Comments on Lissitz and Samuelsen

On Validity Theory and Test Validation

by Stephen G. Sireci

Lissitz and Samuelsen (2007) propose a new framework for conceptualizing test validity that separates analysis of test properties from analysis of the construct measured. In response, the author of this article reviews fundamental characteristics of test validity, drawing largely from seminal writings as well as from the accepted standards. He argues that a serious validation endeavor requires integration of construct theory, subjective analysis of test content, and empirical analysis of item and test score data. He argues that the proposals presented by Lissitz and Samuelsen require revision or clarification to be useful to practitioners for justifying the use of a test for a particular purpose. He discusses the strengths and limitations of their proposal, as well as major tenets from other validity perspectives.

Keywords: construct; content validity; criterion-related validity; testing standards; validity

It has been almost 20 years since the publication of Messick's seminal chapter on validity (Messick, 1989b) and almost 10 years since the publication of the most recent edition of *Standards for Educational and Psychological Testing* (hereinafter, "the Standards"; American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). Thus the time is right to reconsider how we conceptualize what has often been called the most important concept in psychometrics—validity. I am grateful to Robert W. Lissitz and Karen Samuelsen for reminding us of the importance of this topic and how it continues to evolve ("A Suggested Change in Terminology and Emphasis Regarding Validity and Education," this issue of *Educational Researcher,* pp. 437–448). I am also grateful to the editors and editorial board of *Educational Researcher,* not just for the invitation to comment but also for devoting an entire issue to this topic. Given the increased use of tests in education and their consequences for individuals and for society, all educational researchers will benefit from the effort to promote a firm understanding of validity theory and test validation.

I have been a student of validity theory for quite some time and have had the pleasurable job of documenting validity evidence for several testing programs. I describe validation as pleasurable because it is the ultimate challenge for a psychometrician, and there is great satisfaction in overcoming that challenge. To prove our worth in the world, we as researchers must present evidence that the information provided by the tests that we study and help create is useful and scientifically sound. However, as Messick (1989b), Kane (1992, 2006), and many others have observed, it is essentially impossible not only to prove that a test is valid but to prove even that we are measuring what we think we are measuring. Therefore, our job in supporting the use of a test for a particular purpose involves presenting sufficient evidence to defend such use. Any conceptualization of validity theory must acknowledge that what is to be validated is not a test itself but the use of the test for a particular purpose.

Lissitz and Samuelsen summarize several seminal writings in the area of validity theory and test validation. Cronbach and Meehl's (1955) article is a particularly influential work. Both Cronbach and Meehl served on the APA-AERA-NCME Joint Committee that produced the first version of the Standards (APA, 1954), and the purpose of their article was to explain their concept of construct validity. I will not delve much further into the history of validity, but I will add that, in addition to Cronbach, Meehl, Messick, and Kane, my thoughts in this area have also been greatly influenced by Ebel, Guion, Shepard, and many others, as well as by a careful reading of all five versions of the Standards (see Sireci, 1998a, for an additional historical perspective on validity).

My reading of the validity literature over the years has led me to the following conclusions regarding fundamental aspects of validity:

- Validity is not a property of a test. Rather, it refers to the use of a test for a particular purpose.
- To evaluate the utility and appropriateness of a test for a particular purpose requires multiple sources of evidence.
- If the use of a test is to be defensible for a particular purpose, sufficient evidence must be put forward to defend the use of the test for that purpose.
- Evaluating test validity is not a static, one-time event; it is a continuous process.

These fundamental aspects of validity do not come from a single source. Rather, they are common themes that run through the seminal writings in this area (e.g., Cronbach, 1971; Kane, 1992; Messick, 1989a, 1989b; Shepard, 1993), including our authoritative Standards (AERA et al., 1999). Although Lissitz and Samuelsen propose some interesting ideas that identify weaknesses in current

Educational Researcher, Vol. 36, No. 8, pp. 477–481
DOI: 10.3102/0013189X07311609

validity theory, some of their proposals appear to be at odds with what I consider to be fundamental aspects of validity theory and test validation. In my response to their article, I will first comment on the important points they raise with which I agree. Then, drawing from what I consider to be fundamental aspects of validity, I will discuss the points with which I disagree and the areas that are in need of clarification. Finally, I will provide suggestions for how we might best characterize validity theory and go about the process of test validation. In presenting these suggestions I borrow ideas from Lissitz and Samuelsen and many others.

