, DC: American Council on Education and National Council on Measurement in Education.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? New directions for testing and measurement: Measuring achievement over a decade. In *Proceedings of the 1979 ETS Invitational Conference* (pp. 99–108). San Francisco: Jossey-Bass.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement theory and public policy* (Proceedings of a symposium in honor of Lloyd G. Humphreys, pp. 147–171). Urbana: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Crouse, J., & Trusheim, D. (1988). *The case against the SAT*. Chicago: University of Chicago Press.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.

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

Elliott, R., & Strenta, A. C. (1990, January). Is the SAT redundant with high school record in college selection? *Research and Development Update*, pp. 1–8. New York: The College Board.

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

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice. (1978). Adoption by four agencies of uniform guidelines on employee selection procedures. *Federal Register, 43*(166), 38290–38315.

Gesell Institute of Human Development. (1982). *A gift of time . . . A developmental point of view*. New Haven, CT: Author.

Gnezda, M. T., & Bolig, R. (1988). *A national survey of public school testing of pre-kindergarten and kindergarten children*. Washington, DC: National Forum on the Future of Children and Families, National Research Council.

Gnezda, M. T., Garduque, L., & Schultz, T. (Eds.). (1991). *Improving instruction and assessment in early childhood education*. Washington, DC: National Academy Press.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*, 427–439.

Guion, R. M. (1974). Open a new window: Validities and values in psychological measurement. *American Psychologist, 29*, 287–296.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology, 11*, 385–398.

Gulliksen, H. (1950). Intrinsic validity. *American Psychologist, 5*, 511–517.

Hambleton, R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art* (pp. 80–123). Baltimore: Johns Hopkins University Press.

Hambleton, R. K. (1984). Validating the test scores. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 199–230). Baltimore: Johns Hopkins University Press.

Hartigan, J. A., & Wigdor, A. K. (Eds.). (1989). *Fairness in employment testing:*

Validity generalization, minority issues, and the General Aptitude Test Battery. Washington, DC: National Academy Press.
Heller, K. A., Holtzman, W. H., & Messick, S. (Eds.). (1982). *Placing children in special education: A strategy for equity*. Washington, DC: National Academy Press.
Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing* (pp. 109–127). San Francisco: Jossey-Bass.
House, E. R. (1980). *Evaluating with validity*. Beverly Hills, CA: Sage.
Howe, K. R. (1985). Two dogmas of educational research. *Educational Researcher, 14*(8), 10–18.
Human Performance. (1992). 5(1 & 2), 1–169.
Hunter, J. E., & Schmidt, F. L. (1982). Fitting people to jobs: The impact of personnel selection on national productivity. In M. D. Dunnette & E. A. Fleishman (Eds.), *Human performance and productivity: Vol. 1, Human capability assessment* (pp. 233–284). Hillsdale, NJ: Erlbaum.
Kane, M. T. (1986). The future of testing for licensure and certification examinations. In B. S. Plake & J. C. Witt (Eds.), *The future of testing* (pp. 145–181). Hillsdale, NJ: Erlbaum.
Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535.
Kaufman, A. S. (1971). Piaget and Gesell: A psychometric analysis of tests built from their tasks. *Child Development, 42*, 1341–1360.
Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*, 1183–1192.
Linn, R. L. (1982). Ability testing: Individual differences and differential prediction. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies: Part II. Documentation section* (pp. 335–388). Washington, DC: National Academy Press.
Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*(Monograph Supp. 9), 635–694.
Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*, 955–966.
Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012–1027.
Messick, S. (1981a). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin, 89*, 575–588.
Messick, S. (1981b). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10*(9), 9–20.
Messick, S. (1984). Assessment in context: Appraising student performance in relation to instructional quality. *Educational Researcher, 13*(3), 3–8.
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: American Council on Education and Macmillan.
Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*, 229–258.
National Academy of Education Panel. (1992). *Assessing student achievement in the states*. Stanford, CA: National Academy of Education.

