



Comments on Lissitz and Samuelsen

Reconstructing Validity

by Pamela A. Moss

In response to Lissitz and Samuelsen (2007), the author reconstructs the historical arguments for the more comprehensive unitary concept of validity and the principles of scientific inquiry underlying it. Her response is organized in terms of four questions: (a) How did validity in educational measurement come to be conceptualized as unitary, and why? (b) What is construct validity, and how does it provide the basis for a unitary concept of validity? (c) Why has the focus of validity been on the interpretations and uses of test scores rather than on the test itself? and (d) What sort of guidance for test developers and evaluators has been provided within a unitary concept of validity, and how might it be enhanced? The author highlights the role that cases of programmatic validity research can play in representing validity theory and guiding validity inquiry.

Keywords: construct validity; educational measurement; testing; validity

In “A Suggested Change in Terminology and Emphasis Regarding Validity and Education,” Robert W. Lissitz and Karen Samuelsen (this issue of *Educational Researcher*, pp. 437–448) raise a concern that the unitary concept of validity, prominent in the measurement literature (e.g., American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; Messick, 1989b), does not provide adequate guidance for test evaluation. To address this problem, they propose a deconstruction of validity that involves (a) a shift in terminology such that the term *validity* is reserved “for the establishment of the definition of the trait . . . in the test development phase and the analysis of its stability” (p. 444) and (b) an elaboration of what it means to do this kind of work—“a set of approaches . . . that lead researchers through a process that strongly supports or does not support their claims” (p. 444)—which they variously refer to as *content validity*, *internal validity*, and simply *validity*. They also address more briefly what they call the *external* aspects of test evaluation—attention to the intent or purpose of testing, evidence about correlations with other measures, and evidence about the impact or consequences of testing—which, they argue, should be excluded from the definition of validity. Although theoretical considerations have a role to play

in their representation of both internal and external aspects of test evaluation, they advocate separating theoretical and practical considerations at a conceptual level.

The issue of how validity theorists can provide better guidance to those who develop and evaluate tests is an important one, and I believe the field would be well served by following more closely the (instructively different) examples of validation to which Lissitz and Samuelsen point in the work of scholars such as Kane, Mislevy, and Wilson. However, I worry that the conception of validity implied in the terminology that Lissitz and Samuelsen propose appears to move away from a generative understanding of validation as scientific inquiry reflected in the unitary approach, (back) toward a representation of validity in terms of general methodological prescriptions that the unitary approach was intended, in part, to overcome. I am also concerned that Lissitz and Samuelsen’s representation of the unitary concept of validity—based in the work of Cronbach, Messick, and the 1999 Standards (AERA et al.)—glosses over some key elements of the approach and illustrates it polemically with the sorts of examples that Cronbach (1988, p. 12) and Messick (1989b, p. 49) would have criticized as “weak.” Thus Lissitz and Samuelsen stop short of illuminating some of the potential consequences of their shift in terminology for the ways that test developers and evaluators think about and engage in validity work.

In my response I will build an argument for the value of a unitary concept of validity, highlighting some of the key elements from Cronbach, Messick, the Standards, and others that Lissitz and Samuelsen have glossed over. As I will illustrate, a unitary conception of validity is in no way inconsistent with the provision of substantial guidance, nor does it preclude the making of well-reasoned, practical judgments about what can and should be undertaken before (and after) a test is put into operational use. Drawing on the seminal work of Cronbach (1988, 1989), Kane (1992, 2006), and Shepard (1993, 1997), I will sketch a somewhat different direction for a solution to the problem of providing more practical guidance. By working in this direction, we can take full advantage of the promising examples to which Lissitz and Samuelsen point while also maintaining close ties to the generative vision of scientific inquiry reflected in the unitary conception of validity.

My response is organized in terms of four questions:

1. How did validity in educational measurement come to be conceptualized as unitary, and why?
2. What is construct validity, and how does it provide the basis for a unitary concept of validity?

Educational Researcher, Vol. 36, No. 8, pp. 470–476
DOI: 10.3102/0013189X07311608
© 2007 AERA. <http://er.aera.net>

3. Why has the focus of validity been on the interpretations and uses of test scores rather than on the test itself?
4. What sort of guidance for test developers and evaluators has been provided within a unitary concept of validity, and how might it be enhanced?

