

Comments on Lissitz and Samuelsen

Construct Validity: A Universal Validity System or Just Another Test Evaluation Procedure?

by Susan E. Embretson

Lissitz and Samuelsen (2007) have proposed a framework that seemingly deems construct validity evidence irrelevant to supporting educational test meaning. The author of this article agrees with Lissitz and Samuelsen that internal evidence establishes test meaning, but she argues that construct validity need not be removed from the validity sphere. In fact, she argues that doing so could have an adverse impact on the quality of educational tests. She proposes a universal system for construct validity to illustrate how diverse evidence is relevant to claims about measuring examinees' knowledge, skills, abilities, and competencies even when test specifications provide a major source of evidence.

Keywords: construct validity; content validity; educational testing

Validity is a controversial concept in educational and psychological testing. Research on educational and psychological tests during the last half of the 20th century was guided by a threefold distinction of types of validity: criterion-related validity, content validity, and construct validity. Construct validity has always been the most problematic type of validity because it involves theory and the relationship of data to theory. Yet the most controversial type of validity became the sole type of validity in the revised *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). That is, according to the current standards, "validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). Content validity and criterion-related validity are no longer separate types of validity but two of five different kinds of evidence. The current test standards reflect substantial impact from Messick's (1989) thesis of a single type of validity (construct validity) with several different aspects.

Robert W. Lissitz and Karen Samuelsen, in their article "A Suggested Change in Terminology and Emphasis Regarding Validity and Education" (this issue of *Educational Researcher*, pp. 437–448), describe many current pressures on the validity concept in educational testing; in fact, they cite some researchers as believing construct validity to be an unreachable goal. To remedy the situation, Lissitz and Samuelsen propose changes in terminology and emphasis in the validity concept and then relate their distinctions to

specific sources of relevant evidence. These distinctions and the linkages to evidence are extremely useful. But their system leads to a most startling conclusion. Construct validity is irrelevant to defining what is measured by a test!

In Lissitz and Samuelsen's view, content validity becomes primary in determining what an educational test measures. But content validity, as discussed below, is not up to the burden of defining what is measured by a test. In fact, I believe that relying on content validity evidence, as available in practice, to determine the meaning of educational tests could have a detrimental impact on test quality. Furthermore, giving content validity primacy for educational tests could lead to very different types and standards of evidence for educational and psychological tests.

This article is organized into three parts. First, the concept of construct validity and its categorization in Lissitz and Samuelsen's framework will be considered. Second, the limitations of content validity in establishing test meaning will be described and illustrated. Third, construct validity will be conceptualized to incorporate some of the distinctions made by Lissitz and Samuelsen and to define a universal system of validity evidence.

Construct Validity in Lissitz and Samuelsen's Framework

Lissitz and Samuelsen provide an interesting, scholarly, and highly readable history of the validity concept, especially its controversial aspects. Their proposed changes include a taxonomy to classify test evaluation procedures. The taxonomy is stratified by two dimensions: (a) *investigative focus* (internal vs. external sources of information), and (b) *perspective* (theoretical vs. practical). For investigative focus, the internal sources of information are derived from an analysis of the test and its items, and the external sources include the relationship of test scores to other measures and to criteria. For perspective, the theoretical orientation might be interpreted as concern with measuring traits, the practical orientation as concern with measuring achievement. Thus the practical perspective is most relevant to educational and classroom measurement.

Various sources of information are classified within this system. For the theoretical perspective, the internal information includes studies on latent processes involved in item solving; the external information includes nomological-theoretical studies of test scores and other measures. It is worth noting that Cronbach and Meehl's (1955) nomological network is classified as external and theoretical. For the practical perspective, internal information includes

primarily test content and reliability; external information includes test utility and impact.