Must a Unitary Conceptualization of Validity Be Centered on Construct Validity?

Lissitz and Samuelsen argue against the notion promulgated by Loevinger (1957), Messick (1975, 1980, 1989b), Fitzpatrick (1983), and others that all validity is construct validity. I am sympathetic to this view and admire them for speaking out against this unitary conceptualization that has dominated the validity theory literature for decades. The unitary conceptualization of validity as construct validity is theoretically sound and widely endorsed, but like all theories, it has its imperfections. Paramount among these is that it is extremely difficult to describe to lay audiences. In education, tests are commonly used for accountability purposes, and psychometricians are called upon to provide advice to policy makers. Complex philosophical notions of appropriate testing qualities and testing practices are not optimal in such situations. It is hard to describe an underlying latent variable that was "constructed" by educators and that the test is designed to measure. It is far easier to talk about the content domain measured, particularly when it is operationally defined using test specifications. Thus the concept of content validity is much more palatable and understandable than the concept of construct validity to policy makers and the general public. In fact, in educational testing there is often a thin distinction between a construct and a content domain. As Shepard (1993) wrote, "Content domains for constructs are specified logically by referring to theoretical understandings, by deciding on curricular goals in subject matter fields, or by means of a job analysis" (p. 413). Shepard and others have also pointed out that the unitary conceptualization of validity has done little to provide guidance regarding how to validate the use of tests in specific situations. Thus I applaud efforts to bridge the gap between philosophical and pragmatic aspects of test validation.

Another, related weakness of the unitary conceptualization of validity centered on construct validity is that validity theory can be unified without the modifier *construct*. Ebel (1961) suggested using the term *meaningfulness* instead of validity, but that idea never caught on. Later, Kane (1992), drawing from ideas put forth by Cronbach (1971), proposed an *approach* to validation that sidestepped most validity nomenclature and offered practical advice for gathering and analyzing evidence to support the use of a test for a particular purpose. This argument-based approach has been endorsed, at least implicitly, by the Standards (AERA et al., 1999) and by others and provides an example of how validity can be characterized in a general way, without referring to a construct.

Can the term *validity* be used in a general sense without invoking the notion of a construct? I think so, and I think the current version of the Standards accomplished this generality when it defined validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). With respect to validation, the Standards explicitly refer to the concept of a validity argument when they described validation as a process of

> accumulating evidence to provide a sound scientific basis for the proposed score interpretations. . . . Validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use. (AERA et al., 1999, p. 9)

Lissitz and Samuelsen argue against the unitary conceptualization of validity based on construct validity, and they propose new nomenclature to characterize validity and validation. However, their discussion of the literature gives short shrift to the organizing framework put forth in the most current version of the Standards, which was based on the input of just about every major testing organization in the United States and was jointly developed and endorsed by the three major professional associations involved in the testing enterprise. Therefore, before considering Lissitz and Samuelsen's proposals, we should remind ourselves that the Standards specify five "sources of validity evidence . . . that might be used in evaluating a proposed interpretation of test scores for particular purposes" (p. 11). These sources are evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing. I find this framework helpful for gathering, analyzing, and documenting evidence to support the use of a test for a particular purpose, and it is against this framework that I consider the proposals offered by Lissitz and Samuelsen.

Critiquing Lissitz and Samuelsen's Validity Proposals

Lissitz and Samuelsen provide a "systematic structural view of the technical evaluation of tests" (p. 437). In this view, test evaluation is separated into internal and external aspects. Lissitz and Samuelsen's Figure 1 implies that validity is involved only with the internal aspect, but their description of their approach suggests a sequential process in which the validator first looks at the internal aspects of a test and then moves to the external aspects, if the situation calls for it. My understanding of the model is that the authors want us first to evaluate aspects of a test independent of its application and any theory and to separate validation of the construct from validation of the test itself. As they put it,

> An inquiry into the validity of a test should first concern itself with the characteristics of the test that can be studied in relative isolation from other tests, from nomothetic theory, and from the intent or purpose of the testing. (p. 437)

