



Comments on Lissitz and Samuelsen

## Reconsidering Issues in Validity Theory

by Joanna S. Gorin

Lissitz and Samuelsen (2007) propose a new framework for validity theory and terminology, emphasizing a shift in theory and practice toward issues of test content rather than constructs. The author of this article argues that several of Lissitz and Samuelsen's critiques of validity theory focus on previously considered, but subsequently discarded, validity conceptualizations. In addition, she suggests that Lissitz and Samuelsen's conceptualization returns to methods shown historically to be problematic for score use and interpretation. In doing so, she highlights developments in validity theory and practice centering on cognitively based examinations of test scores that have contributed to increased understanding of score meaning and stronger validity arguments.

**Keywords:** construct validity; measurement theory; test development; validity theory

As a faculty member in a graduate program in educational psychology, I teach the introductory graduate-level measurement course annually to first-year students from various departments in the College of Education. Consistently, the most challenging class of the semester for me to teach (and for students to understand) is the one devoted to validity. Regardless of their prior coursework in measurement or their specific disciplines within education, students all enter the course believing that validity (whatever their conceptualization of the word may be) is a critical issue for testing. In fact, most, if not all, use the terms *validity* and *valid test* in conversation to refer to any "good" use of tests. 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), correctly point out that most first-year graduate students have a general, and simple, definition of validity. They see it as the extent to which a test measures what it is supposed to measure. Beyond this definition, however, students rarely understand how validity is judged; needless to say, they have little idea how it is established. Despite my best attempts to describe the holy trinity, the unified framework, or argument-based approaches to validity, few students emerge from the class with confidence that they could evaluate validity when developing, using, or even selecting tests.

Therefore, I appreciated the authors' provocative article, which challenges us to clarify and question some of our implicit

assumptions about validity theory and practices. Regardless of how those of us in the measurement field may perceive validity theory, it is of little use if it remains inaccessible to the larger educational and psychological community.

I begin with the authors' invitation to generate my own definition of validity. Validity is the extent to which test scores provide answers to targeted questions. It originates in a question about individuals, groups, programs, and systems. In the context of the question, data are provided by tests. Through a process of empirical analysis, the test scores are interpreted in terms of the original question. Empirical data collection is a critical component of this process, primarily for establishing the cognitive and psychometric properties of test scores. Nevertheless, data collection is different from validation. As noted by Mitroff and Sagasti (1973) in Messick's (1989) discussion of validity, "Data are *not* information; information is that which results from the interpretation of data" (p. 123). Validation is an information-gathering or evidence-building process that culminates in a test of hypotheses about score meaning (Borsboom & Mellenbergh, 2007). Information and evidence are evaluated relative only to the question(s) to be answered by score interpretation. The greater the extent to which the available evidence supports use of the test scores to answer to my assessment questions, the greater my confidence in using the test.

The central thesis of Lissitz and Samuelsen's suggested changes to validity theory and terminology might be summarized as follows: Validity is an issue of the content of a test, which is an internal test property and should be examined as such. They propose a decidedly controversial shift in validity theory—an approach that is often beneficial for advancing new ideas—but I remain unconvinced that their suggested changes move the domain in a good direction. In many ways, in fact, Lissitz and Samuelsen's conceptualization may be seen as a move backward, toward methods that have been shown historically to be problematic for measurement theory and practice.

Let us consider three related aspects of the proposed validity framework and terminology:

- Validity is an issue of the internal content of tests, not the external definition of constructs.
- Validity is established through appropriate operational definitions and item development procedures. Empirical tests of the success of these procedures address separate issues.
- Validity is not influenced by the way in which scores are to be used. It should not be examined differently for tests that have different uses (e.g., diagnosis, placement, or research).

Educational Researcher, Vol. 36, No. 8, pp. 456–462  
DOI: 10.3102/0013189X07311607  
© 2007 AERA. <http://er.aera.net>

The majority of my commentary examines potential consequences, both positive and negative, of these proposed changes in validity theory. In doing so, I present a differing (and more optimistic) view of current validity theory. Finally, I conclude by considering Lissitz and Samuelsen's claim that current validity theory "has little to offer to test construction in educational testing" (p. 440), and I address the question whether their framework better meets the needs of test developers and the larger measurement community. Although I limit my discussion primarily to measures of cognitive abilities, I would argue that for noncognitive measurement, which involves traits that are often less well defined or associated with observable behaviors, many of the arguments may be even stronger.