Given the general readership of *Educational Researcher*, I think it is important to acknowledge that Lissitz and Samuelsen's title implies a considerably broader conversation than the one in which we are actually engaged. We are not talking about "validity and education," a topic that would span the wide range of research discourses through which educational research is practiced, or even "validity and educational assessment," which involves multiple sources of evidence relevant to students' learning (AERA et al., 1999, p. 3) and interpretive practices that spill over the boundaries of educational measurement (Moss, 2007; Moss, Girard, & Haniford, 2006). Rather, we are talking about validity in the context of the development and evaluation of standardized tests, where the goal is to produce scores that are (arguably) comparable across individuals and contexts.¹ That said, because understandings of validity shape the sorts of testing practices likely to be found sound, and these practices, in turn, contribute in direct and indirect ways to students' opportunities to learn (Moss, Pullin, Haertel, Gee, & Young, in press), *Educational Researcher* provides a fruitful forum for inviting multidisciplinary engagement in this debate.

An Argument for a Unitary Concept of Validity

My argument for a unitary concept of validity is organized in terms of the four questions listed earlier. I discuss key features articulated by proponents of a unitary approach, point to apparent differences in interpretation that I have with Lissitz and Samuelsen, and for the fourth question, suggest an alternative direction for enhancing the guidance available to test developers and evaluators.

1. How did validity in educational measurement come to be conceptualized as unitary, and why?

An understanding of the unitary concept of validity—and the consequences of abandoning it—requires an understanding of what, in fact, was being unified and why.² Developments in validity theory in educational measurement can be traced across successive editions of two seminal publications: (a) *Standards for Educational and Psychological Testing* (as it is now called), jointly sponsored by AERA, APA, and NCME, and (b) the validity chapter in *Educational Measurement*, sponsored by the NCME and the American Council on Education. Five editions of the Standards have been published—in 1954/1955 (when separate documents were developed for psychological and achievement tests), 1966, 1974, 1985, and 1999—and a sixth is currently being planned (see www.teststandards.org).³ Each edition of the Standards is drafted by a committee of measurement scholars jointly appointed by the sponsoring organizations. The process of developing an edition generally takes multiple years, as drafts are submitted to the sponsoring organizations and the field for review and comment. There have been four validity chapters written since 1951, by Cureton (1951), Cronbach (1971), Messick (1989b), and Kane (2006). The authors of these chapters typically are selected by an editorial board for their

sustained and seminal contributions to validity theory. Drafts of the chapters are peer reviewed; however, unlike the Standards, the chapters are expected to represent the individual authors' perspectives on validity. Although there has been a lively discourse in the measurement literature about validity, these documents tend to be treated as somewhat canonical (as reflected, for instance, in their widespread citation in measurement textbooks), albeit not without dissent, some radical, some more nuanced, and they tend to serve as the representation of validity within or against which alternative conceptions must be situated (as our dialogue illustrates).

When the first edition of the Standards was published in 1954, validity was conceptualized in terms of different aims of testing, each associated with a different type of validity investigation. This conception of validity persisted in similar form through the first three editions of the Standards (1954/1955, 1966, and 1974). The authors of the 1966 Standards characterized three types of validity:

- *Content validity* demonstrated how well a test "samples the class of situations or subject matter about which conclusions are to be drawn."
- *Criterion validity* compared test scores with "one or more external variables considered to provide a direct measure of the characteristic or behavior in question." (In the 1954/1955 Standards, predictive and concurrent validities had been represented as separate types of [criterion] validity, depending on whether the criterion was administered at the same time as the test in question or at a later time.)
- *Construct validity* served the aim of inferring "the degree to which the individual possesses some hypothetical trait or quality (construct) . . . that cannot be observed directly" by determining "the degree to which certain explanatory concepts or constructs account for performance on the test . . . through studies that check on the theory underlying the test" (APA, 1966, pp. 12–13).

