

# Dialogue on Validity

## A Suggested Change in Terminology and Emphasis Regarding Validity and Education

by Robert W. Lissitz and Karen Samuelsen

This article raises a number of questions about the current unified theory of test validity that has construct validity at its center. The authors suggest a different way of conceptualizing the problem of establishing validity by considering whether the focus of the investigation of a test is internal to the test itself or focuses on constructs and relationships that are external to the test. They also consider whether the perspective on the test examination is theoretical or practical. The resulting taxonomy, encompassing both investigative focus and perspective, serves to organize a reconceptualization of the field of validity studies. The authors argue that this approach, together with changes in the rest of the terminology regarding validity, leads to a more understandable and usable model.

**Keywords:** assessment design; construct validity; content validity; test development

The drumbeat for change in the way we approach validity and our expectations regarding validation studies has been louder of late. Much of the dissatisfaction with Messick's unitary concept of validity is based on the notion that his rather global view of the topic is impractical. Brennan (1998) said, "In my experience those who are actually responsible for validation almost always require detailed and concrete guidance for conducting validation activities, and the 'unitary' notion is simply not helpful for them" (p. 7). John Fremer (2000) echoed this sentiment when he said, "We have elevated the concept of construct validation to so high a level that it seems an 'out of reach' goal" (p. 1). Borsboom, Mellenbergh, and van Heerden (2004) add, "The concept that validity theorists are concerned with seems strangely divorced from the concept that working researchers have in mind when posing the question of validity" (p. 1061).

If that is the case, and we think it is, where shall we go from here? Kane (2004), recognizing that the difficulty of applying validity theory to testing programs is "exacerbated by the proliferation of many different kinds of validity evidence and by the lack of criteria for prioritizing different kinds of evidence" (p. 136), has put forth an

argument-based approach to validity. Borsboom et al. (2004) also note that current validity theory "fails to serve either the theoretically oriented psychologist or the practically inclined tester" (p. 1061) and advance a simplified conception of validity. Linda Crocker (2003), in her presidential address to the National Council on Measurement in Education (NCME), suggests returning to the notion of content validity. As she says, "When scores are used for educational accountability, the 'load-bearing wall' of that [validity] argument is surely content representativeness" (p. 7).

We concur with her renewed focus on content for educational accountability. One purpose of this article is to provide a list of approaches to determining content validity. We also believe that a deconstruction of construct validity is in order, and a second purpose of this article is to provide a more communicative vocabulary for types of validation, maintaining their essential separateness of intent, focus, and verification procedures. The reader will note that we spend most of our time discussing content validity, but it is not a unified theory of all validity as the current construct validity movement attempts to be. A suggested vocabulary for these nonunified areas is provided at the end of the article. The variations in types of validation should not be ignored. Each type is important in its own area of application.

To orient the reader, we will first outline the systematic structural view of the technical evaluation of tests in terms of the study of reliability and validity. Figure 1 represents the structure of the problem as we see it, which includes two sets of considerations. One involves test development and analysis of the test itself; we refer to this as an *internal* matter. The second involves evaluations relating the test to other measures involved in a theory, the relationship of the test to criteria that measure its utility for specific *external* purposes, and the impact of the testing. We will say more about this as the article unfolds, but our thesis is that the internal characteristics should be determined to be the content validity of the test and that these do not depend on external factors, which we clarify and rename toward the end of the article. In other words, we are suggesting that an inquiry into the validity of a test should first concern itself with the characteristics of the test that can be studied in relative isolation from other tests, from nomothetic theory, and from the intent or purpose of the testing.

We will then provide an overview of our approach to the procedures that involve examination of the test characteristics. Figure 2,

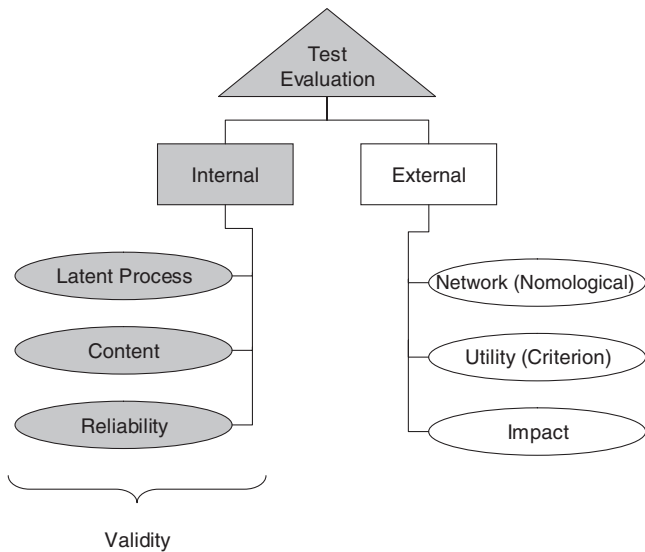


FIGURE 1. *The structure of the technical evaluation of educational testing.*

		Perspective	
		Theoretical	Practical
Investigative Focus	Internal	Latent Process	Content and Reliability
	External	Nomological Network	Utility and Impact

FIGURE 2. *Taxonomy of test evaluation procedures.*

which is complementary to Figure 1, indicates that there are essentially two primary perspectives (*theoretical* and *practical*) and that there are also two primary investigative focuses (internal and external) brought to bear on any examination or evaluation of a test. This two-by-two table reflects (guides) the structure defined in Figure 1. What we are trying to do here is to conceptualize the work on validity and reliability as fitting into a system. However, we are not trying to suggest a unitary definition. We reject the latter approach, recognizing that there are multiple purposes for examining validity and reliability and that it is helpful to think systematically about the nature of that work.

Later in this article, we present two tables to organize the questions that guide the typical approaches to evaluating important test characteristics. There is an extensive literature on this effort, and the tables help to illustrate the concerns and the richness of the published work on these matters. To justify the examination of validity, a brief historical detour may be helpful. This digression may also prove useful in reinforcing the idea that validity is a concept that evolved through the 20th century and is continuing to do so.