### **Validity Is About Internal Content, Not Externally Derived Constructs**

Both in the title of their article and throughout the text, Lissitz and Samuelsen present a strong argument for validation that stresses test content rather than constructs. The crux of their argument is that within a traditional validity framework, constructs can be defined (and thus validated) only through analyses of external measurements. According to their argument, the question of whether a test measures what it purports to measure is a question of the meaning of test scores, which should be established by examining the internal properties of the test. Content is internal and therefore relevant to validity; constructs are external and consequently addressed in other test development procedures and analyses.

If we instead stipulate that constructs can be defined with terms internal to the measurement system, the differentiation between content and constructs becomes less clear. Lissitz and Samuelsen would have us believe that contemporary construct-centric theories of validity perpetuate Cronbach and Meehl's (1955) original view of constructs and construct validity. Most important, Lissitz and Samuelsen contend that constructs can be known only within a larger nomological network in terms of external variables and their observed relationships. Yet measurement and validity theorists' notions of constructs have evolved beyond this perspective. Specifically, as the influence of cognitive psychology on psychometrics and assessment has grown, researchers and test developers have realized the utility of componential models of cognition to describe test score meaning (Baxter & Glaser, 1998; Carroll & Maxwell, 1979; Embretson & Gorin, 2001; Kyllonen & Christal, 1990; Pellegrino, 1988; Snow & Lohman, 1989). For example, the construct measured by a cognitive test can be described in terms of the cognitive and metacognitive strategies, and multiple alternative paths to correct answers that include sequencing and execution of cognitive subprocesses. These formulations are internal theories that define test constructs and therefore, even within Lissitz and Samuelsen's framework, are relevant for validity examinations. Although they briefly acknowledge the existence of such theories, they fail to grasp the fact that it is the internal theories of the construct (i.e., the relationships among subprocesses and dimensions comprised by the construct), not those of the nomological network, that are the focus of construct validity in contemporary measurement. Embretson's (1983, 1995) work distinguishing construct representation and nomothetic span as separate aspects of construct validity is more reflective of currently held conceptualizations of

constructs and construct validity than those that Lissitz and Samuelsen address. Lissitz and Samuelsen's argument against construct validity would be strengthened if their comparisons were made to currently held conceptualizations. To compare the proposed content validity framework with the views of Cronbach and Meehl is to argue against a stance that is generally considered to have evolved.

Beyond the distinction between content as internal and constructs as external, there is little differentiation between Lissitz and Samuelsen's notion of content and other definitions of construct. Thus what Lissitz and Samuelsen present as a radical change in validity terminology is more appropriately characterized as an issue of semantics or perhaps terminological preference. For several reasons, I would argue that in choosing between *content* and *construct*, use of the latter to describe the whole of internal processes underlying test scores better serves the broader measurement community. One reason is that, from a practical perspective, it is important to look beyond the educational assessment context. Contrary to Lissitz and Samuelsen, I would argue that constructs exist across all assessment contexts, whereas content does not. Take, for example, measures of language impairment administered by speech pathologists and school psychologists. The typical tasks, including sentence repetition, fast mapping, nonword repetition, and rapid naming, are content free; adequate score meaning derives from the representativeness of the skills required of the items, independent of content (Bishop, North, & Donlan, 1996; Conti-Ramsden, 2003; Denkla & Rudel, 2007; Devescovi & Caselli, 2007; Dollaghan, 1987; Dollaghan & Campbell, 1998). A parsimonious theory of validity suitable for all possible contexts rather than limited to discipline-specific models would be preferable. A more flexible vocabulary for validity theories would allow for cross-disciplinary discussions. Although many scholars are critical of various aspects of Messick's (1989) unified view of validity, its flexibility in terminology and implementation is perhaps one of its hallmarks, a point to which I will return later.