As fleshed out in the 1966 Standards, studies of construct validity

check on the theory underlying the test. The procedure involves three steps. First, the investigator inquires: From this theory, what hypotheses may we make regarding the behavior of persons with high and low scores? Second, he gathers data to test these hypotheses. Third, in light of the evidence, he makes an inference as to whether the theory is adequate to explain the data collected. (APA, 1966, p. 13)

When Loevinger (1957) used the term *ad hoc* to criticize content, predictive, and concurrent validities (in the passage that Lissitz and Samuelsen cite, p. 440), she was raising a concern about identifying validity in terms of particular procedures associated a priori with particular aims of testing. Content, predictive, and concurrent categories, she argued, were possible supporting evidence for construct validity, which subsumed them and much more. Thus the categories were not logically distinct or of equal importance. Moreover, the presentation implied that the categories represented options rather than components of validity. Only construct validity, she argued, provided a scientifically useful basis for establishing validity. Her argument prefigured the move toward

a unitary concept of validity reflecting the scientific principles of construct validity that Lissitz and Samuelsen are seeking to deconstruct.

Thus construct validity (Cronbach & Meehl, 1955) was initially viewed as a type of validity to be brought to bear when no content domain or criterion variable was sufficient to operationally define the meaning of test scores. In his 1971 validity chapter, Cronbach gave construct validity far more centrality in his general conception of validity than had the Standards. And he highlighted the important role of evaluating counterhypotheses—viable alternative interpretations of score meaning—for guiding validity research. Although he maintained the relevance of the now familiar three types of validity inquiry (content-, construct-, and criterion-related) as aspects of a validity inquiry, he likened validity research to the evaluation of a scientific theory as characterized in *construct validity*, and he argued that most educational tests entailed constructs: “Whenever one classifies situations, persons, or responses, he uses constructs” (p. 462). The 1985 Standards (AERA, APA, & NCME) moved in this direction, articulating more prominently a unified conception of validity that draws on multiple kinds of evidence to evaluate the inferences and uses of test scores. The authors renamed the traditional three categories to emphasize their role as types of evidence (construct-, content-, and criterion-related *evidence*) rather than types of validity: “An ideal validation includes several types of evidence, which spans all three of the traditional categories” (p. 9). The third validity chapter (Messick, 1989b) represented the completion of the move to a unified conception of validity as scientific inquiry into score meaning, as articulated in the description of construct validity; the 1999 Standards reflect this vision. Similarly, Kane’s (2006) fourth validity chapter continued in this tradition, reflecting (as had all the others) some revisions to the conceptual framework guiding validity work. As cited by Lissitz and Samuelsen, the authors of the 1999 Standards defined validity as “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” and named five categories of evidence for researchers to consider in building a validity argument. The categories were for evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing.⁴ The authors of the Standards noted that “professional judgment guides decisions regarding the specific forms of evidence that can best support the intended interpretation and use” (AERA et al., 1999, p. 11).

2. What is construct validity, and how does it provide the basis for a unitary concept of validity?

As Lissitz and Samuelsen point out, representations of validity as a unitary concept have evolved from representations of construct validity. However, their characterization of construct validity appears to overemphasize one particular kind of evidence—correlations among different measures—and to downplay the role of theory-based hypothesis testing that underlies the representation of construct validity. Although they acknowledge, briefly, that construct validity entails the tying of constructs to observables (which, I note, is central to their conception of internal validity), their other references to construct validity focus on the evidence of the relationship among measures.

Cronbach and Meehl (1955) characterize the process of construct validation as follows:

Construct validation takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are a means of confirming or disconfirming the claim. . . . The fundamental principles are these: 1. Scientifically speaking, to “make clear what something *is*” means to set forth the laws in which it occurs. We shall refer to the interlocking system of laws which constitute a theory as a *nomological network*. 2. The laws in a nomological network may relate (a) observable properties or quantities to each other; or (b) theoretical constructs to observables; or (c) different theoretical constructs to one another. These “laws” may be statistical or deterministic. (p. 290)

Thus, for Cronbach and Meehl, “Construct validity is not to be identified solely by particular investigative procedures, but by the orientation of the investigator” (p. 282). In light of Lissitz and Samuelsen’s presentation, it is important to note that Cronbach and Meehl listed multiple kinds of evidence as relevant to construct validity within a hypothesis-testing framework, including content validity, studies of process, interitem correlations, intertest correlations, test-criterion correlations, studies of stability over time, and stability under experimental intervention (see pp. 287–288, 300). (Later, Cronbach, 1988, also addressed the role that evidence from test development and from previously published studies plays in building a validity argument; p. 4.) Furthermore, Cronbach and Meehl highlighted the crucial role that theory-based hypothesis testing plays in guiding validity research: “High correlations and high stability may constitute either favorable or unfavorable evidence for the proposed interpretation, depending on the theory surrounding the construct” (p. 300). It is the theory that guides the selection and interpretation of evidence, including evidence of content validity. Thus construct validity subsumes the kinds of evidence that Lissitz and Samuelsen have characterized as content or internal validity but within a theoretically informed hypothesis-generating framework.