### A Brief History of Validity

Around 1915, although a formal definition of validity did not exist, the concept that would later be known as *criterion-related validity* was born. Over the following decade or so, those involved in psychological testing labored to quantify the relationships

between test scores and some criterion those scores were meant to predict. Reports such as the following were common. There was a “correlation of .50 between an accuracy score and grade of office work actually performed, while the *validity* [italics added] of a speed score obtained from the test was .42” (Dubois, 1970, p. 88). During that period, “tests which we would call ‘valid’ were regarded as ‘trustworthy’ or as ‘having diagnostic value’” (p. 46). It was not until the 1930s that validity was formally defined in the psychological literature. Two books, in particular, summarized the thinking on validity at that time: L. L. Thurstone’s (1931) *The Reliability and Validity of Tests* and J. P. Guilford’s (1936) *Psychometric Methods* (Dubois, 1970). Later, articles on the subject began to appear in the newly founded journal *Educational and Psychological Measurement* (e.g., Peterson, 1944). In these books and journal articles, establishing validity was linked to a correlation with some criterion. This is an example of what is called, in Figure 2, a practical perspective with an external focus.

Most validity studies of the 1930s and 1940s were predictive in nature, but another type of criterion-related validity developed as well: *concurrent validity*. Like predictive validity, concurrent validity calls for the correlation of test scores with some criterion, the difference having to do with when the criterion scores are observed. These two forms of validity mirrored the purpose of the testing commonly done at that time. A test was developed with the intention of predicting some external behavior. The validation was performed to establish that the test was achieving that purpose, in other words, that it had utility for that purpose.

Work on IQ testing began late in the 19th century and continued into the 21st. We will not try to capture the history of that work, except to note that it arose from several sources, including interest in eugenics (Galton, 1869) and interest in schooling and the identification of students who were particularly challenged (Binet & Simon, 1916/1983). The work by Binet and Simon is particularly interesting because of the clear influence of what today would be called *content validation*, along with criterion validity. The initial selection of tasks for the Binet and Simon tests came from an analysis of students’ activities in school. It was what might be called an abstracted result of a “job analysis” of the job of a student (i.e., the sort of thinking that was required of students in the French schools) that dictated the content of the initial forms of IQ testing. The measure of success of the tests was the ability to predict teachers’ judgments of their students’ abilities, using what became known as a *concurrent criterion validity* scheme.

In the late 1940s and early 1950s another type of validity began to be discussed: *content validity*. Rulon (1946), the acting dean of the Harvard Graduate School of Education, was looking at validity as it applied to educational testing. He defined an obviously valid test as

one in which the material presented to the student is the kind of material which constitutes the objectives of instruction, and in which the operation required of the student by the test situation is the operation which the school is trying to train the student to perform on such material. (p. 295)

Examples of obviously valid tests were tests for driving licenses and piano playing ability, as well as for penmanship, arithmetic, and sewing. Rulon believed that there was no need for external

criteria in cases such as these because the measure of an individual's acquired knowledge or skills served as its own criterion. In this situation, experts in the subject matter, who could determine whether the domain and the desired cognitive processes had been adequately sampled, could establish validity.

Cureton (1951) introduced the term *content validity* and contrasted this new type of validity with that based on criterion relationships:

We may, alternatively, ask those who know the job to list the concepts which constitute job knowledge, and to rate the relative importance of these concepts. Then when the preliminary test is finished, we may ask them to examine its items. If they agree fairly well that the test items will evoke acts of job knowledge, and that these acts will constitute a representative sample of all such acts, we may be inclined to accept these judgments [of validation]. (p. 664)

In modern industrial or organizational parlance, this form of validity is driven by answering the question of whether the test is consistent with a job and task analysis. In large-scale testing, the key questions are whether the test covers the relevant instructional or content domain and whether the coverage is at the right level of cognitive complexity. It might also be noted that in this special case, where the test is actually a sample of the criterion, a correlation between the two (performance on the test and performance on the job) would be an example of both criterion validity and equivalent forms reliability, simultaneously. This can happen in practice when the test is constructed by doing a job analysis and then the personnel director captures the job elements in both the criterion and the test. A simple example of such a practice is when a typing sample is used to test for typing skill during personnel selection and then at a later time the same employee's typing is evaluated to form the criterion. Later, we will briefly return to this observation that validity and reliability are not completely distinct concepts. The reader will note that this observation is incorporated in Figures 1 and 2.

Lennon (1956) formally defined, and perhaps extended, content validity when he characterized it as

the extent to which a subject's responses to the items of a test may be considered to be a representative sample of his responses to a real or hypothesized universe of situations which together constitute the area of concern to the person interpreting the test. (p. 295)

This focus on person response processes, rather than simply on test items, set the conception of content validity apart from others and provided a basis for considering tests of differential item function and factor analyses (as examples) under the umbrella of content validity.

Lindquist (1951), in his essay "Preliminary Considerations in Objective Test Construction," hypothesized a bridge between the seemingly disparate criterion-related and content validities. He discussed the relationship between test and criterion behaviors as it applies to content tests. He noted that although the test constructor attempts to devise a test that is as much like the criterion as possible, there are challenges.

The test series, however, does not contain all of the elements comprising the criterion series; rather, only the more discriminating, or

the more readily reproducible, or the more crucial, or the more readily measurable, or the more relevant of the elements of the criterion series are selected for the test, and these elements may, in important respects, be quite differently distributed than in the natural or criterion situations. (pp. 146–147)

Lindquist went on to articulate some of the irrelevant variations and factors introduced because of the lack of overlap between criterion and test. In doing so, he presaged future discussions of construct-irrelevant and construct-relevant variance components.

Writers of the late 1940s and early 1950s began to evidence what, in retrospect, might be described as dissatisfaction with the manner in which validity was defined, moving from empirical approaches to those that were more theoretical in nature (Sireci, 1998). For example, in 1946 Guilford discussed two different types of validity, practical and factorial. By his definition, practical validity was criterion related and factorial validity was based on the factor loadings (we would consider this to be one of the ways of formally approaching content validity). In 1949, Cronbach also spoke of two types of validity, but he defined them on the basis of the analyses involved. He classified the analyses as either practical or judgmental.