More important, a return to the use of content validity as the whole of validity theory threatens to stifle many of the recent advances in test design resulting from construct-centric models of validity. Historically, the use of content validity tools such as operational definitions as indicators of score meaning has been tried and discarded. For the most part, the methods associated with test development based on operational definitions, like those proposed under the content classification of Lissitz and Samuelsen, have led to empirically unsubstantiated claims regarding score meaning. Contemporary construct validity theories advance methods, including those borrowed from psychology, that increase the level of detail and the empirical nature of score descriptions (Borsboom & Mellenbergh, 2007; Embretson & Gorin, 2001; Gorin, 2006; Leighton, 2004; Leighton & Gierl, 2007; Mislavy, 2006; Snow & Lohman, 1989; Yang & Embretson, 2007). One example is the use of cognitive models of test constructs. Unlike the behavioral descriptions characteristic of operational definitions, cognitive models specify individual processes that describe item solutions. Strong theories of test scores (where *theory* refers not to the nomological network but to the internal processes of the test) in the form of cognitive models provide a testable hypothesis regarding score meaning. From this vantage, test score analysis can be more scientific, generating measurable evidence regarding test scores

and their interpretation. Test developers and researchers have used cognitive models of test items to improve the quality of validity arguments (Embretson & Gorin, 2001; Gorin, 2006; Yang & Embretson, 2007), to streamline item development procedures (Bejar, 1993; Embretson, 1998, 1999; Enright & Sheehan, 2002; Irvine, 2002), and to augment typical score reports generated from tests (Briggs, Alonzo, Schwab, & Wilson, 2006; Huff, 2006; VanderVeen et al., 2006). Although they are often more difficult to develop than operational definitions, cognitive models of test constructs have shown great potential for describing tests. The process may be unfamiliar to many educators and test development experts; however, the return on the investment is favorable.

### Validity Is Established Through Operational Definitions and Test Development Procedures

More than a theoretical framework to conceptualize validity, educators and psychologists want practical tools for validation. In addition to the shift from construct emphasis to content focus, Lissitz and Samuelsen address the procedures for validity investigations and sources of validity evidence. In their view, “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) processes*” (p. 442). For the most part, Lissitz and Samuelsen give weight to results from experts’ domain and task analyses and to completeness of test specifications regarding content representativeness as sufficient evidence of validity. They place less emphasis on substantive examinations of student processing and cognition. The justification for their procedures, consistent with their overall emphasis on content validity, is to restrict validity analyses to internal test properties—in other words, to preclude analyses that involve any measurements external to the test. They reject the notion that analyses including entities external to the test can serve as evidence regarding internal score meaning. In support of their argument, Lissitz and Samuelsen suggest that in current validity theory, correlations among measures trump all other sources of evidence and provide sufficient justification for test use. Regarding a hypothetical test of students’ knowledge, skills, and abilities, the authors state that

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 . . . 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. (p. 443)

I am unable to see where Messick’s unified framework of validity suggests that correlations with other tests are the only or even the most important source of evidence. Corroboration of the content through expert review, qualitative analysis of the content in a field, and other methods (which are included within the unified validity framework) can provide useful evidence of the representativeness of a set of items for a domain. However, it is idealistic to suppose that expert review of the items’ content will result in scores that reflect only the intended skills. Correlations with external variables provide one source of additional evidence to consider in the larger validity argument.

Test developers who invest thousands upon millions of dollars to hire content experts and item-writing specialists can attest to the fact that even with great effort, unintentional dimensions are often reflected in student responses to items. In general, Lissitz and Samuelsen’s view on appropriate evidence for validity is far more restrictive than those that they wish to discard. Sources of validity evidence include response process analysis, experimental manipulation, correlational analysis, group comparisons, and content analysis (Borsboom & Mellenbergh, 2007). Whereas Messick (1989) suggests that “construct validity embraces almost all forms of validity evidence” (p. 16), Lissitz and Samuelsen’s suggested methods for establishing validity are limited to examinations of content representativeness, targeted mostly in the early stages of item development. Given the cost of test development and the high stakes often associated with score use, however, it seems prudent to evaluate the effectiveness of the test development methods in generating useful and meaningful test scores.