In building their argument against a unitary concept of validity, Lissitz and Samuelsen construct hypothetical examples of the (atheoretical) use of the correlational evidence that bear little resemblance to the scientific orientation that Cronbach and Meehl and their successors had in mind:

[1] In the previous example of a seventh-grade mathematics test, . . . its claim [of validity] . . . 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 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. (Lissitz & Samuelsen, p. 442)

[2] 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. (Lissitz & Samuelsen, p. 443)

These contentious examples of the use of correlational evidence illustrate, at best, the sort of “weak” approach to construct validity

that Cronbach (1989) criticized as “rak[ing] together miscellaneous correlations” (p. 155). He called instead for reporting “incisive checks into rival hypotheses, followed by an integrative evaluative argument” (p. 155). A researcher engaged in a strong program of construct validity would design and interpret correlational studies—or any other sort of evidence, including content-related evidence—in light of the fit with and challenge to their theoretically informed anticipations.

3. *Why has the focus of validity been on the interpretations and uses of test scores rather than on the test itself?*

Perhaps most fundamental to the understanding of validity is the question of what is being validated: a test, or an interpretation and use of test scores? Lissitz and Samuelsen raise this fundamental issue when they criticize Messick’s stance:

We also find Messick’s (1989[b]) 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. (p. 442)

The argument that validation focuses on interpretations and uses of tests rather than on tests themselves has been prominent in the measurement literature for a long time. Here is the argument as Cronbach framed it in 1971:

The phrase validation of a test is a source of much misunderstanding. One validates, not a test, but an interpretation of data arising from a specified procedure. A single instrument is used in many different ways—Smith’s reading test maybe used to screen applicants for professional training, to plan remedial instruction in reading, to measure the effectiveness of an instructional program, etc. Since each application is based on a different interpretation, the evidence that justifies one application may have little relevance to the next. Because every interpretation has its own degree of validity, one can never reach the simple conclusion that a particular test “is valid.” (p. 447)

This position has been prominently stated in the testing standards and validity chapters ever since. One can trace similar arguments back to Cronbach and Meehl (1955, p. 297) and to the earliest Standards, which acknowledge that the same test can be used for different aims (e.g., APA, 1954, p. 13). The National Research Council (2002) even uses the development of this understanding as an example of the “accumulation of knowledge” in its report *Scientific Research in Education*:

At first, validity was viewed as a characteristic of the test. It was then recognized that a test might be put to multiple uses and that a given test might be valid for some uses but not for others. That is, validity came to be understood as a characteristic of the *interpretation and use* of test scores, and not of the test itself, because the very same test (e.g., reading test) could be used to predict academic performance, estimate the level of an individual’s proficiency, and diagnose problems. Today, validity theory incorporates both test interpretation and use (e.g., intended and unintended social consequences). (p. 35)

Although I would not have represented an evolving philosophical perspective as “accumulation of knowledge,” I endorse

the National Research Council’s positive evaluation of the perspective as representing an important development that acknowledges the ways in which tests are used and that supports sound testing practice. To argue this point from one of Lissitz and Samuelsen’s own examples, they cite off-the-shelf tests that are used in different states (presumably with different state content standards) as instances where content-related evidence of validity is crucial. Wouldn’t the interpretation grounded in the standards and the associated judgments of validity likely be different, depending on the standards to which the published test was being compared? (It is noteworthy that the example also calls into question the use of the terms *internal* and *external* for distinguishing aspects of test development and evaluation.)

4. *What sort of guidance for test developers and evaluators has been provided within a unitary concept of validity, and how might it be enhanced?*

Lissitz and Samuelsen’s well-taken concerns about the feasibility of the vision of construct validity in Cronbach and Meehl (1955) have been voiced by numerous scholars, including Cronbach (1989) himself:

The idealized strong program is most appropriate to a scientific perspective that reaches centuries into the future. No science can live only for the day when truth becomes crystal clear. Social and behavioral scientists in particular are obligated to help their contemporaries think through problems and evaluate proposed solutions. (p. 163)

Cronbach (1971) highlighted the role of “plausible counterinterpretations” to score meaning for directing research toward possibly vulnerable parts of a theory and avoiding the interminable “plodding verification of every sentence written about a construct” (p. 464). Later, he suggested a set of criteria by which test evaluators might prioritize validity questions: (a) prior uncertainty about the issue, (b) information to be yielded by a feasible study compared with how much uncertainty will remain, (c) cost of the investigation in terms of time and dollars, and (d) leverage for achieving consensus about the use of the test in the relevant audience (Cronbach, 1989, p. 165). Shepard (1993) highlighted the crucial role of purpose in guiding validity inquiry and helping researchers set priorities in addressing validity questions most relevant to the context of use.