Although the term *construct* can be traced back to an article by MacCorquodale and Meehl in 1948, the concept of construct validity dates back to the 1955 article by Cronbach and Meehl titled "Construct Validity in Psychological Tests." In that seminal work, Cronbach and Meehl wrote:

Construct validity is ordinarily studied when the tester has no definite criterion measure of the quality with which he is concerned, and must use indirect measures. Here the trait or quality underlying the test is of central importance, rather than either the test behavior or the scores on the criteria. (p. 282)

It is important to recognize that we are drawing a distinction between the assessment of a construct and the development of an argument regarding construct validity. What needs to be kept in mind is that construct validity, as the full article by Cronbach and Meehl makes clear, not only is the measurement of the qualities (constructs) that are under consideration but also includes the nature of the network that relates these qualities to each other. In other words, without the *nomological network*, also known as a theory that specifies the relationships between a construct of focus and other constructs, we do not have construct validity. In Figure 2, this is what we call a theoretical perspective with an external focus. The Cronbach and Meehl article presents the measurement of the constructs (the links that represent the grounding or measurement of the constructs to observables) as well as the links that relate the constructs to each other (which represent the theory that relates these constructs). Our reading and interpretation of this article indicates that construct validity is the combination of the study of a construct (or trait) and its relationships to other constructs, not just the study of a construct in isolation. The study of nomological theory, in other words, is critical to construct validity; the term, which sounds as if it refers to a single construct, is in our opinion an unfortunate choice. Moreover, the article purposely confounds the two focuses—construct determination and theory development. This is particularly clear in item 5 of the "Recapitulation":

When a predicted relation fails to occur, the fault may lie in the proposed interpretation of the test or in the network. Altering the network so that it can cope with the new observation is, in effect, redefining the construct. (p. 300)

At a later point in our article, we suggest what we hope are more communicative terms for several aspects of validity. The excellent article by Embretson (1983) is another effort that we believe was intended to distinguish these two focuses, which she calls *construct representation* and *nomothetic span*. Her discussion on construct representation focuses on task decomposition, which examines the mechanisms that are responsible for the behavior observed by the choice of testing materials (e.g., test items). This is what we call a theoretical perspective with an internal focus (Figure 2).

It is clear that Cronbach and Meehl (1955) considered construct validity to be a fourth kind of validity, the other three being *content*, *predictive*, and *concurrent*. Although Cronbach and Meehl's article on construct validity, with its discussion of the nomological network, moved the validity discussion forward on a theoretical plane, it did little to answer the question of how to operationalize construct validation. Campbell and Fiske (1959) developed a strategy to examine the definitions of variables in their article on the multitrait-multimethod matrix. They were "primarily concerned with the adequacy of tests as measures of a construct" (p. 100), which they believed was a necessary precursor to studying a trait and its relationship to other traits (theory) as Cronbach and Meehl had suggested. To have confidence in the measure of a construct, Campbell and Fiske posited that it was essential to consider both convergent and divergent (or discriminant) forms of evidence in validation and to separate the construct-relevant variance from the method variance associated with the measurement process.

In 1957 Jane Loevinger wrote that, "since predictive, concurrent, and content validities are all essentially ad hoc, construct validity is the whole of validity from a scientific point of view" (p. 636). Although this idea gained acceptance when championed by Messick (1975, 1988, 1989) in his writing over several decades, it remained a hotly debated issue. Robert Ebel embraced one end of the spectrum of opinions on this topic and Messick the other in a series of arguments and rebuttals over the years. In a sense, Ebel (1983) thought of *ability* as more of a catchall term referring to a series of behaviors, and that belief led him to focus on "intrinsic rationale validity" (p. 8) or content validity. Messick, on the other hand, claimed that *ability* referred to an underlying trait, and that view led him to believe content coverage to be more of a test construction issue than a serious validity argument. As one dismissive of the concept of content validity, Messick (1989) argued that "content validity should not be viewed as residing in the test but, if anywhere, in the relation between the test and the domain of application" (p. 41). To some extent he is clearly right. A fourth-grade math test must reflect the domain that is determined by some process to constitute fourth-grade math. We argue later in this article that content validity may have been an ad hoc process at one time, but as we try to show in Table 1, it does not have to be. In other words, this content validation process is critical and can be formally assessed and should be. It is the content validation process that ensures that the mathematics test reflects the domain of interest.

This combination of theory and construct specification is captured in the *Standards for Educational and Psychological Testing* (hereinafter, the Standards; American Educational Research Association [AERA], American Psychological Association [APA], & NCME, 1999), which states: "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). This combination of theory and evidence regarding the constructs, as expressed in the notion of the nomological network and now construct validity, has been accepted by many as the overarching unified concept in the world of validity. One of our concerns is that the combination of theory and evidence regarding the constructs leads to a confounding that makes pursuit of validity (or theory) very difficult. It is the separation of these two intentions—construct definition and theory building—that we advocate and believe will improve the understanding and utility of both concepts. We and others (e.g., Borsboom et al., 2004, in psychology) argue that construct validity as it currently exists has little to offer to test construction in educational testing. The approach that is being developed should initially be focused on the content elements of the assessment (what we call *internal* and *theoretical*), their relationships, and the student behavior and cognitions that relate to those elements as they are being processed (i.e., cognitive theories of cognitive processes). When this stage is completed, the researcher should move on to external and theoretical. Perhaps Holland (1988) best expressed this view:

I do not mean to imply that the search for the causes of a phenomenon is a useless endeavor. Indeed, it is a driving force that motivates much of science. Rather, I mean that a logical analysis of the search for causes follows from an analysis of the measurement of causal effects, and it is not logically prior to this more basic activity. Defensible inferences about the causes of an effect are always made against a background of measured causal effects and relevant theories. (p. 451)

The same approach, laid out in our article, is first to study the measuring device (internal focus) and then to study the nomological network (external focus, theoretical perspective).

### Other Conceptions of Validity

In a very critical and interesting essay on the topic of validity, Borsboom et al. (2004) state: "We do not see the need for a unified validity concept (Ellis & Blustein, 1991; Messick, 1989; Moss, 1992; Shepard, 1993) because we think there is nothing to unify" (p. 1069). We, too, find the focus on construct validity to be a mistake, especially in the field of educational measurement, although Borsboom et al.'s focus is on psychology. In educational measurement, the test often has components that we need to understand, but the test stands in relative isolation. We deal in a test space where algebra (as an example) may be complex and deserve considerable attention as a set of subtests or subdomains. The issue of defining algebra is considered apart from any other latent trait such as English ability, and hence the development of a nomological network is not really an issue, even for future study. Test standards, in typical large-scale educational testing, have been written to try to specify the domain in question. A test then can be defined to measure that particular content defined by