Although briefly acknowledged by Lissitz and Samuelsen, most current construct-centered validity frameworks also strongly emphasize the early stages of item development. Test development frameworks evolving from construct-oriented validity theories, including Embretson’s cognitive design system (CDS; Embretson, 1994, 1998) and Mislevy’s evidence-centered design framework (ECD; Mislevy, Steinberg, & Almond, 2003), highlight the need for rigorous efforts early in the test development process (as do Lissitz and Samuelsen). However, where Lissitz and Samuelsen stop at documentation of test development procedures, ECD and CDS complete the process. Lissitz and Samuelsen argue that once appropriate item development methods have been implemented and a psychometric model applied, score validity is established. But to argue that an item has its meaning innately and that we therefore need no further justification except to state the logic by which the item was written and statistically parameterized is to ignore the need for any empirical support for score interpretation (beyond the authority of the development specialists). The aforementioned test design frameworks proceed where Lissitz and Samuelsen fall short. In CDS as well as ECD, principled design of items, including detailed operational definitions, qualitative data analysis, and expert judges, only increase the likelihood that scores will measure the intended processes. To establish validity, the success of these efforts must subsequently be examined. It is the result of this analysis that provides evidence of validity.

One of Lissitz and Samuelsen’s most contentious points is that correlations are useless for matters of score validity. To illustrate their point, they offer several examples of physical measurements of length and height. In the length example, they reject the need for correlations to establish the meaning of *length*, a concept that they argue exists purely by the application of a ruler to an object. However, one glaring difference between measurement of physical traits, such as length, and measurement of those of interest in psychology, such as mathematics ability, is readily apparent. Length is both observable and objectively definable; there is an agreed-upon meaning of length and of the correspondence between the numbers read from a ruler and length. A ruler is an accepted test of length because it has been validated; that is, it has been shown to support inferences that are accurate. In contrast, most latent psychological traits, including abilities and knowledge, are difficult to measure not only because they are latent

(unobservable) but also because there is not a universal definition. This is evident both in educational settings—for which different test developers and users could argue that a variety of skills are or are not a part of a particular ability—and in psychological contexts, for which competing models (i.e., different dimensional structures) may be held by competing theorists. The validity of a measure of these types of variables (i.e., nonphysical, unobservable) must inherently be evaluated relative to the definition of the trait adopted by those interpreting and using the test scores. In the measurement of length, we know that its standing is not related to width or other variables. The point is that for the latent traits, we do not know where the relationships exist and where they do not. If the ruler is broken or error prone, the inaccuracies of the measurement can be directly observed; no test to determine this fact is needed. The only means of determining if our educational and psychological tests are “broken” is to empirically question the accuracy of score-based interpretations. According to Borsboom and Mellenbergh (2007),

The reliance on a process of measurement and the associated measurement model usually involves a degree of uncertainty; the researcher assumes, but cannot know for sure that a measurement procedure is appropriate in a given situation. . . . They make assumptions. Like all assumptions, these can be questioned. (pp. 85–86)

This questioning, according to Borsboom and Mellenbergh, is the “problem of test validity” (p. 86).

In a more educationally relevant example, Lissitz and Samuelsen describe a situation wherein a math test and an English test are highly correlated with one another. This correlation, they would argue, is irrelevant in terms of the validity of scores from either test. The abilities and skills measured by the math test are established by the fact that the processes required to solve the test questions include mathematical processes. Similarly, the English test is valid if it includes skills relevant to a language arts curriculum. According to Lissitz and Samuelsen, the correlation between a math and English test should not be interpreted as suggesting that they measure the same concept. In fact, that is exactly what the correlation may suggest. The shared variance suggests that there is some common underlying trait contributing to answers on both tests. It may be test-wiseness, a method factor due to the use of a single item format, a shared amount of verbal fluency that influences responses to both math and English test items, or even a general intelligence factor such as *g*. Further investigation is then warranted to determine the common underlying cause and the extent to which it weakens interpretations of the scores from either of the tests as measures of English or mathematics. Although spurious correlations between unrelated variables are observed in the social sciences, not all correlations are meaningless and uninterpretable. In fact, the majority of correlations suggest relationships that can have substantive explanations that are useful for identifying the construct of interest.