The distinction between internal and external sources of validity is reminiscent of an earlier conceptualization of construct validity (Embretson, 1983) as consisting of two aspects: construct representation and nomothetic span. In Embretson (1983), construct representation concerned identifying the theoretical mechanisms that underlie test performance (i.e., the processes, strategies, and knowledge); nomothetic span concerned the network of relationships of test scores with other variables. Thus construct representation was concerned with the meaning of test scores, nomothetic span with the significance of test scores. On the surface, it seems that Lissitz and Samuelsen's designation of internal sources as providing primary information about what is measured by the test (pp. 437, 441–442, 446) coincides with the notion of construct representation. However, it is important to note, Lissitz and Samuelsen's distinction between internal and external concerns types of *test evaluation procedures*, not aspects of validity. In fact, Lissitz and Samuelsen "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" (p. 444). Thus the external sources of evidence do not provide information about test validity.

In Lissitz and Samuelsen's taxonomy, the concept of construct validity no longer would be regarded as validity because they regard it as relying on external methods. So rather than just decentralize construct validity, they have removed the concept from the validity sphere! Lissitz and Samuelsen "argue that construct validity as it currently exists has little to offer to test construction in educational testing" (p. 440). In their 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. (p. 446)

Obviously this is a radical change in conceptualization.

A critical element in this deconstruction of construct validity is its classification as external evidence. This classification does not reflect Cronbach and Meehl's (1955) conceptualization because they did include internal sources of evidence, namely, studies of internal structure, studies of change, and studies of processes. Within the nomological network, these sources would be classified as test-to-construct evidence. Thus construct validity, in its initial conceptualization, did include the internal sources of information viewed as central to test meaning by Lissitz and Samuelsen. Thus construct validity need not be decentralized for this reason.

However, Cronbach and Meehl (1955) did not regard the internal sources of information as having precedence in defining the construct; instead, these were simply other sources of evidence about the construct, which included the external relationships. Lissitz and Samuelsen's view of construct validity as concerning only external sources of information may characterize some current practices, especially for psychological tests. For many tests, especially those that are not based on contemporary

model-based procedures, it is far less expensive to reconceptualize test meaning on the basis of external evidence than to develop new test forms that better measure the originally intended construct. Concern about the strong role of external sources in establishing test meaning motivated Embretson's (1983) distinctions in construct validity to give internal sources the primacy role. If internal sources are primary, then item and test design principles (Embretson, 1998) become central in establishing test validity.

If construct validity were construed to include internal sources of information, it would become central to establishing test meaning for psychological tests in Lissitz and Samuelsen's framework. Lissitz and Samuelsen's Table 2 lists several questions that are relevant to item and test design to establish test meaning under the theoretical perspective, which presumably is relevant only to psychological tests. Relying on internal information to establish test meaning requires a scientific and theoretical foundation for item and test design principles. That is, scientific evidence and theory are needed on how item-type-specific features and testing procedures affect knowledge, skills, and abilities (KSAs). Principles for test design are emerging for some item types, including popular test items such as paragraph comprehension (Freedle & Kostin, 1993; Gorin & Embretson, 2006) and mathematical problem solving (Embretson, 2004; Enright, Morley, & Sheehan, 1999; Singley & Bennett, 2002), as well as item types on ability tests, such as spatial items (Bejar, 1993) and nonverbal reasoning items (Embretson, 1998). Although research on item and test design principles is an active area, a much larger foundation is needed to support test meaning from this kind of evidence.

But even if the concept of construct validity were extended to include internal evidence, it would still not be considered appropriate to establish meaning for educational tests in Lissitz and Samuelsen's framework. That is, educational tests are classified in the practical perspective in their framework, and thus test meaning depends primarily on content-related evidence. Therefore, I next consider issues surrounding the use of content-related evidence in establishing test meaning.

The Limitations of Content Validity in Defining Test Meaning

Lissitz and Samuelsen's framework separates the sources of test meaning for educational and psychological tests. Test meaning for educational tests, which are classified in the practical perspective, can be supported by reliability evidence and by content validity evidence; test meaning for psychological tests arises from evidence of latent processes (i.e., KSAs) and item interrelationships. It is argued here that content validity and reliability are not sufficient evidence to establish meaning for educational tests. One might also argue that latent process evidence is not sufficient for psychological tests, but that is not the primary concern here.