My agreement with the most recent writings on the topic of validity, including the Standards (AERA et al., 1999), leads me to believe that validity can be evaluated only with respect to a specific testing purpose, and so I cannot agree with Lissitz and Samuelsen that we should approach validity independent of the testing context and purpose of the testing. I am a big believer in validity evidence based on test content, but I cannot imagine convening a committee of subject matter experts to evaluate and rate test items without informing them of the testing purpose. Even if such studies were conducted, how would the results be evaluated if not with respect to test specifications designed for a

particular purpose? Furthermore, if a test were found to have good content coverage for a specific purpose (e.g., sixth-grade math in Massachusetts), would it still be content valid for another purpose (e.g., adult mathematics literacy)? Therefore, although I agree that much validity evidence can be gathered by evaluating test content, I do not agree that validation can proceed without being referenced to the purposes for which test scores are being used.

With respect to the conceptualization of Lissitz and Samuelsen's internal aspect of validity, I believe further explanation is needed. How are latent processes internal to the test, and how can reliability be considered part of content validity? Lissitz and Samuelsen's description of the internal aspect of validity makes it sound as if reliability and validity were properties of a test, but they are not. Reliability and validity refer to interpretations of test scores, not the test itself. Estimates of test score reliability can be of several types designed to answer different questions (e.g., How consistent are the scores over time, forms, item types, etc.?), and many different estimates may be found based on different types of examinees and estimation designs. The authors even seem to argue that cognitive processes are part of content validity, which is also unclear to me.

Further explanation is also needed regarding Lissitz and Samuelsen's external aspect of validity. I can see how theory, criteria, and impact are external to a test, but I think an integrated evaluation of these factors is needed to evaluate fully the utility of a test for a particular purpose. I am not suggesting that all sources of evidence need to be investigated to evaluate test use and interpretation. The testing purpose and types of inferences derived from test scores should determine the most relevant evidence to gather. What I am suggesting is to keep all potential sources of evidence on the table for full, simultaneous consideration. In this view, I agree with Messick's (1989b) definition of validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). I believe integration of theory and evidence is needed in a validation endeavor, and I do not see the utility of separating test evaluation further into the different aspects proposed by Lissitz and Samuelsen. If aspects of validation are separated, as sometimes they must be to explain the contribution of different types of evidence, I much prefer the traditional notions of content, criterion-related, and construct validity or the five sources of validity evidence currently described in the Standards.

With respect to Lissitz and Samuelsen's taxonomy of test evaluation procedures, I see the logic in distinguishing theoretical and practical perspectives, as well as internal and external investigative focuses. However, I believe that both perspectives and focuses are needed in a validation endeavor, which is consistent with the idea that validation involves integrating empirical evidence and theory relevant to test development and use. As the Standards describe,

> A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses. . . . Ultimately, the validity of an intended interpretation . . . relies on all the available evidence relevant to the technical quality of a testing system. This includes evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all examinees. (AERA et al., 1999, p. 17)

Although the guidance provided in the Standards is not perfect, I believe that this integrated characterization is more beneficial to those who seek to evaluate tests than is the framework being proposed by Lissitz and Samuelsen.

Before discussing some of the important points raised by Lissitz and Samuelsen, I want to mention a few other aspects of their article that I believe are in need of revision or clarification. First, I noted that the stated purposes of their article were to "provide a list of approaches to determining content validity . . . [and] provide a more communicative vocabulary for types of validation" (p. 437). I think the authors need to do more to accomplish their first purpose. There are numerous approaches to content validation, and the authors do not discuss any of them (see Crocker, Miller, & Franks, 1989, or Sireci, 1998b, for descriptions of some of these methods). One approach that has emerged since the Standards were published is based on alignment methodology (see Bhola, Impara, & Buckendahl, 2003; Webb, Alt, Ely, Cormier, & Vesperman, 2005). I believe that this approach needs to be mentioned in any discussion of validating educational tests used for accountability purposes. Lissitz and Samuelsen provide samples of questions asked regarding content (see their Table 1), but the questions fall short of a list of content validation approaches, and some of the questions listed under "latent process" could also be grouped under "content."