External correlation-based methods, including multitrait-multimethod analyses, are useful because they reveal patterns that support conclusions regarding what is actually measured. Without evidence relating consistency and covariance among item responses, we can only hypothesize that tasks measure certain content/constructs; we cannot empirically evaluate it. In the previous example of math and English test scores, the authors’ assertion that test development

should proceed from a notion (one might even say a theory) about how individuals differ in mathematics is no doubt true. But their conclusion that a test designed from this perspective necessarily leads to items that reflect these, and only these, differences is unsubstantiated. Unintentional sources of variance intrude in item responses, even with the most careful item construction. The purpose of building a validity argument is to determine the extent to which the item development procedures are effective in generating items that reflect individual differences on the focal construct.

I suspect that the authors’ dissatisfaction with correlations as evidence of validity stems from their acceptance of the historical definition of constructs as existing solely relative to other variables in the nomological network. They argue that to examine the relationship among test scores and other measures is merely “theory building,” which is outside the scope of validity investigations. The point that Lissitz and Samuelsen overlook in this argument is the distinction among possible purposes for examining correlations among measures. Theory building for theory building’s sake, which Lissitz and Samuelsen presume to be the goal of correlation-based validity analyses, is indeed outside the scope of validity investigations. However, using theories to form hypotheses regarding expected patterns of correlations, if tests are representations of specified constructs, is good argumentation:

Construct validity also subsumes content relevance and representativeness as well as criterion-relatedness, because such information about the content domain of reference and about specific criterion behaviors predicted by the test scores clearly contributes to score interpretations. (Messick, 1989, p. 17)

Messick goes on to say that when considered in a larger context of score meaning, as one of several sources of evidence, correlations between test scores and criterion measures contribute to construct validity.

Finally, although I may not agree that test development procedures and internal evaluation of content are sufficient for test validity, I concur with Lissitz and Samuelsen that, in practice, little attention is paid to internal evidence of score meaning. Researchers writing items to measure their variables of interest, teachers writing items to measure classroom learning, and even test developers are typically so interested in collecting student responses that they overlook some critical steps in test and scale development. Much of the recent literature on test development, including but not limited to examinations of validity, have devoted significant attention to the importance of domain-specific and general cognitive theory. Contrary to the argument advanced by Lissitz and Samuelsen, contemporary views of test design, as detailed in the National Research Council’s (2001) *Knowing What Students Know: The Science and Design of Educational Assessment*, show that hypothesis testing with strong theories of the construct can improve the quality of educational test scores and interpretations. Psychometricians examining high-stakes tests such as the SAT, GRE, and TOEFL, have shown that cognitive theory, when joined with rigorous psychometric methods, can facilitate efficient item generation (Gorin, 2005; Enright & Sheehan, 2002), augment diagnostic score information (Huff, 2006; VanderVeen et al., 2006), and provide additional evidence

regarding the meaning of tests scores for individuals and groups of students (Gierl, Tang, & Wang, 2005; Sheehan & Ginther, 2001).

### Validity Is About Scores, Not Interpretations

One of the most significant changes in contemporary validity theory, and I would argue one of the most beneficial to test development and use, has been the emphasis on validity in terms of “appropriate inferences” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). The role of inference and interpretation has added complexity to validity examinations. Validity evidence, if considered only in terms of the test itself, is meaningless. As stated succinctly by Messick (1989), “Evidence takes form only in answer to questions” (p. 16). One piece of evidence may be relatively unpersuasive, given a particular question of interest. For another question, however, the same piece of evidence could be quite convincing. The complexity of the validity study comes from determining which evidence is most persuasive and how to gather such evidence. Lissitz and Samuelsen hold that the evidence of validity is independent of the question or inference of interest, that it is no more or less than the meaning of test scores—a view that I had hoped was of interest only in discussions concerning the history of educational measurement.