Examining the shortcomings of content validity and reliability evidence for educational tests requires first establishing what is included. The reliability concept in the Lissitz and Samuelsen framework is generally multifaceted and traditional, including both item interrelationships and the relationship of test scores over conditions or time. Reliability does also include differential item functioning (DIF) and adverse impact. Perhaps it could be argued that adverse impact and DIF could be considered as external information because

external groups are involved. However, more important is the nature of their conceptualization of content validity and its adequacy, along with reliability evidence, to support test meaning.

The concept of content validity has evolved over time. In the current test standards (AERA et al., 1999), content-related evidence is linked to constructs, in that it concerns “the relationship between a test’s content and the construct it is intended to measure” (p. 11). Because Lissitz and Samuelsen strive to decentralize construct validity’s role, their view is more consistent with the prior test standards (AERA, APA, & NCME, 1985), in which content validity was a type of evidence that “demonstrates the degree to which a sample of items, tasks or questions on a test are representative of some defined universe or domain of content” (p. 10). However, Lissitz and Samuelsen’s view of content validity is broader than the prior test standards (AERA et al., 1985) because they add two important elements: cognitive complexity level and test development procedures. They summarize the key issues in large-scale (educational) testing as “whether the test covers the relevant instructional or content domain and whether the coverage is at the right level of cognitive complexity” (p. 439), which is consistent with many current educational test blueprints. Test development procedures are added to content validity in Lissitz and Samuelsen’s Table 2 because information about item writer credentials and item quality control are included.

On many levels of educational testing, content specifications, in the form of test blueprints that often include item complexity level, have become increasingly central to guiding test content. These blueprints specify percentages of test items that should fall in various categories. For mathematical achievement, for example, the test blueprint for the National Assessment of Educational Progress (NAEP; Martin, Olson, & Wilson, 2006) includes five content strands as well as three levels of complexity. Currently, the majority of states employ similar strands in their year-end and federally mandated testing of mathematical achievement. More detailed specifications within strands is also common, and information about test development procedures is available for many such tests.

But are blueprints and other forms of test specifications, in conjunction with reliability evidence, sufficient to establish meaning for an educational test? For several reasons, I believe they are not.

First, it should be recognized that the structure of a content domain is, in itself, a theory. These structures change over time. For example, the NAEP framework, particularly for cognitive complexity, has evolved (Martin et al., 2006). Changes in domain structure also could evolve in response to recommendations of panels of experts. For mathematics achievement, contemporary views and scientific evidence on mathematics and mathematical learning, such as those being considered by the currently constituted National Mathematics Panel, could lead to changes in the basic strands and their relative emphasis. Views on complexity level also may change on the basis of empirical evidence, such as item difficulty modeling, task decomposition, and other methods.

Second, scant evidence is available that items can be reliably classified into the blueprint categories. Certain factors in an achievement domain may make such categorizations difficult. For example, in mathematics a single real-world problem may involve algebra and number sense as well as measurement content. Thus the item potentially could be classified into three of

the five strands. Similarly, classifying items for mathematical complexity also can be difficult, given the somewhat abstract definitions of the various levels in many systems.

Third, practical limitations on large-scale testing may lead to unrepresentative samples of the content domain. More objective item formats, such as multiple choice and limited constructed response, have long been favored because they can be reliably and inexpensively scored. But these formats may not elicit the deeper levels of reasoning that experts believe should be assessed for the subject matter.

Fourth, using content specifications, along with item writer credentials and item quality control, may not be sufficient to assure high-quality tests. Leighton and Gierl (2007) view content specifications as one of three models for making inferences about examinees’ thinking processes. A major weakness in using the cognitive model of test specifications for inferences is that no evidence is provided that examinees are in fact using the presumed skills and knowledge to solve items.