An important issue at work in our dialogue is what expectations are set a priori and what are left to professional judgment in light of the particular circumstances—purposes, intended interpretations, formats, contexts, and so forth—of test development and implementation. As I indicated in my introduction, I worry about the consequences of framing validity in terms of “a set of approaches . . . that lead researchers through a process” (Lissitz & Samuelsen, p. 444). Similarly, the National Research Council’s (2002) *Scientific Research in Education* criticized proposed federal legislation for “mandating a list of ‘valid’ scientific methods . . . erroneously assum[ing] that science is mechanistic and thus can be prescribed” (p. 130). Against the language of the proposed legislation, the National Research Council argued that “it is the [self-regulating norms of the] scientific *community* that enables scientific progress, not . . . adherence to any one scientific *method*” (p. 19). Although the approaches that Lissitz and Samuelsen cite positively provide

educative examples of good validity work, the frame in which they are located risks losing the generative, flexible, and critical principles of scientific inquiry.

For an alternative strategy through which sound (feasible) expectations for validity research might be promoted—one that would allow us to take advantage of the promising examples to which Lissitz and Samuelsen point but within a flexible framework of scientific inquiry—I draw initially on the work of Cronbach (1988, 1989), Shepard (1993), and Kane (1992, 2006). Kane and Shepard, building on Cronbach's conception of validity argument, called for researchers first to specify what I will call a comprehensive plan or research agenda for validity work and then to use it to guide decision making about what evidence to pursue. As Kane (2006) described it, the plan ("interpretive argument") lays out "the network of inferences and assumptions leading from the observed performance to the conclusions and decisions based on the performances" (p. 23). Although he called for the plan to be comprehensive, he noted that it was unlikely that all inferences would be evaluated (or the associated hypotheses tested): Some would be taken for granted (e.g., that students can read the assignment); some might be deemed unfeasible; and some might be deemed unnecessary in light of the purposes of testing. These are not decisions that can be made in advance; the details of the interpretive argument depend on "the specific interpretation being proposed, the population to which the interpretation is applied, the specific data collection procedures being used, and the context in which measurement occurs" (Kane, 1992, p. 529). For instance, although low-stakes tests may require only evidence gathered during the development stage, high-stakes tests appropriately require a more extensive evaluation of the fully developed test in use. The comprehensive plan is used to guide decisions about validity research. The availability, for public and professional review, of the plan, the evidence produced, the evidence deemed unnecessary or impractical, the rationale, and the conclusions—the validity argument—would provide one means to encourage accountable practice. The 1999 Standards adopted a similar perspective, calling for the development of "a set of propositions that support the proposed interpretation for the particular purposes of testing" (AERA et al., 1999, p. 9) but were far less explicit about the potential role of a more comprehensive plan for illuminating gaps in the evidence.

But how might representations of validity theory—which tend, as Kane notes, to be somewhat abstract—provide guidance to support test developers and researchers in developing and carrying out sound plans of validity work? Many different kinds of guidance have been provided under the unitary conception of validity in educational measurement, and, I acknowledge, there is much more that could be done. The kinds of guidance provided include (a) lists of categories of types of evidence, inferences, or aspects of validity, often illustrated with examples of the kinds of studies that might be undertaken (e.g., AERA et al., 1999; Cronbach, 1988; Kane 1992, 2006; Messick, 1989b); (b) principles to guide choices among the myriad kinds of evidence that are arguably relevant to a given interpretation or use (e.g., Cronbach, 1989; Kane, 2006; Shepard, 1993); (c) standards or guidelines about the nature of evidence that should be made available to enable professional judgment (e.g., AERA et al., 1999; Educational Testing Service, 2002); (d) outlines of "interpretive arguments" (Kane, 2006) or comprehensive plans