With respect to Lissitz and Samuelsen's second purpose, I do not yet see the utility in the vocabulary they propose. In fact, the term *internal validity* is already used and widely known in the field of research design (i.e., Campbell & Stanley, 1963), and ascribing another meaning to that term is likely to foster confusion rather than clarity.

As I understand them, the new vocabulary and framework proposed by Lissitz and Samuelsen are anchored in the belief that a test can be evaluated independent of a notion of the construct measured. This is a purely operationalist perspective, which has merit but also limitations. For example, in their section "Other Conceptions of Validity," they discuss the issue of defining an algebra test. They argue that "the test and its adequacy do not depend on the relationships to other tests defining other domains" and "the development of a nomological network is not really an issue, even for future study" (p. 440). Once the testing purpose and context are considered, as they must be in any serious validation effort, it is hard to agree with these statements. If the algebra test correlates highly with English-language proficiency, for example, the potential for bias against English-language learners is present and should be evaluated. Yes, such an evaluation of potential bias would start with an analysis of test content, but bias could be missed without the statistical finding to urge us to search more diligently. Thus, although I agree that validity evidence based on test content is paramount for educational tests, it is hard for me to conceptualize an atheoretic paradigm of test validation.

One last point with which I take issue is the importance of evaluating testing consequences. Lissitz and Samuelsen describe this as "sometimes an important consideration and sometimes worth studying" (p. 444). I think evaluation of testing consequences is a necessary part of any serious validation effort. First, it should be noted that Messick (1989b) did not use the term *consequential validity* but rather talked about social considerations in testing, including the consequential basis of test interpretation and test use. More important, however, if we are to evaluate the validity of inferences derived from test scores, the implications and effects of those inferences (i.e., the consequences) must also be studied. Unintended negative consequences might not point to problems in a test per se but may indicate problems within a testing system. For example, a computer-based test may have sufficient evidence of validity for its intended purpose, but if educational programs are turning away students because they do not have enough computers, that consequence has important implications for the future of the testing program.

Summary of Points of Agreement and Disagreement

I have been critical of many of the arguments raised by Lissitz and Samuelsen, and so it is important that I return to the many issues on which we agree. First of all, I agree that the unitary conceptualization of validity has undermined the importance of content validity in evaluating the utility and appropriateness of tests used in educational contexts (Sireci, 1998a). Second, I acknowledge the difficulty that this theory has presented for practitioners whose job it is to document the validity of a test for a particular purpose. Third, I agree that the notion of validity has evolved over the years and that the notion of construct validity has been particularly confusing to many. Lissitz and Samuelsen have raised several important points about the problems in holding a view of test validation that is focused primarily on theoretical or criterion-related analysis at the expense of actually looking at the test items.

Where I depart from Lissitz and Samuelsen has to do with the importance of a theory of the construct measured in test validation. Specifically, I think theory is needed to design validation studies and to rule out the plausible rival hypotheses that may explain test performance. As Campbell and Fiske (1959) demonstrated, analysis of discriminant validity helps to illuminate test score interpretations above and beyond the information provided solely through convergent evidence (see also Zumbo, 2007). I also prefer Campbell and Fiske's characterization of reliability and validity as ends of a continuum (i.e., reliability as monomethod-monotrait validity) rather than the view that reliability is an internal aspect of validity. I also depart from Lissitz and Samuelsen in that I do not see validity as an inherent property of a test, as their proposal seems to imply. Finally, it seems that much of their theory is relevant mainly to educational tests. In other testing contexts, such as personality assessment and projective testing, certainly construct theory must take a more prominent role.

The Continuing Evolution of Validity

The unitary conceptualization of validity is philosophically sound and can be applied to virtually any testing program. However, this philosophical victory has come at the expense of ease of communication and a denigration of traditional vocabulary that more easily communicated important characteristics of test quality. The current version of the Standards (AERA et al., 1999) does a good job in describing validity in both philosophical and pragmatic terms and in categorizing different types of validation evidence into sources of validity evidence. However, like validation itself, validity theory is not static. As the use, importance, and consequences of educational and psychological tests increase, so do the questions regarding test quality, fairness, and utility. Troublesome validity issues that are not adequately addressed in the current version of the Standards include evaluating the comparability of different versions of a standardized assessment (e.g., test accommodations for individuals with disabilities, translated versions of tests), the role of test-curriculum framework alignment in educational accountability systems, the different roles of stakeholder groups in the evaluation of testing consequences, and how specific statistical procedures (e.g., differential item functioning) and research designs (e.g., comparison of instructed and noninstructed groups of students) fit into the five sources of validity evidence.