Consider, for example, the results from a recent validity study on the NAEP mathematics assessment (Daro, Stancavage, Ortega, DeStefano, & Linn, 2007) for Grades 4 and 8. A group of mathematicians examined items from NAEP and from some representative state accountability tests. Only a small percentage of items (i.e., 3%–7%) were deemed to be flawed, which is encouraging. However, a much larger percentage of items were deemed marginal (23%–30%). The mathematicians also were clear that many of the items they classified as marginal exhibited construct-irrelevant difficulties that could affect performance for some test takers. These difficulties included problems with pattern specifications, unduly complicated presentation, unclear or misleading language, and excessively time-consuming processes. It is notable that these marginal items previously had survived both content-related and empirical methods of evaluation. Although empirical item tryout probably leads to the detection of many flawed items, apparently it is not sufficient to detect marginal items. To accommodate the inherent heterogeneity of most achievement domains, the standards for evaluating indices of item discrimination are set relatively low. Given the high percentages of items deemed marginal in NAEP and representative state tests, it is easy to imagine that some items could have acceptable discrimination indices for the wrong reason. That is, they are related to other items in the test with similar marginal features.

The problem of irrelevant KSAs may be the most important limitation of relying on content validity evidence that is not linked to constructs. To provide a concrete context for irrelevant sources of performance, it is instructive to consider some actual test items. To illustrate marginal mathematics items, the two items in Figure 1 are variants of actual disclosed items from national and state tests. These items were written to correspond to test blueprints in mathematics, and they had presumably met some standards for item quality. Yet a logical-theoretical analysis by a mathematician, following procedures similar to those involved in the NAEP validity study, reveals some irrelevant features for mathematics.¹

Both problems can be regarded as involving excessive or irrelevant wording. That is, they contain words or phrases that may distract, confuse, or mislead the student. In the first problem, an eighth-grade item, the intended mathematical task is for the student

- 1) Mr. Wilson and 3 friends dined at a popular restaurant. The bill was \$77 and they left a \$15 tip. Approximately what percentage of the total bill did they leave as a tip?
- A) 10 %
 - B) 13 %
 - C) 15 %
 - D) 20 %
 - E) 25 %
- 2) A total of 80 players were in a football league. There were 10 players on each team. Which number sentence is in the same fact family as $80 \div 10 = 8$?
- A) $8 \times ? = 80$
 - B) $8 \times 10 = ?$
 - C) $? \times 80 = 10$
 - D) $8 \times 80 = ?$

FIGURE 1. *Mathematics items from standardized tests with misleading wording.*

to find out what percentage 15 is of 77. But the wording does not necessarily lead students to this conclusion. Some students might not know what a tip is, having never have been to a restaurant in which there is tipping. And \$77 for four people might seem so unreasonable to some students that they will look for a catch. Also, the term *bill* is used in one place, and the term *total bill* in the question. The total bill can reasonably be regarded as total paid (i.e., \$92), so the mathematical task becomes what percentage is 15 of 92. Thus C, not D, becomes the best answer. In the second problem, a fourth-grade item, the description of football is irrelevant; for those who know that there are 11 players on a football team, it is also misleading. Perhaps worse, *fact family* is not a mathematical concept. Students can understand the inverse relation between multiplication and division without understanding or using the term *fact family*.

These items are not isolated cases, unfortunately. One can visit state websites to find many other examples of excessive or confusing wording. Furthermore, these examples concern only irrelevant verbal content. Similar analyses of items for irrelevant spatial content and irrelevant logical skills can be readily found.

Circumventing these irrelevant sources of item performance requires more sources of evidence than those listed by Lissitz and Samuelsen under the internal and practical perspective for educational tests. First, evidence listed under Lissitz and Samuelsen's theoretical perspective is also relevant to detecting irrelevant sources of item performance. These methods include cognitive analysis (e.g., item difficulty modeling), verbal reports of examinees, and factor analysis. Second, although Lissitz and Samuelsen did not regard external sources of evidence as relevant to test meaning, in fact they may provide needed safeguards. For example, Lissitz and Samuelsen regarded the correlation of an algebra test with a test of English as irrelevant to the meaning of an educational test. But if this correlation is too high, it may suggest a failure in the system of internal evidence that supports test meaning.