Consider an exaggerated example of a test of fifth-grade mathematics. If a group of fifth-grade math teachers examined the content of the test and judged it to include all mathematics skills included in the fifth-grade mathematics curriculum, according to Lissitz and Samuelsen the test would be valid. If the principal decided to use the test to select students for the Language Arts Gifted and Talented program, Lissitz and Samuelsen would argue that the test is still valid, but has little utility. Is it necessary, or even logical, to make such a distinction? Based on the existing notion of validity as a judgment regarding the appropriateness of interpretations or uses of test scores for a given purpose, the fifth-grade mathematics test, regardless of the psychometric quality of the items or the representativeness of the curriculum, is not valid for placing students in the gifted program.

Now consider a more realistic placement testing application. When considering various tests for use in placement, a test user would typically consider a valid test to be one that accurately places individuals in the appropriate course or field. Several tests might exist, all of which have been shown to reflect a similar set of skills or abilities that are representative of the desired content. In selecting a specific test, however, evidence regarding the association between the scores on the test and success in subsequent performance is desirable. According to Lissitz and Samuelsen, such considerations are not a part of validity; rather, they pertain to issues of test utility: first, because the question of interest is one of test use, not score meaning, and second, because the analysis includes external measures.

Despite Lissitz and Samuelsen’s insistence that score meaning resides in tests and does not change, validity is more than score meaning. Even by the simplest definition, validity is the extent to which a test measures what it purports to measure. Evidence of the meaning of scores is only half of the validity equation. Ferrara et al. (2004) frame validity in terms of the alignment of two types of information: the intended knowledge, skills, processes, and

strategies (KSPS) about which the test user would like to make inferences and the observed KSPS, which are applied by the examinees when solving an item. Validity can be judged in terms of the relationship between these two types of information. Gorin (2006) revises the terminology, referring to the *intended construct* that is specified a priori in terms of a construct definition and the *enacted construct* that is actually measured. Validity investigations provide evidence of the alignment between these two constructs.

### Practicality of the Validity Frameworks

With relatively few exceptions, the issues pertinent for validity analysis presented by Lissitz and Samuelsen are also those of interest to construct-oriented validity models. The differences are more in the details. Lissitz and Samuelsen distinguish theoretical versus practical and internal versus external characteristics of tests, some of which they argue are relevant for validity investigations and others for utility. However, their recommended process for validation remains vague, at best. The strongest criticism of current validity theory, specifically the unified view, is the lack of specific guidelines for test developers interested in implementing the models. Although Lissitz and Samuelsen underscore the point that their framework is not a unified theory, the distinction is once again one of semantics. Lissitz and Samuelsen’s differentiation between internal and external investigative focuses, as well as between practical and theoretical perspectives, harkens back to Messick’s (1989) facets of validity. The proposed change amounts more to a reshuffling of analyses and sources of evidence than to a unique perspective on validity. Furthermore, even with Lissitz and Samuelsen’s text and figures, which purport to make clear the process of validation, both the structure and process of the classification are unclear. If we are to consider the changes proposed by Lissitz and Samuelsen, the advantages of the process must be made evident. Clarification of their framework and contradictions between the text and the figures should be resolved if these ideas are to take hold. For example, within their framework, differential item functioning (DIF) is an issue of reliability. Given that DIF is often considered a matter of dimensionality and generalizability, the latent processes classification would seem more appropriate. Their divergence from this tradition may require further explanation. More glaringly, Lissitz and Samuelsen contradict themselves on a key issue, the role of correlations in validity. Although in their text they adamantly discourage the use of external correlations as validity evidence, in their list of potential sources of evidence regarding the internal factors to consider for content validity, they include “convergent/divergent evidence from correlations” (Table 1, p. 441) as a potential source of evidence of the theoretical perspective of internal test properties. If they are uncertain as to where specific sources of evidence belong in their framework, it is likely that practitioners will be as well.