In my opinion, the current state of validity theory and standards for test validation is far from perfect, but it is pretty darn good. The argument-based approach to test validation requires us to prioritize validity questions and gather evidence to defend the way test scores are currently used. It acknowledges that validity can never be unequivocally established but also that we need to put forth enough evidence to make a convincing argument that the interpretations made on the basis of test scores are useful and appropriate. That makes a lot of sense to me and is the reason why this approach is stressed in the Standards as well as in the courts (Sireci & Parker, 2006).

Problems of validity nomenclature have pervaded our field since the inception of validity and will continue to affect our thinking and practices. The traditional sources of validity evidence—content, criterion related, and construct related—seem useful to me, and I still use these terms. However, Messick and others correctly pointed out that all forms of evidence could be subsumed under the rubric of construct validity. Messick argued that all forms should be considered aspects of construct validity because *validity* referred to interpretations of test scores, not the test itself. That argument is philosophically sound, but it does not stop me from talking about content and criterion-related validity when I discuss educational testing with students and lay audiences. I applaud efforts to refine our validity vocabulary, but I am not in favor of expanding it unless some type of parsimony is reached. Lissitz and Samuelsen attempt such parsimony and clarification, and perhaps others will be persuaded by their arguments. At this juncture, however, I find an argument-based approach to validation and the sources of validity evidence promulgated by the Standards to be more useful and compelling for evaluating educational tests.

The important lesson I take from Lissitz and Samuelsen is that we must consider test content as a necessary and integral source of information whenever we evaluate the use of a test for a particular purpose. As they describe, it is only through analysis of test content that we can adequately explain empirical relationships among test scores and other variables. Lissitz and Samuelsen imply that a focus on construct validity has impeded investigations of test content, and I agree (Sireci, 1998a). In my opinion, Messick was disturbed by testing programs that presented a validity argument based solely on subjective interpretations of test content and so he demoted content validity to a much lower status than construct validity. Lissitz and

Samuelsen remind us that content validation is a necessary step in validating the utility of an educational test. I agree with that, although I would argue that validity evidence based on test content is necessary but not sufficient. A serious effort to validate use of an educational test should involve both subjective analysis of test content and empirical analysis of test score and item response data.

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.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin* [Supplement], *51*(2, part 2).

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*(3), 21–29.

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

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Crocker, L. M., Miller, D., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education, 2,* 179–194.

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

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

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist, 16,* 640–647.

Fitzpatrick, A. R. (1983). The meaning of content validity. *Applied Psychological Measurement, 7,* 3–13.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527–535.

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. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30,* 955–966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012–1027.

Messick, S. (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. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education/Macmillan.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405–450.

Sireci, S. G. (1998a). The construct of content validity. *Social Indicators Research, 45,* 83–117.

Sireci, S. G. (1998b). Gathering and analyzing content validity data. *Educational Assessment, 5,* 299–321.

Sireci, S. G., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement: Issues and Practice, 25*(3), 27–34.

Webb, N. L., Alt, M., Ely, R., Cormier, M., & Vesperman, B. (2005). *The WEB alignment tool: Development, refinement, and dissemination.* Washington, DC: Council of Chief State School Officers.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 45–79). Amsterdam, Netherlands: Elsevier Science.

AUTHOR

STEPHEN G. SIRECI is a professor of education and director of the Center for Educational Assessment at the University of Massachusetts, Amherst, School of Education, 156 Hills South, Amherst, MA 01003–4140; *sireci@acad.umass.edu.* His research focuses on test development, test evaluation, validity theory and applications, fairness issues in testing, computer-based testing, cross-lingual assessment, and assessing special populations such as students with disabilities and linguistic minorities.

Manuscript received September 4, 2007
Accepted September 17, 2007