Finally, it is important to consider a possible consequence of classifying educational and psychological tests as involving different systems of evidence, as in Lissitz and Samuelsen. This type of separation was a real possibility in the early 1990s, when testing was changing rapidly and no revision of the joint test standards had been initiated. In 1992 the Committee on Psychological Tests and Assessments (of which I was a member) was concerned about the development of separate standards and initiated the process that eventually led to the development of the

current standards.² Perspectives did indeed differ between the three organizations; however, once initiated, it became important that the standards applied to any test, from any perspective.

Construct Validity as a Universal System and Unifying Concept

In this section, a validity framework is presented to show that construct validity need not be decentralized. The framework is consistent in many ways with the current test standards (AERA et al., 1999) and Cronbach and Meehl's (1955) conceptualization, as well as with Lissitz and Samuelsen's distinctions and elaborations. With further development, the framework may provide a useful conceptualization of validity evidence. In the framework, validity is conceptualized as a universal and interactive system. That is, the system is universal because all sources of evidence are included and may be appropriate for both educational and psychological tests. The system is interactive because the adequacy of evidence in one category is influenced or informed by adequacy in the other categories.

Eleven categories of evidence about test meaning and significance, defined in Table 1, are included in the system. It is important that these categories be conceived broadly so that they can be applied to both educational and psychological tests. Consistent with most validity frameworks and the current test standards (AERA et al., 1999), it is postulated that tests differ with regard to which categories in the system are most crucial to test meaning, depending on the test's intended use. Even so, most categories of evidence are potentially relevant to a test.

Some of the categories of evidence presented here correspond directly with Lissitz and Samuelsen's framework. For example, two external sources of evidence, Utility and Impact, are taken directly from their framework. Another external source, labeled Other Measures in Table 1, is similar to the evidence that Lissitz and Samuelsen call *nomological information*, but it is relabeled here to allow the information to play varying roles (including the feedback role). Other categories are elaborations of Lissitz and Samuelsen's distinctions. For example, the category Psychometric Properties includes all the evidence in Lissitz and Samuelsen's *reliability* category, as well as their *latent process* category as related to a specific test. Scoring Models is a separate category in this universal system so that the impact of decisions about dimensionality, guessing, elimination of poorly fitting items, and so forth is highlighted for its impact on scores and their relationships. The Test Specifications category is construed broadly, like Lissitz and Samuelsen's *content* category, to include test blueprints, item writer guides, item writer credentials, test administration procedures, and so forth. It is not quite the same as their content category, however. Two major sources of evidence are separated from Test Specifications, namely, *domain structure* and *item design principles*. These categories consist of scientific evidence, knowledge, and theory that exist apart from a particular test. Domain Structure is postulated to be the knowledge and theory about the domain as indicated by scholarly publications of subject matter and curriculum experts. Item Design Principles, on the other hand, consists of scientific evidence and knowledge about how features of items affect the KSAs applied by examinees.

Figure 2 presents a schema of the validity system. Solid paths in the system indicate direct impact; dashed paths indicate feedback. Internal and external sources of evidence are indicated by the lines at the bottom of the figure.

Table 1
Evidence Categories in the Validity System

Category	Examples
Logical/Theoretical Analysis	Theory of the subject matter content, specification of areas and their interrelationships.
Latent Process Studies	Studies on content interrelationships, impact of item design features on psychometric properties and response time, impact of various testing conditions, etc.
Practical Constraints	Available test administration methods, scoring mechanisms (raters, machine scoring, computer algorithms), testing time, locations, etc.
Item Design Principles	Formats, item context, complexity, and specific content as determining relevant and irrelevant basis for item responses.
Domain Structure	Specification of content areas and levels, as well as relative importance and interrelationships.
Test Specifications	Blueprints specifying domain structure representation, constraints on item features, specification of testing conditions.
Psychometric Properties	Item interrelationships, DIF, reliability, relationship of item psychometric properties to content and stimulus features, reliability.
Scoring Models	Psychometric models and procedures to combine responses within and between items, weighting of items, item selection standards, relationship of scores to proficiency categories, etc.
Utility	Relationship of scores to external variables, criteria, and categories
Other Measures	Relationship of scores to other tests of knowledge, skills, and abilities
Impact	Consequences of test use, adverse impact, proficiency levels, etc.