Nuttall, D. L., & Goldstein, H. (1990). *The 1988 examination results for ILEA* (mimeograph). London: London School of Economics.
Popper, K. R. (1962). *Conjectures and refutations: The growth of scientific knowledge*. New York: Harper & Row.
Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Belknap Press of Harvard University.
Ruch, G. M. (1929). *The objective or new-type examination*. Chicago: Scott, Foresman.
Schmidt, F. L., & Hunter, J. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529–540.
Scott, W. D. (1917). A fourth method of checking results in vocational selection. *Journal of Applied Psychology, 1*, 61–66.
Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research, 46*, 407–441.
Shepard, L. A. (1989). A review of research on kindergarten retention. In L. A. Shepard & M. L. Smith (Eds.), *Flunking grades: Research and policies on retention* (pp. 64–78). London: The Falmer Press.
Shepard, L. A. (1990). Discussion (In response to Anne Anastasi, What is test misuse? Perspectives of a measurement expert). In *The uses of standardized tests in American education: Proceedings of the 1989 ETS invitational conference* (pp. 37–44). Princeton, NJ: Educational Testing Service.
Shepard, L. A., & Graue, M. E. (in press). The morass of school readiness screening: Research on test use and test validity. In B. Spodek (Ed.), *Handbook of research on the education of young children* (2nd ed.). New York: Macmillan.
Shepard, L. A., & Smith, M. L. (1988). Escalating academic demand in kindergarten: Counterproductive policies. *Elementary School Journal, 89*, 135–145.
Shepard, L. A., & Smith, M. L. (1989). Academic and emotional effects of kindergarten retention in one school district. In L. A. Shepard & M. L. Smith (Eds.), *Flunking grades: Research and policies on retention* (pp. 64–78). London: The Falmer Press.
Skinner, B. F. (1945). The operational analysis of psychological terms. *Psychological Review, 52*, 270–294.
Smith, M. L., & Shepard, L. A. (1988). Kindergarten readiness and retention: A qualitative study of teachers' beliefs and practices. *American Educational Research Journal, 25*, 307–333.
Society for Industrial and Organizational Psychology, Inc. (1987). *Principles for the validation and use of personnel selection procedures* (3rd ed.). College Park, MD: Author.
Stricker, L. J., Rock, D. A., & Burton, N. W. (1991). *Sex differences in SAT predictions of college grades* (College Board Report No. 91-2 and ETS RR No. 91-38). New York: College Entrance Examination Board.
Tenopyr, M. L. (1977). Content-construct confusion. *Personnel Psychology, 30*, 47–54.
Thorndike, R. L. (1949). *Personnel selection test and measurement techniques*. New York: Wiley.
Toulmin, S., Rieke, R., & Janik, A. (1979). *An introduction to reasoning*. New York: Macmillan.
Tyler, R. W. (1934). *Constructing achievement tests*. Columbus: Bureau of Educational Research, Ohio State University.
Tyler, R. W. (1989). Constructing achievement tests. In G. F. Madaus & D. L.

Stufflebeam (Eds.), *Educational evaluation: Classic works of Ralph W. Tyler* (pp. 17–86). Boston: Kluwer Academic Publishers.

U.S. Department of Labor. (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 for the U.S. Employment Service* (USES Test Research Report No. 44). Washington, DC: Author.

U.S. Department of Labor. (1983b). *Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the General Aptitude Test Battery* (USES Test Research Report No. 45). Washington, DC: Author.

von Mayrhauser, R. T. (1992). The mental testing community and validity: A prehistory. *American Psychologist, 47*, 244–253.

Wigdor, A. K., & Garner, W. R. (Eds.). (1982). *Ability testing: Uses, consequences, and controversies: Part I. Report of the committee*. Washington, DC: National Academy Press.

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 75–107). Hillsdale, NJ: Erlbaum.

Willingham, W. W. (1985). *Success in college*. New York: College Entrance Examination Board.

Willingham, W. W., & Breland, H. M. (1982). *Personal qualities and college admissions*. New York: College Entrance Examination Board.

Wilson, K. M. (1983). *A review of research on the prediction of academic performance after the freshman year* (College Board Report No. 83-2). New York: College Entrance Examination Board.

Wylie, R. C. (1974). *The self-concept: Theory and research on selected topics* (Vol. 1). Lincoln: University of Nebraska Press.

Wylie, R. C. (1979). *The self-concept: Theory and research on selected topics* (Vol. 2). Lincoln: University of Nebraska Press.

Yalow, E. S., & Popham, W. J. (1983). Content validity at the crossroads. *Educational Researcher, 12*(8), 10–14.

Manuscript Received August 10, 1992
Revision Received October 26, 1992
Accepted October 26, 1992