I approached Lissitz and Samuelsen’s article seeking discussions of the challenges in understanding the nuances of validity theory. Moreover, I hoped to find (as promised by Lissitz and Samuelsen early in the article) specific tools to guide the use of validity theory in practice. This theme appears only transiently throughout the article. The thrust of the critique was aimed at defeating the notion of construct meaning in favor of content review and a shift away from the importance of score interpretations. Although I have dedicated much of my commentary to the strengths of the unified validity theory, its complexity undeniably

interferes with its utility. Lissitz and Samuelsen paint a picture of general dissatisfaction with current validity theory; barring selected issues, including the notion of consequences of test use as part of validity theory, I am more optimistic about the usefulness of existing frameworks for test development. Rather than discarding the existing validity models that have advanced methods of test development and the quality of assessment, perhaps the authors' suggestion to "change the definition of construct validity to minimize the role of the nomological network that deals with the development of the theory" (pp. 441–442) is more valuable. I would argue that a review of current validity theory suggests that such a move has already occurred. Both the unified and argument-based models of validity, when compared with original conceptions of construct validity, stress the importance of multiple sources of validity evidence, one of which might include correlations with external factors. Correlations and regressions with external variables as evidence of the external aspect of validity constitute only one source of evidence for only one of the six validity aspects. To argue that the correlations play the load-bearing role in the unified view of validity is a misrepresentation.

Returning to my opening comments regarding students' difficulties in applying validity theories, let us consider both the existing and the proposed validity frameworks. Although I applaud Lissitz and Samuelsen for their efforts to improve the accessibility of validity theory for practitioners, I offer a different solution to test developers' struggles. Rather than unnecessarily simplifying validity theory to accommodate only those analyses and evidence that current psychometricians and test developers are adept at producing, perhaps the real change should come in the types of analyses for which people are trained. Recent test design research suggests that methods and theories of cognitive psychology can improve score meaning (Gorin, 2006). These methods should become a part of the psychometrician's repertoire. Or, alternatively, test developers should invite experts in cognitive psychology to assist with item and test design. Neither of these recommendations is new. Leighton's (2004) reference to the hybrid psychometrician as the ideal measurement specialist alludes to the need for professionals with more diverse expertise. Validity theory has advanced since the 1950s. Perhaps it is time for the measurement curriculum to catch up.

## Conclusion

Validity is not a box to be checked yes or no. Nor is it a checklist that once marked is a fait accompli. Validity is a judgment, and like all judgments it is relative and ever evolving. It can be, and should be, evaluated in light of new evidence and desired interpretations, making validity and validation an ongoing process. For any assessment situation, there are typically multiple available tests. What are the most persuasive forms of evidence for various types of interpretations? What are the appropriate procedures for gathering these pieces of evidence? Answering these questions is the key to validity evaluations. Currently, we have left the resolution of these questions up to the test users, the majority of whom are ill prepared to systematically address them. As assessment specialists and measurement theorists, our responsibility is to provide tools and recommendations for how these judgments can be made. Our theoretical validity frameworks, including the one proposed by Lissitz and

Samuelsen, offer information regarding the types of validity evidence that are most useful for various testing situations. Empirically based recommendations to help guide test developers in selecting the most useful analysis are critically lacking. How do factors such as the nature of the construct to be measured and the intended use of the test affect choices in sources of evidence? Should validation be approached similarly for tests with established statistical and psychometric properties that are to be used in new ways and for newly developed instruments about which little is known? For readers who hoped to find some assistance with these questions, I have carefully avoided that daunting task in this commentary. Perhaps in a follow-up to this important issue of *Educational Researcher*, measurement scholars and psychometricians will tackle that larger challenge.

## NOTE

A word of thanks to Roy Levy and Samuel Green. Many of the arguments presented arose from informal conversations about validity, and I want to credit their efforts and thoughts.