**Table 1**  
**Internal Factors That Should Be Considered for the Systematic Evaluation of Content Validity**

Perspective	Sample of Questions Asked	Potential Sources of Evidence
Practical		
Content	Does the assessment encompass the full range of the content standards? Does the assessment properly reflect the cognitive complexity of those standards? Is the same emphasis reflected in the assessment as in the standards and in the classroom? Are the items appropriate for the purpose of the assessment? Are the items properly constructed? Are there criteria and mechanisms in place for scoring the items?	Analysis of the curriculum Creation of a table of specifications or test blueprint Documentation of match between items and blueprint Documentation that students have the opportunity to learn Documentation of the qualifications of the item writers and raters (including their training) Review of items and scoring rubrics for quality Examination of item characteristics (difficulty, discrimination, option selection) from pretest, pilot, or field test
Reliability	When matched on ability, do students from different racial groups perform similarly? Genders? Are the test items appropriate for students with limited English proficiency? Students with disabilities? Do the different test forms provide the same information? Does the test provide the same information on different occasions? How accurate are the test scores?	Bias and sensitivity review DIF analysis Review of the accommodations offered Analysis of the scores of accommodated students Test-retest reliability Parallel-forms reliability Internal consistency reliability Examination of the standard error of estimate for total score and subtest scores Rater consistency
Theoretical		
Latent process	Are the tasks eliciting the expected knowledge, skills, and abilities from the students? Do tasks that are supposed to work together do so? Do tasks that are supposed to provide unique information do so? Are the item difficulties at the expected levels, and are the distracters functioning as expected?	Review of the item performance data Examination of pattern of intercorrelation of items Examination of item, testlet, and total test score relationships Convergent/divergent evidence from correlations Factor analysis Results of verbal reports or think-alouds Cognitive analysis of student responses

Note. DIF = differential item functioning.

the standards, meaning that the test is a combination of tasks and these tasks are the operational definition that is captured (or is supposed to be) by the name and description of the domain to be measured by the test.

In this context, the test and its adequacy do not depend on the relationships to other tests defining other domains. A test and sometimes a modeling procedure, such as item response theory (IRT) or a cognitive analysis, jointly define in a behavioral and psychometric way the trait of interest. Referring to Figure 2, this particular test process may be practical and internal, although it may become theoretical, depending on the psychometric work being done; however, it remains internal. Deming (1986) summed up one valuable consequence of operationalization when he said, "An operational definition is one that reasonable men can agree on. An operational definition is one that people can do business with" (pp. 276–277). Our only addition to this is that in the era of latent variable analysis, there is a second operationalization that must be considered—the one that is defined by the psychometric model—and that model must also be

explicit and must be one that reasonable men and women can agree on. Sometimes that model includes a consideration of the scoring of the test items as well as the sampling of the items from some test bank. Clearly, to prepare a test, administer it, and then report the results involves a series of steps whose consideration, although important, would take us well beyond the focus of this article.

The following, oft-cited quote by Messick (1994) defines the key questions in his approach to validation:

A construct-centered approach would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (p. 17)

The use of the term *construct* has been confusing because it appears in the expression *construct validity* in the Standards (AERA, APA, & NCME, 1999) and in much of Messick's treatise on this topic but also is used in the terms *construct centered* or *construct representativeness* to refer to what had been originally and continues to be the result of a process called *content validation*. One could argue that we have a choice: either expand and deepen our understanding of content validity to include the idea that we are constructing a construct or change the definition of construct validity to minimize the role of the nomological network that deals with the development of the theory that relates multiple constructs to each other.

We feel that the essence of content validity has been the assessment of the process of the creation of instruments that measure something of interest (in Figure 1, what we call *content*) as well as the *latent* (cognitive) *process*. The adequacy of that creative process and its result is essentially what we are evaluating when we assess content validity. In our system, we suggest adding reliability to this process because it is an issue that is internal to the test (see Figure 1); that is, it does not involve any other tests or external measures. What we mean by other tests or measures does not include a parallel or equivalent form of the test that is one's focus. These multiple forms of a test are meant to be intersubstitutable, and the items are conceived as random selections from the same domain. That is very different from the situation where two tests are intentionally designed to test different domains and would never be substituted for one another.

The term *construct* usually is reserved, in other contexts, for the concept that is defined within the test or assessment device and the application of psychometric models in the case of latent variables and does not, by itself, include or imply the development of any theory that relates that construct to other constructs that arise external to the test that is being developed. As we discussed above in reference to the original article by Cronbach and Meehl (1955), construct validity is the integration of measuring constructs (which is essentially a matter of content validity) and the development of theory having to do with the formal relationships between different constructs, including some that are external to the test. Notice that we are using the term *theory* the way Cronbach and Meehl used the term when they were talking about the nomological network that relates multiple constructs to each other. This is in distinction to the development of a cognitive model that relates internal elements of the test, for example. This cognitive model may even be a theory. Construct validity is external and theoretical. In other words, the Messick (1994) quote is clearly a reference to establishing content validity, as we would define that term and, in the absence of theory relating multiple external constructs, does not refer to construct validation (p. 17). In Figure 2, it is the combination of theoretical and internal.

As we indicate above, early validation efforts (e.g., predictive validity) were basically efforts to find some, often conceptually unrelated, variable with credibility (termed *criterion*) that correlated with the test results and hence bolstered our faith in the importance of the test results. This approach lodges the validity in the relationship that is found between the test behavior and that behavior on the external variable, that is, the criterion. Whether one could understand why there was such a relationship

was irrelevant; hence it was practical and often atheoretical. Perhaps a short illustration of an obvious nature will help here. Suppose a group of researchers declare their belief in a concept called length. They then feel obligated to define what they mean by *length*. They agree that length is the result of the application under certain conditions (e.g., perhaps controlling temperature) of a certain ruler to measure the number of unit lengths that fit an object, and a unit length is specified by that ruler. So now the researchers have a "ruler" that measures length. It seems doubtful to us that anyone would feel more confident in this test of length if they measured a number of lengths and correlated these with something they called widths, or any other variable for that matter. In other words, the correlation or even some experimental approach would be useful to establish a relationship between length and width but would do nothing to verify (validate) or negate (invalidate) the concept of length, as defined by these researchers. Such issues of nomothetic span are important to determining theories that involve such concepts as length and width, but they do not verify the definitions of those terms.

This claim about length is the same as the claim for a test assessing the latent mathematics skills of a seventh grader, for example. Correlating these IRT-derived and scaled scores with a test that claims to be measuring English, and finding a very high correlation, does not support any claim that the two tests are measuring the same concept. The truth or falsity of that claim comes from an examination of the items and the way the tests were constructed, along with any psychometric theory that created the latent variables, if the focus is on such variables. For example, did the mathematics test specifications include coverage of material required by curriculum specialists in mathematics? As Borsboom et al. (2004) state, "One has to start with an idea of how differences in the attribute will lead to differences in test scores" (p. 1067). In our example, one can ask how people differ who do or do not know seventh-grade mathematics. The development of a test with a set of items (or the associated latent variable derived through application of a psychometric theory) that measures this difference is more likely to be a content-valid test for seventh-grade mathematics. Notice that even here we are not appealing to any external indicator to support a claim of content validity. We argue that the difference between a latent variable and a manifest variable may be great, but the differences do not enable—nor do they forbid—our use of the concept of content validity for constructs. Also, a high correlation between two different traits may be an important observation, but again, that observation by itself does not negate the content validation that took place.