Note. DIF = differential item functioning.

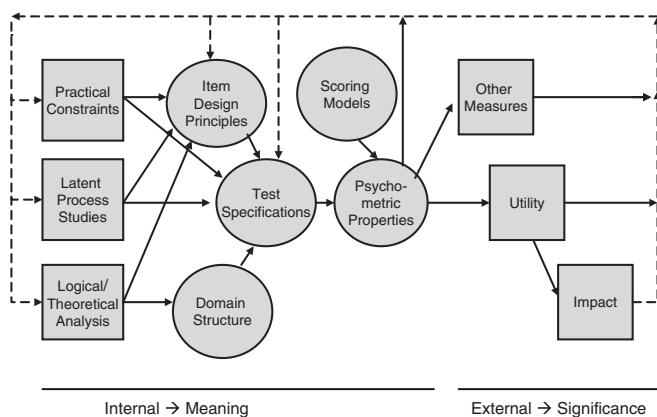


FIGURE 2. *A universal system for validity.*

Perhaps the most essential category in determining test meaning in Figure 2 is Test Specifications because, in conjunction with Scoring Models, specifications determine the psychometric properties of the test and, in turn, the external relationships of test scores. However, the Test Specifications category is preceded by Domain Structure and by Item Design Principles, both of which involve scientific evidence, knowledge, and theory. Domain Structure, in turn, is preceded by Logical/Theoretical Analysis. Item Design Principles, on the other hand, is viewed as arising from scientific evidence about KSAs (i.e., Latent Process Studies, as well as Logical/Theoretical Analysis). The category Practical Constraints is included in the system because such constraints, in part, determine the item types for which it is important to develop design principles.

Finally, feedback loops are included from all external sources of evidence and Psychometric Properties to the categories that precede them. The external relationships of test scores are viewed as informing not only the Test Specifications that guide a particular test but

also Item Design Principles and views of Domain Structure. For example, consider a test of fourth-grade pre-algebra skills that was speeded to place emphasis on automatic numerical processes. External evidence of strong adverse impact for certain groups of children may lead to questioning either the relationship of item speededness to automaticity, an item design issue, or to questioning the domain structure that placed heavy emphasis on the automaticity of numerical skills.

Several aspects of the system also are consistent with Lissitz and Samuelsen's distinctions and elaborations. First, test meaning is determined by internal sources of information. Second, test significance is determined by external sources of information. Third, content aspects of the test are central to test meaning. Test specifications, which include test content and test development procedures, have a central role in determining test meaning. Furthermore, test specifications also directly determine the psychometric properties of tests, including reliability information.

However, the system does differ in some important ways from the Lissitz and Samuelsen framework. First, a broader system of evidence is viewed as relevant to supporting test specifications. That is, general evidence about item design principles, as well as general knowledge about domain structure, supports the test specifications for a particular test. These two categories of evidence, in turn, are preceded by further scholarly activities, empirical studies of latent processes, and logical/theoretical analyses of the domain, as well as by consideration of practical constraints. Second, the broader system of evidence includes both theoretical and practical elements. Thus, for an educational test, item design principles include scientific evidence and theory about latent processes to provide evidence that examinees' responses are relevant to the intended domain. Similarly, the domain structure that supports the test specifications is regarded as a theory in this system. Third, interactions among components arise because internal evidence leads to expectations for external evidence. And, conversely, external evidence provides important information

about the adequacy of the evidence from internal sources. That is, although external evidence is not given a major role in determining test meaning per se, it provides potentially important information about the adequacy of the internal evidence: When hypotheses about external evidence are not confirmed or unintended consequences of test use are observed, potential inadequacies in the evidence supporting test meaning are indicated.

Thus this purposed universal system of validity includes multiple sources of validity evidence and allows interaction between internal and external sources of information.

Conclusion

On the basis of their distinctions and elaborations, Lissitz and Samuelsen have removed the concept of construct validity from establishing the meaning of an educational test. They believe that only internal evidence should support test meaning. For educational tests, the relevant internal evidence would consist of content validity and reliability information. For psychological tests, internal evidence would consist of latent process studies. Because they regard construct validity as consisting only of external evidence, such validity is not regarded as relevant to establishing test meaning.