for validity research for particular types of interpretations and uses of tests (e.g., an algebra placement test), accompanied by examples of the types of evidence that might be or have been developed under the plan (Kane, 1992, 2006; Shepard, 1993); (e) descriptions and (critical) analyses of actual programs of validity research, associated with a particular test or construct (e.g., Cronbach, 1989, p. 150; Shepard, 1993, pp. 432–443); and (f) frameworks illustrated with extended examples (Kane, 2004, 2006; Mislevy, Steinberg, & Almond, 2003; Wilson, 2005) that take us from conceptualization through test development and implementation (some of which Lissitz and Samuelsen have productively cited).

The kind of work cited above under categories (d), (e), and (f) has made crucial (if underutilized) contributions to the representation of validity. These studies go beyond isolated examples of categories of evidence (as, for instance, the 1999 Standards provide) to provide extended examples of how programs of validity work might unfold—how a validity argument can be built across multiple sources of evidence and challenges to the intended interpretations and uses of tests. They are rich examples from which we can learn. That said, the examples tend to represent either extremely mature or prototypical programs of research and development that may be aspirational for many developers and users. The educational measurement field would be well served, I believe, if we also had multiple exemplars, critically analyzed, of programs of validity research for common testing uses. One recommendation I would make to the committee revising the 1999 Standards is to consider, if possible, either the development of such cases of validation or the commissioning of their development by working groups of professionals, to show what sound programmatic work consistent with the Standards might look like. Such cases could be created from elements of existing programs of research so that no particular program of work is identified or singled out for special attention. The interpreters of the cases could also build arguments for whatever additional evidence they consider essential for a given purpose and what may be more aspirational. Reaching consensus about programs of validity research serving particular purposes in particular contexts will likely be more difficult than reaching consensus about abstract standards or principles, but having such professionally debated exemplars would go a long way toward illustrating what validity principles actually mean in practice and identifying important gaps in conventional programs of test development and evaluation. The cases would thus provide examples of what sound validity work might look like in various testing contexts and would provide sets of principles that test developers and validity researchers could use to learn from and reason with in their own (always partially unique) contexts of work. I should note, as well, that our current dialogue about how best to represent validity theory in educational measurement would be enhanced if we were to explore, comparatively, the implications of different choices in concrete contexts of practice.

Concluding Comments

A theory of validity represents both a philosophical perspective and a set of conceptual tools that shape our thinking and action (Moss et al., 2006). A change in terminology almost inevitably implies a change in practice. The separations that Lissitz and Samuelsen propose for the way we think about validity—between the theoretical

and the practical, the internal and the external, the purpose of testing and the development of the test—risk losing the power of the principles of scientific inquiry to guide the evaluation and postponing efficient attention to the particular kinds of evidence most relevant to the uses of a test and the interpretations implied therein. Although some testing practices may be quite adequately supported by the kinds of activities that Lissitz and Samuelsen want to locate under the label of internal validity, for others, particularly those that affect the rights and life chances of individuals or the resources and viability of institutions, a higher standard of scrutiny is essential.

I worry as well that circumscribing the term *validity* to include only certain kinds of evidence risks releasing those responsible for large-scale testing programs from the obligation to ask challenging questions about the fit between the tested domains and the domains to which we want to generalize in education. As Kane (2006) noted,

The target domains of most interest in education are not restricted to test items or test-like tasks, although they may include this kind of formal performance as a subset. A person's level of literacy in a language [for instance] depends on his or her ability to perform a variety of tasks in a variety of contexts, ranging from the casual reading of a magazine to the careful study of a textbook or technical manual. These performances can occur in a variety of locations and social situations. (p. 31)

Questions about the educational importance of what is measured—about the extrapolation from a test domain to the domain of cognition or action outside the testing situation to which we want to generalize (Mislevy, Gee, & Moss, in press)—is a point that has been made in one way or another in each of the validity chapters, beginning with Cureton (1951), who cautioned that if “immediate [tested] objectives lack ultimate relevance, such tests retard educational progress instead of stimulating it” (p. 654). Although empirical evidence relevant to this issue may not be feasible in every circumstance, we can at least expect that such questions be routinely raised and point to examples of how those questions can be empirically addressed, so that the limitations as well as the benefits of testing practice are illuminated. Doing so could, I believe, go a long way toward promoting more judicious and educationally sound uses of tests.