We also find Messick's (1989) assertion that validity cannot reside in the test (p. 41) to be essentially incorrect and confusing. We argue that it does, in fact, reside in the definition of the test, the development phase, and any psychometric theory that gave rise to that test and its associated constructs, whether latent or manifest. The development phase (what Messick called *construct centered* and we call *content valid*) becomes critical and is one of the ways we know what those test characteristics are. In the previous example of a seventh-grade mathematics test, that test remains a seventh-grade mathematics test regardless of its domain of application, as we indicated above. Its claim to be a seventh-grade mathematics test (see Table 1 for examples of processes that support such a claim) is met or refuted by the process used

to create the test and an analysis of the test items and their relationship to socially constructed definitions of what we mean by saying that a test is a measure of such material. It is a practical perspective with an internal focus. It is not dependent on correlations with student height, running speed, English grammar, or any other measure that resides in some application to some domain that is external. As Borsboom et al. (2004) note, "The idea that validity consists in the correlation between a test and a criterion has obstructed a great deal of understanding and continues to do so" (p. 1065). Borsboom (2006) also makes the point we are trying to make here, that the psychometric model, to the extent that there is one—and there usually should be—is also critical to the construct definition that results. This becomes an added element of the content validity as we are defining the term. In other words, the content validity assessment should include some attention to the psychometric model that is applied to that test development activity. For example, whether the test development involves a multivariate IRT model (e.g., Reckase, 1997) or the one-, two-, or three-parameter IRT model is relevant to the content validation.

Examples from current testing in education are also valuable in understanding this argument. Probably the most common use of educational testing is in the classroom (P-16 and graduate school). How does a teacher know that his or her classroom test is valid? It would be very difficult to find an example of a validation study that provided convincing support for that classroom test, unless the study was essentially content validation. For example, the teacher might compare the test to the table of specifications that was (or should have been) prepared as an initial phase of the test construction; if those match, content validity will be incrementally supported by that evidence. The teacher might also compare the test to the curriculum that was dictated for the course (*Debra P. v. Turlington*, 1981; Phillips, 1993). This sort of validation process is also called *curricular validity* (Mehrens & Lehmann, 1991; Ornstein, 1993; Plake, Impara, & Buckendahl, 2004; Smisko, Twing, & Denny, 2000). The teacher might also look at the item statistics that result from analysis of the item behavior. Some of these are essentially reliability measures: for example, stability within a test (i.e., interitem correlations) or item-versus-test-total-score consistency. The teacher might also compare his or her test content to that required by national associations that provide guidance to teachers of mathematics. These and other internal processes are summarized in Table 1.

On the other hand, analysis that involves the correlation of the test results to the results of an IQ test might be interesting if one is trying to develop some theory. But if the design, construction, and psychometric modeling of the original test had been examined and found worthy (i.e., content validated) as to its content characteristics and the knowledge, skills and abilities, or attitudes (KSAs) it was evaluating, it is unlikely that anyone would suggest that we throw out those results in favor of ones resulting from an IQ test, even if there were a strong correlation. Even more important, if you had another test for the course that was not a good match to the content standards and had been poorly constructed and did not use a defensible psychometric model, but had a higher correlation with IQ or some other criterion, it is very unlikely, in our opinion, that the user would prefer this second test. This would also be true whether it was part of a nomological network (i.e., a theory) or not.

Another very common example for testing that rests primarily on content validity for its justification is the state assessment testing for Adequate Yearly Progress (No Child Left Behind Act, 2001). In some cases these are state-developed tests and in some cases they are off-the-shelf tests such as TerraNova (CTB McGraw-Hill, 2005) or the Iowa Test of Basic Skills (Riverside Publishing, 2005). Another example is the determination of high school graduation eligibility, perhaps by the New York State Regents Examination or the Maryland High School Assessments, and the determination of student status in the United States (e.g., by the National Assessment of Educational Progress). In each case, the justification for the legitimacy of the testing does not come from a correlation with another external measure (e.g., a criterion, as in traditional criterion validity) or from the existence of a theory that relates the latent trait to some other variable of interest that is external to the test. The heart of the justification of these tests comes from an argument that they constitute an assessment of material for which it makes sense to hold a student responsible and is supported by a series of analyses including perhaps quite complicated modeling (see Kennedy & Wilson, in press, and Mislevy, Steinberg, Almond, Haertel, & Penuel, 2003, for examples). As is stated in the Arkansas Student Assessment and Educational Accountability Act of 2004,

The purpose of this subchapter is to provide the statutory framework necessary to ensure that all students in the public schools of this state have an opportunity to demonstrate grade-level academic proficiency through the application of knowledge and skills in the core academic subjects consistent with state curriculum frameworks, performance standards and assessments.

In the state of Maryland, another example of this emphasis on content specification (content validity) is the Maryland Learning Outcomes (MLOs) that govern and guide the appropriateness of the state tests. There is actually no nomothetic span relating to domains outside the test itself, especially as in divergent validity considerations comparing the test to others to show that their results are different.

The argument that justifies the use of a test to meet these and other state requirements is laborious and complex and comes from efforts such as a systematic comparison of the test specifications with the required learning outcomes (e.g., the MLOs; Maryland State Department of Education, 2005) dictated by the state for all children or the arduous effort to see if there is support for the belief that the test is judged to be assessing the essentials of the textbook(s) that the school is using. In the best of these validation efforts there is also a psychometric theory underlying the development of the measures that result from this effort, whether those measures are latent (i.e., thetas) or manifest (i.e., transformed  $z$  scores). In our opinion, it is obvious that these are essentially content validation arguments and go far beyond any appeal to a simple understanding of their lowest level of operationalization for their justification, such as listing the items in the test. It is comprehensive, modern operationalization at a very sophisticated level of analysis that must be marshaled to lead to a judgment of whether this test has been validated.