## REFERENCES

- 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.
- Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice*, 17(3), 37–45.
- 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.
- Bishop, D. V. M., North, T., & Donlan, C. (1996). Nonword repetition as a behavioral marker for inherited language impairment. *Journal of Child Psychology and Psychiatry*, 130, 391–403.
- Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 85–118). New York: Cambridge University Press.
- Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, 11(1), 33–63.
- Carroll, J. B., & Maxwell, S. E. (1979). Individual differences in cognitive abilities. *Annual Review of Psychology*, 30, 603–640.
- Conti-Ramsden, G. (2003). Processing and linguistic markers in young children with specific language impairment (SLI). *Journal of Speech, Language, and Hearing Research*, 46, 1029–1037.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Denkla, M. B., & Rudel, R. G. (2007). Rapid automatized naming (R.A.N.): Dyslexia differentiated from other learning disabilities. *Neuropsychologia*, 14, 471–479.
- Devescovi, A., & Caselli, M. C. (2007). Sentence repetition as a measure of early grammatical development in Italian. *International Journal of Language and Communication*, 42(2), 187–208.
- Dollaghan, C. A. (1987). Fast mapping in normal and language-impaired children. *Journal of Speech and Hearing Disorders*, 52, 218–222.
- Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, 41, 1136–1146.
- Embretson, S. E. (1983). Construct validity: Construct representation vs. nomothetic span. *Psychological Bulletin*, 93(1), 179–197.

- Embretson, S. E. (1994). Application of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107–135). New York: Plenum Press.
- Embretson, S. (1995). Developments toward a cognitive design system for psychological and educational tests. In D. Lubinsky & R. Dawis (Eds.), *Assessing individual differences in human behavior: New concepts, methods and findings* (pp. 17–48). Palo Alto, CA: Consulting Psychologist Press.
- 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. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, 64, 407–433.
- Embretson, S. E., & Gorin, J. S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, 38(4), 343–368.
- Enright, M. K., & Sheehan, K. M. (2002). Modeling the difficulty of quantitative reasoning items: Implications for item generation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 129–157). Mahwah, NJ: Lawrence Erlbaum.
- Ferrara, S., Duncan, T. G., Freed, R., Vélez-Paschke, A., McGivern, J., Mushlin, S., et al. (2004, April). *Examining test score validity by examining item construct validity: Preliminary analysis of evidence of the alignment of targeted and observed content, skills, and cognitive processes in a middle school science assessment*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Gierl, M. J., Tang, X., & Wang, C. (2005). *Identifying content and cognitive dimensions on the SAT* (College Board Research Report No. 2005–11). New York: College Board Press.
- Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, 42, 351–373.
- Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, 25, 21–35.
- Huff, K. (2006, April). *Using item difficulty to inform descriptive score reports*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.
- Irvine, S. H. (2002). The foundations for item generation for mass testing. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 3–24). Mahwah, NJ: Lawrence Erlbaum.
- Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, 14, 389–433.
- Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, 23(4), 6–15.
- 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.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.
- Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–305). Westport, CT: American Council on Education/Praeger.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3–67.
- Mitroff, I. I., & Sagasti, F. (1973). Epistemology as general systems theory: An approach to the design of complex decision-making experiments. *Philosophy of Social Sciences*, 3, 117–134.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Pellegrino, J. W. (1988). Mental models and mental tests. In H. Wainer & H. I. Brown (Eds.), *Test validity* (pp. 49–59). Hillsdale, NJ: Lawrence Erlbaum.
- Sheehan, K. M., & Ginther, A. (2001, April). *What do passage-based multiple-choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on the current TOEFL reading section*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for education measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–331). New York: American Council on Education/Macmillan.
- VanderVeen, A., Huff, K., Gierl, M., McNamara, D. S., Louwerse, M. M., & Graesser, A. C. (2006). Developing and validating instructionally relevant reading competency profiles measured by the critical reading section of the SAT. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theory, interventions, and technologies* (pp. 137–174). Mahwah, NJ: Lawrence Erlbaum.
- Yang, X., & Embretson, S. E. (2007). Construct validity and cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 119–145). New York: Cambridge University Press.

#### AUTHOR

JOANNA S. GORIN is an assistant professor at Arizona State University, Mary Lou Fulton College of Education, P.O. Box 870211, Tempe, AZ 85287; [joanna.gorin@asu.edu](mailto:joanna.gorin@asu.edu). Her research focuses on cognitively based approaches to assessment design and psychometric modeling.

Manuscript received September 4, 2007

Accepted September 17, 2007