I want to thank Lissitz and Samuelsen and the editors of *Educational Researcher* for encouraging dialogue on this important issue. Our understanding of validity in educational measurement—and our ability to promote sound testing practice consistent with ambitious educational goals—is only enhanced when scholars have the courage to raise bold proposals and offer them for public critique and dialogue.

NOTES

¹It is important not to conflate standardization with multiple-choice tests. Essentially, *standardization* refers to the aspects of an assessment that are common across individuals and contexts. Although standardized assessments include multiple-choice tests, they can also include complex performance assessments, such as multimedia portfolios or observation systems, where features like guidelines, criteria, and procedures for combining evidence are standardized.

²This section draws heavily on Moss, Girard, and Haniford, 2006.

³See American Psychological Association (APA), 1954; American Educational Research Association (AERA) & National Council on Measurements Used in Education (NCME), 1955; APA, 1966, 1974; AERA, APA, & NCME, 1985, 1999. Although not apparent from the name of the copyright holder, AERA, APA, and NCME jointly sponsored all five editions.

⁴Because Lissitz and Samuelsen associate the term *consequential validity* with Messick, it is important to note that Messick (1996) actually eschewed that term or any term that implied different types of validity. He referred instead to the *consequential aspect* or *basis of validity* and, indeed, offered a more circumscribed understanding of its role in validity judgments than Lissitz and Samuelsen describe. In Messick's (1989a) view,

It is not that adverse social consequences of test use render the use invalid but, rather, that adverse social consequences should not be attributable to any source of test invalidity such as construct irrelevant variance. If the adverse social consequences are empirically traceable to sources of test invalidity . . . then the validity of the test use is jeopardized. . . . If the social consequences cannot be so traced . . . then the validity of the test use is not overturned. (p. 11)

The 1999 Standards adopted Messick's position, although others, myself included, have argued for a broader role (see Moss, 1992, 1995, 1998, for an overview of the different positions). It is also important to note that, although Messick dismisses the term *content validity* for a similar reason, having to do with the implication of validity “types,” he considers the work that falls under that label as crucial (albeit insufficient) within a unitary conception of validity.

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.
- American Educational Research Association & National Council on Measurements Used in Education. (1955, January). *Technical recommendations for achievement tests*. Washington, DC: National Education Association.
- American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin* [Supplement], 51(2, part 2).
- American Psychological Association. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: Author.
- American Psychological Association. (1974). *Standards for educational and psychological tests*. Washington, DC: Author.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., 443–507). Washington, DC: American Council on Education.
- Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.
- Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory and public policy* (pp. 147–171). Urbana: University of Illinois Press.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.

- Educational Testing Service. (2002). *Standards for quality and fairness*. Princeton, NJ: Author.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–535.
- Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement*, 2(3), 135–170.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.
- Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36, 437–448.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3(Monograph Supplement 9), 635–694.
- Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5–11.
- Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.
- Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment*. Washington, DC: National Center for Education Statistics.
- Mislevy, R. J., Gee, J. P., & Moss, P. A. (in press). On qualitative and quantitative reasoning in validity. In K. Ericson & M. Wolff-Roth (Eds.), *Generalizing from educational research: Beyond the quantitative-qualitative opposition*. Mahwah, NJ: Lawrence Erlbaum.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–67.
- 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. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice*, 14(2), 5–13.
- Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, 17(2), 6–12.
- Moss, P. A. (Ed.). (2007). *Evidence and decision making: The 106th yearbook of the National Society for the Study of Education, Part I*. Malden, MA: Blackwell.
- Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education*, 30, 109–162.
- Moss, P. A., Pullin, D., Haertel, E. H., Gee, J. P., & Young, L. J. (Eds.). (in press). *Assessment, equity, and opportunity to learn*. New York: Cambridge University Press.
- National Research Council. (2002). *Scientific research in education* (R. J. Shavelson & L. Towne, Eds.). Washington, DC: National Academy Press.
- Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405–450.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5–8, 13, 24.
- Wilson, M. R. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

AUTHOR

PAMELA A. MOSS is a professor at the University of Michigan, School of Education, 610 East University Avenue, Ann Arbor, MI 48109-1259; pamos@umich.edu. Her areas of specialization are at the intersections of educational assessment, validity theory, and interpretive social science.

Manuscript received September 21, 2007

Accepted September 21, 2007