**Table 2**  
**External Factors That Should Be Considered for the Systematic Evaluation of Educational Assessments**

Perspective	Sample of Questions Asked	Potential Sources of Evidence
Practical		
Utility	Are the outcomes from the test related to those from other recognized assessments of similar content? Are the outcomes from the test predictive of outcomes in a desired domain? Can interventions be found that are related to high achievement on the test?	Correlations Regression analysis Hierarchical linear modeling Cost-benefit analysis
Impact	Are the results from the test being used properly? Are the results from the test meeting the needs of the stakeholders? Are changes in practice or policy compromising the original intent of the test or parts of it?	Examination of the decisions being made with the test data Examination of the false positive and false negative error rates Conducting evaluation studies of utilization
Theoretical		
Nomological network	Is the latent trait measured by the test different from the trait(s) measured by other tests? Is the latent trait measured related to other traits of interest by a statistical or psychometric theory? Are there expected differences in the latent trait for different manifest groups? Does the latent trait change over time?	Multitrait-multimethod matrix Analysis of variance Confirmatory factor analysis Structural equation modeling Growth curve modeling

### A Suggested Change in Emphasis and Vocabulary

Following the line of reasoning presented above and the summary portrayed in Figure 1, we now turn to evaluations of a test that involve variables external to the instrument itself, for example, a test's relationship with other tests or other variables, constructs, or latent or manifest traits measured by some device external to the test in question and perhaps involving statistical or psychometric theory. There are three such external activities that are common and very important. Some approaches to these external questions are included in Table 2, and the following discussion helps to clarify the suggested change in terminology.

#### *Utility Determination*

This issue concerns exploring the usefulness of a test for some specific purpose. A person might ask if the test can be used to estimate some external variable (also known as a criterion) that is more difficult to obtain or can only be obtained after some decision must be made, when it is too late to be useful. For example, admissions counselors in a college might ask whether the XYZ test has utility for selecting applicants, in the sense that persons who do well on the XYZ test tend to do well on some measure of success in college. Clearly, the absence of such a demonstration has nothing to do with the validity of the test developer's claim to have carefully measured a particular KSA. The absence of a correlation might suggest that some other measurement process or a different trait might be more useful, but we suggest reserving the word *validity* for the establishment of the definition of the trait (i.e., KSA) in the test development phase and the analysis of its stability. This approach can be reduced to essentially an atheoretical, strictly empirical approach to testing and validation. By itself, discovering an association contributes no understanding of either the criterion or the test construct that is used

to predict it. In the current terminology of our field, this sort of study is often labeled *criterion-related evidence of validity* (AERA, APA, & NCME, 1985, 1999). We suggest using a new phrase to apply to a successful study of such a matter: "This study has determined that there exists a high level of test utility for that purpose."

#### *Theory Support*

The class of theory support studies, again, has no direct impact on the internal validity of an assessment device but is a very important application of interest and may indirectly cause us to consider using a different instrument. For example, if researchers have a belief that by manipulating some variable they should see some change in another variable measured by the test or assessment device in question or if researchers study some dynamic system and believe they should see some particular specified configuration of causes and effects (whether called a nomological network or a relationship and often appears as a set of arrows in a structural equation model), we would say they are engaging in work to support a theory. The failure to verify a theory involving a particular construct as indicated by a particular test and choice of psychometric theory should have, obviously, no impact on the belief in the validity of that test as a measure of some construct, domain, variable, or trait. On the other hand, this negative evidence might lead a researcher to think about other possible constructs, even with similar names, but with different definitions. For example if the variable *height* did not yield a very satisfying relationship in one's theory of basketball playing, it might occur to a theoretician that *differential height* (as previously defined by a careful process that measured the difference between static height and jump height) might be a better variable to use in one's theory of effective basketball playing. In other words, we are not arguing

that researchers should not consider the variables they are using in their theories, but we are suggesting that lack of empirical evidence in support of a theory (a nomological network, in Cronbach and Meehl's lexicon) is not an indicator of inadequate content validity evidence for the test device. Our notion of validity clarifies the confusion about whether the theory or the test is at fault when construct validity is not found. In our system, the fault is in the theory because theory development would always be the result of a two-stage process in which determination of the content validity sense of the measure precedes the use of the test in a theory. This approach is consistent with that of Holland (1988), whom we quoted earlier as taking the position that the first thing to do is to determine the measure and only after that is done can one move to building a theory involving that measure.

There is another argument related to the traditional construct validity literature that we feel is quite relevant here. This has to do with how theories are in fact supported or refuted. In our opinion, Messick's conceptualization and extensive discussion offers little in the way of useful advice to researchers interested in theory development or in the process of supporting or rejecting an established theory (see Borsboom et al., 2004, as well, for comments on this issue). We believe that work on experimental design (e.g., Cook & Campbell, 1979), structural equation modeling (e.g., Jöreskog, 1973), and a host of other statistical approaches that dictate methods of collecting and analyzing data as support for a theory are much more fruitful fields to study if the student wishes to learn something of use related to theory development. Theory validation is very important, but its furtherance is not facilitated by reading the history of validation. This is not true of content and criterion validity and reliability theory, where there exists a set of approaches that lead researchers through a process that strongly supports or does not support their claims.

### *Impact Evaluation*

The class of impact evaluation studies, which involve the consequences of using a test or assessment device—again having no direct relationship to internal validity—is also an important kind of research. Messick (1989) referred to the validity at issue in such studies as *consequential validity*. We think the impact of a test is sometimes an important consideration and sometimes worth studying, but if a test is later shown to have some impact that is unintended or unwanted, that observation should not be considered relevant to the question of whether the test is valid. In a way, the consequential validity concern is related to the argument that the ends justify the means. In other words, consequential validity suggests that there is a belief that if a test leads to unacceptable or unwanted ends, the test (the means) was not acceptable or valid. The argument can also go in the other direction, where positive consequences can be seen as justifying the use of a particular assessment device. We reject that argument. The reader interested in this issue might refer to Linn (1997), Reckase (1998), or Moss (1998) for some converging and diverging opinions.

Impact evaluation is not a characteristic that helps determine content validity, but it can be an important consideration. Sometimes testing is done so that a certain impact on the environment is achieved. In personnel selection and placement, for example, the adverse impact of testing is of considerable interest to the company involved or to the courts. In such a case,

a concern with the impact on hiring and job success is very critical. Impact evaluation should be assessed in such a situation. Another example is the case where testing is introduced to change what teachers do in the classroom. If that is the impact that was intended, then verifying that it actually occurs would be an important step.