In this article I have tried to show that construct validity need not be decentralized. Cronbach and Meehl's (1955) concept of construct validity did include both internal and external sources of evidence, and if viewed this way, the concept would fit in Lissitz and Samuelsen's framework for tests from the theoretical perspective. However, perhaps more significant is the designation of different types of evidence for theoretical as opposed to practical perspectives. For educational tests, relying solely on the types of internal evidence in the Lissitz and Samuelsen framework could adversely affect educational test quality, given the current conceptualization and implementation of content-related evidence. Other sources of internal evidence, including test design principles and scholarly views of domain structure, are extremely important in supporting the relevancy of examinee responses to the intended content domain. I gave examples from major educational tests to show how items that survive quality control procedures like those described by Lissitz and Samuelsen may evoke construct-irrelevant response processes for the examinees.

Because both test design and domain structure involve elements of theory and scientific evidence, their inclusion is more appropriately conceptualized as the purview of construct validity. Furthermore, I have suggested that external evidence is not irrelevant to test meaning, especially if such results lead to questioning whether examinees' responses are based on the KSAs targeted by the test. To unify these diverse aspects of relevant evidence, I have presented a universal system of validity with both theoretical and practical elements. This system shows how construct validity can provide internal evidence for educational test meaning as well as an important role for external evidence.

Consider again the often-cited view of Messick (1989) on 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)

What kind of validity best characterizes the approach that Messick envisioned?

Lissitz and Samuelsen argue that Messick's definition is more appropriate for what has traditionally been part of content validity than for construct validity, as intended by Messick. They reason as follows:

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. (p. 441)

I agree with Lissitz and Samuelsen that the role of external evidence in establishing test meaning should be minimized and that internal evidence should be strongly emphasized. However, given the need for multiple sources of evidence to establish internal test meaning, including theoretical components even for educational tests, I believe that the choice is clear. The validity system that they seek is, after all, best labeled as *construct validity*.

NOTES

¹Thanks to Bert Fristedt, Department of Mathematics, University of Minnesota, for the analysis of these examples.

²The joint standards might not have been created but for the initiative of Frank Farley, who was president of both the American Educational Research Association and the American Psychological Association in the early 1990s.

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.
- Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Fredriksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323–359). Hillsdale, NJ: Lawrence Erlbaum.
- Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing Program.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Daro, P., Stancavage, F., Ortega, M., DeStefano, L., & Linn, R. (2007). *Validity study of the NAEP Mathematics Assessment: Grades 4 and 8*. Washington, DC: NAEP Validity Studies Panel, U.S. Department of Education.
- Embretson, S. E. (Whitely). (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93(1), 179–197.
- Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3, 380–396.

- Embretson, S. E. (2004, October). *A cognitive model of mathematical reasoning*. Paper presented at the annual meeting of the Society for Multivariate Experimental Psychology, Naples, FL.
- Enright, M. K., Morley, M., & Sheehan, K. M. (1999). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education, 15*(1), 49–74.
- Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing, 10*, 133–170.
- Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 30*, 394–411.
- Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26*(2), 3–16.
- Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*, 437–448.
- Martin, W., Olson, J., & Wilson, L. (2006). *Mathematics framework for the 2007 National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board, U.S. Department of Education.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp.13–103). New York: American Council on Education/Macmillan.
- Singley, M., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 361–384). Mahwah, NJ: Lawrence Erlbaum.

AUTHOR

SUSAN E. EMBRETSON is a professor of psychology and director of quantitative psychology at the Georgia Institute of Technology, School of Psychology, 654 Cherry Street, Atlanta, GA 30332; susan.embretson@psych.gatech.edu. She was the recipient of the 2001 Distinguished Scientific Lifetime Contribution Award, American Psychological Association, Division of Measurement, Evaluation and Statistics. Her research interests include measurement, psychometric methods, cognition, and individual differences.

Manuscript received September 18, 2007

Accepted September 20, 2007