### **Methods to Establish Test Characteristics**

We believe that psychometricians must adopt a new way to think about validity and a higher level of expectation related to the determination of content validity. We believe that if the primacy of content validity were recognized, then psychometricians would begin to develop much more advanced procedures for systematically determining the adequacy of the test development and test definition process. It is important to recognize that, although we believe progress can be made in this effort, the field is not without powerful approaches already, including the ones listed in Table 1. These are approaches that we believe are helpful for each of the internal areas pictured in Figure 1.

Considerable work has already begun on the development of content validity as an assessment enterprise. Examples focused on the content include the work by Nitko (2004); see, for example, his Table 3.2 (pp. 42–43) as an illustration of an effort to organize this work. Another content-focused example is the work by Downing and Haladyna (1997); see their Tables 2 and 3 (pp. 64 and 73, respectively). A third example of a systematic approach to the theory underlying the performance within a test, which we characterize as a type of content validity, is the work of Mislevy et al. (2003), although the authors would certainly disagree with our labeling of their approach. Our basis for this characterization of their approach is that their conceptual assessment framework specifies variables that characterize students and schemas for getting evidence about those students. Our discussion above should clarify that we would put the work of Mislevy et al. in the camp of content validity and not in construct validity, because the focus of much of their work is essentially inside the test of interest. We see much of their work as being in the spirit of Rulon (1946), Lindquist (1951), and Lennon (1956) and their emphasis on the operations and person response processes. The area of the cognitive analysis of a test is one of the most productive and promising areas in psychometric application today. Related work by Kennedy and Wilson (in press) is another excellent example of an effort that looks deeply at the cognitive aspects of the thinking that a student engages in when coming to grips with a test. This modeling includes valuable insights into the content of the test and how it is or should be scored, as well as the learning and thinking processes of students. Another example is the argument-based approaches of Kane (2004) and Mislevy et al. that focus on evidence and the need to specify an argument during the test construction phase rather than consider validity to be a post hoc or ad hoc endeavor. We can hope, as does Borsboom (2006), that psychologists (and, we would add, educators) will come to see the value of sophisticated psychometric work applied to the study of educational processes, including—and perhaps, especially—assessment.

In our opinion, the driving forces for theory development and verification come from three directions: (a) the specific domain of interest; (b) the literature on research design and statistics,

particularly structural equation modeling, which will help guide the specification of the connections and the verification of these connections that constitute the theory; and (c) psychometric theory, which can do much the same thing. As indicated above, some of the approaches to this area are included in Table 2.

### Summary and Conclusion

Clearly, there is an iterative connection between theory building and content or construct specification that cannot be ignored. We believe it is important to treat both phases as critical to the development of good theory and to the application of theory to education and psychology. We suggest that the heart of the validity question—and the primary effort that a person committed to test validation must make—is the study of the test construction process, including the specification of the psychometric theory associated with the assessment device. We hope that we have shown that minimizing the importance of this effort was a mistake.

There is, of course, another characteristic of tests, as we have indicated above, that is not dependent on the relationship to other external measures, constructs, traits, and so forth, and that is reliability. Whether we look at internal consistency or test-retest reliability, or at the consistent classification of a student as proficient, these statistics are essentially a characteristic of the test and do not depend on any empirical or theory-driven study of relationships with other measures external to that test, no matter how important or interesting these other measures may be. In this sense, reliability is also a fundamental characteristic of a test that does not depend on other external measures for its meaning. We certainly acknowledge that the level of the reliability coefficient is in part a function of the sample of students who took the test. If the students were more different in their abilities, then, other things being equal, that would lead to a higher reliability coefficient. One might argue that reliability derives from the test data, and of course that is true if one takes a simple computational viewpoint of the matter. You cannot get a coefficient if you do not have data and do not subject the data to analysis. We suggest that people should not take such a viewpoint: Their focus should be on the test itself and the process that led to its development. Following this reasoning, we believe that the profession should adopt a very different way of conceptualizing the validity issue and a very different terminology for talking about this problem. Together, we suggest that these essentially internal characteristics (reliability and content validity) be called the *internal validity* of the test, and all other characteristics be considered essentially external matters. Figure 1 summarizes our conceptualization of the matter.

In this new formulation of validity, the test definition and development process (what is currently known as *content validity*) and test stability (what is currently known as *reliability*, or sometimes *generalizability* [Brennan, 1983]) become the critical descriptors of the test. They also become the primary justification for its existence and acceptance for use. They exist independent of, or regardless of, the application of the test or the use of the test in some theoretical formulation. The quality of each of these forms of validity is what determines the degree of internal validity of the test as a measure of a trait, construct, or academic achievement level or any KSA. We argue that this is the fundamental issue for users of tests (including nonpsychometricians,

such as teachers in a classroom, reading specialists, and even esoteric theorists or policy analysts) and for judges in a trial involving test results. Content validity, or internal validity, should be acknowledged as the critical initial characteristic to consider when evaluating the quality of a test.

We expect a ruler or a weighing scale or any other measure to be well defined and to give reproducible results when applied appropriately, in the absence of change in the entity being measured. It is not unreasonable, in our opinion, to expect the same from psychological or educational measurement devices. This is certainly what users of tests want and expect from the tests that their clients take. Certainly, rulers differ in quality: Some give more stable results than others do, and these should be said to be more valid indicators of the traits in question. Of course, in testing, there are sources of variance that one does not encounter in the application of a ruler. For example, test items may change from form to form and across time; however, they should always be under the control of the administrator and always consistent with the design plan for the test. This is true whenever multiple forms are constructed or administration depends on spiraling, to name just two examples. This will clearly affect the statistics associated with the testing and must be kept in mind. We do not believe, though, that such a consideration changes the central truth of what we are saying. Another concern is that a ruler is a manifest and therefore well-defined measurement device. Our point is that the same should be true of psychological and educational measures, and we believe that an emphasis on the content validity process, including the psychometric theory in the case of latent variables, will provide us with a definition of a construct that can approach the universality of a ruler.

We are not saying that the relationship of length to some other variable, construct, or trait is not important. It certainly is, but it is not a fundamental characteristic of the assessment device or its related data. The internal characteristics of an instrument do not depend on the relationship to some external variable to define the instrument's validity, although such external variables can be important, as we have indicated above. The other characteristics—what are currently called *criterion* and *construct validity* determinations—are very important, but the user of these techniques should, we argue, recognize that they answer fundamentally different questions. It is for this reason that they should not be presented as a unified theory of validity. These questions should not draw the researcher's attention away from a focus on internal validity.

In summary, we have tried to suggest a reorientation of the validity problem and a more useful terminology that we hope will make the validation process less of a “nebulous enterprise” (Langenfeld & Crocker, 1994, p. 154). We have suggested that thinking clearly about the content of the assessment is the first step and the most basic step. We have suggested that criterion validity (what we call *utility*) is a useful concern, for very limited purposes where a specific criterion can be specified and a theory does not exist. Finally, we have taken the position that theory is important but has gained little or no benefit from the 50 or more years of ruminating about construct validity. Again, we are not suggesting that theory is unimportant, although for most applications of psychometrics in education there is no external theory being developed and hence no opportunity or necessity to think hard about that issue.

We have also suggested one approach to revising the terminology for evaluating tests, and we hope that this article will serve to further the discussion. Whatever the new validity paradigm that emerges, it must be one that professionals can understand and become comfortable with and one that will prove to be valuable to them. It seems clear to us that the time has come for such a development.

## NOTE

We thank Gregory Camilli, coeditor of *Educational Researcher's* Features section, and several reviewers for their very helpful comments.

## REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Arkansas Student Assessment and Educational Accountability Act of 2004, Arkansas Code § 6–15–404 (2004).
- Binet, A., & Simon, T. (1983). *The development of intelligence in children*. Salem, NH: Ayer. (Original work published 1916)
- Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71, 425–440.
- Borsboom, D., Mellenbergh, G., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071.
- Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing Program.
- Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice*, 17(1), 5–9.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validity by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.
- Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice*, 22(3), 5–11.
- Cronbach, L. J. (1949). *Essentials of psychological testing*. New York: Harper.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- CTB McGraw-Hill. (2005). *TerraNova, The Second Edition (CAT/6)*. Monterey, CA: Author.
- Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.
- Debra P. v. Turlington, 644 F.2nd 397 (5th Cir. 1981).
- Deming, W. E. (1986). *Out of the crisis*. Cambridge, MA: MIT Press.
- Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, 10(1), 61–82.
- Dubois, P. H. (1970). *A history of psychological testing*. Boston: Allyn & Bacon.
- Ebel, R. L. (1983). The practical validation of tests of ability. *Educational Measurement: Issues and Practice*, 2, 7–10.
- Ellis, M. V., & Blustein, D. L. (1991). The unificationist view: A context for validity. *Journal of Counseling and Development*, 69, 561–563.
- Embretson, S. E. (Whitely). (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93(1), 179–197.
- Fremer, J. (2000). Promoting high standards and the “problem” with construct validation. *NCME Newsletter*, 8(3), 1.
- Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences*. London: Macmillan.
- Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6, 427–439.
- Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equation models. *Sociological Methodology*, 18, 449–493.
- Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences*. New York: Seminar Press.
- Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement*, 2(3), 135–170.
- Kennedy, C., & Wilson, M. (in press). Using progress variables to map intellectual development. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting*. Maple Grove, MN: JAM Press.
- Langenfeld, T. E., & Crocker, L. M. (1994). The evolution of validity theory: Public school testing, the courts, and incompatible interpretations. *Educational Assessment*, 2(2), 149–165.
- Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement*, 16, 294–304.
- Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 119–158). Washington, DC: American Council on Education.
- Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, 16(2), 14–16.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3(Monograph Supplement 9), 635–694.
- MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, 55, 95–107.
- Maryland State Department of Education. (2005). *Learner outcomes and indicators*. Retrieved June 5, 2007, from <http://www.mdk12.org/mspp/mspp/whats-tested/learneroutcomes>
- Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). Fort Worth, TX: Holt, Rinehart & Winston.
- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955–966.
- 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: Lawrence Erlbaum.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23.
- Mislevy, R. J., Steinberg, L. S., Almond, R. G., Haertel, G., & Penuel, W. (2003). Leverage points for improving educational assessment. In B. Means & G. Haertel (Eds.), *Evaluating the effects of technology in education* (pp. 149–180). New York: Teachers College Press.
- Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, 62(3), 229–258.
- Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, 17(2), 6–12.
- Nitko, A. J. (2004). *Educational assessment of students* (4th ed.). Upper Saddle River, NJ: Pearson.
- No Child Left Behind Act of 2001, Pub. L. No. 107–110 (2001).
- Ornstein, A. C. (1993). Norm-referenced and criterion-referenced tests: An overview. *NASSP Bulletin*, 77(555), 28–39.

- Peterson, S. (1944). The word-dexterity test, a better measure of college aptitude. *Educational and Psychological Measurement*, 4(4), 307–313.
- Phillips, S. E. (1993). *Legal implications of high-stakes assessment: What states should know*. Oak Brook, IL: North Central Regional Educational Laboratory.
- Plake, B. S., Impara, J. C., & Buckendahl, C. W. (2004). Technical quality criteria for evaluating district assessment portfolios used in the Nebraska STARS. *Educational Measurement: Issues and Practice*, 23(2), 12–16.
- Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25–36.
- Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice*, 17(2), 13–16.
- Riverside Publishing. (2005). *Iowa Test of Basic Skills*. Rolling Meadows, IL: Author.
- Rulon P. J. (1946). On the validity of educational tests. *Harvard Educational Review*, 16, 290–296.
- Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405–450.
- Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, 45, 83–117.
- Smisko, A., Twing, J. S., & Denny, P. (2000). The Texas model for content and curricular validity. *Applied Measurement in Education*, 13(4), 333–342.
- Thurstone, L. L. (1931). *The reliability and validity of tests*. Ann Arbor, MI: Edwards.

#### AUTHORS

**ROBERT W. LISSITZ** is a professor of education and director of the Maryland Assessment Research Center for Education Success, Department of Measurement, Statistics, and Evaluation, University of Maryland, College Park, MD 20742; [rlissitz@umd.edu](mailto:rlissitz@umd.edu). His research focuses on psychometrics, applied statistics, value-added modeling, test linking, and standard setting.

**KAREN SAMUELSEN** is an assistant professor in the Research, Evaluation, Measurement, and Statistics Program, Department of Educational Psychology and Instructional Technology, University of Georgia, 325S Aderhold Hall, Athens, GA 30606; [ksam@uga.edu](mailto:ksam@uga.edu). Her research focuses on mixture models, especially as they pertain to the measurement of differential item function.

Manuscript received August 8, 2006

Revisions received November 6, 2006, and February 26, 2007

Accepted July 24, 2007