



Response to Comments

## Further Clarification Regarding Validity and Education

by Robert W. Lissitz and Karen Samuelsen

We would like to begin by thanking the editors of *Educational Researcher's* Features section for the opportunity to present our ideas on validity, a topic that we consider to be of utmost importance ("A Suggested Change in Terminology and Emphasis Regarding Validity and Education," this issue, pp. 437–448). We would also like to thank the distinguished discussants—Susan E. Embretson, Joanna S. Gorin, Robert J. Mislevy, Pamela A. Moss, and Stephen G. Sireci—for their thoughtful and challenging comments (also in this issue of the journal) regarding our article. We wish we had more time and space to respond to all of the interesting remarks and excellent suggestions for improvement in our understanding of validity. Because the editors have restricted us to 2,500 words, our response will be limited to a brief summarization of our position emphasizing three elements that some of the discussants interpreted somewhat differently from our intention or about which we have a truly different opinion. Those elements are (a) the unitary theory, (b) validity terminology, and (c) the construct-versus-content debate. We will spend little time addressing specific points raised by the discussants, although the temptation is great.

### The Unitary Theory

As we indicated in our article, we are attempting to move away from a unitary theory focused on construct validity and to reorient educators to the importance of content validity and the general problem of test development. The motivation for our article appears, by coincidence, quite explicitly in the comment articles by Gorin and Sireci. As stated by Gorin, "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" (p. 456). Sireci adds the following about the unitary theory: "Paramount among these [imperfections] is that [the unitary conceptualization of validity as construct validity] is extremely difficult to describe to lay audiences" (p. 478). Conversations with students and, more important, with sophisticated professionals working on state and other tests and related activities, have compelled us to draw the same conclusion. As Einstein reportedly once said, "You do not really understand something unless you can explain it to your grandmother." Construct validity, as it is currently articulated, seems to flunk the grandmother test.

We hope it is clear that we are not proposing to substitute a new unified theory of validity for the one promulgated by Messick and others who have elevated construct validity to such a lofty position. Our figures and tables are intended to show that there are quite different elements in any consideration of what has been termed *validity*, along with a wide variety of techniques that may be applied to the various concerns and considerations that arise in attempting to deal with validity. We have also tried to show that the concept of reliability is not a completely different concept, as often presented, but shares many characteristics with validity, particularly in the area that we term *utility*, traditionally called *criterion validity*. As Lindquist (1951) indicated (quoted in our article, p. 439), when a test is constructed to predict a specific criterion, the test typically is constructed to be as much like the criterion as possible. This practice is similar to, if not the same as, that of developing equivalent forms and the assessment of them with reliability measures. (Embretson has expanded on our figures, particularly by adding further considerations to the development phase. Her way of depicting this is worth further study, although we are not clear that it is an easier depiction than would result if we appended her additional considerations to those covered in our tables.)

Sireci has a marvelous sense of perspective based on intimate knowledge of the standards, their continued development, and his own valuable research in the field. However, we believe he is a bit too positive about the unitary theory when he describes it as "theoretically sound" and "philosophically sound," and we are not sure what such claims mean. As Sireci effectively states, "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" (p. 480). Messick's view is, at best, dated, given that the work by Mislevy and others in the field today is quite sophisticated, carrying forward the process of establishing content validity (internal and theoretically based). Unfortunately, we believe that describing someone's work as achieving "construct validity" is quite uninformative if, like Messick and many others, one considers the term to be a truly unifying concept that includes everything related to these matters. It would be somewhat like describing a person you know to someone who does not know the person by saying that he or she is human.

Educational Researcher, Vol. 36, No. 8, pp. 482–484  
DOI: 10.3102/0013189X07311612  
© 2007 AERA. <http://er.aera.net>

## Validity Terminology

Too much focus on terminology and too little focus on concrete advice in the work by Messick and others is perhaps the reason for the current state of affairs. Several of the discussants confirm this observation, sometimes explicitly and sometimes implicitly. Part of the problem with construct validity is that it is actually a compound concept, including both a construct-centered activity (what we call an internal focus) and a nomological activity (what we call an external focus). This problem was very well articulated by Embretson in 1983 and in her response to our article in this issue. These very helpful explications and that by Moss in her response to our article are useful in explaining what Cronbach and Meehl (1955) meant when they discussed the concept of construct validity in their groundbreaking article. Our position—like the position of Cronbach and Meehl that is found in the reference to Embretson (1983) in Gorin's article (p. 457)—is that there are two separate aspects of construct validity and both are necessary to meet the definition of the concept of construct validity. Thus focusing on a construct does not mean engaging in classical or modern construct validation. In other words, we tried to explain that this form of encompassing (unifying) validity is not the same as what one means by terms such as *construct centered* or *construct definition*. We believe that there is much overlap between our article, the 1983 article by Embretson, and her current comment on our article (and we are gratified to see that Embretson agrees). Embretson, in her excellent review of our work, indicates that in the development phase of a test, doing what we call *external examination activity* is not nearly as important as the internal study that is critical to that phase. Naturally, we agree with that contention. She also indicates, in her response to our article, that studying external factors can lead to insights about what one is doing with regard to test development and hence should not be ignored completely. Perhaps our article communicated too sharp a distinction between what one learns in the external and the internal exercises regarding a test. We thank Embretson for emphasizing that point and noting the iterative nature of test development and application for many tests.

As we indicated, Mislevy's work is, in the terminology of our system, an excellent example of what people should be doing in test development and test analysis (internal, theoretical, and practical content validity). Our two tables are an attempt to summarize some of the work that has been applied to these various forms of validity. Moss's suggestion that exemplars be developed to illustrate best practice is an idea worth considering and one we would support. We believe that it will be hard to do, especially if one wants exemplars of construct validity that are not essentially illustrations of content, or criterion, validity. We disagree that focusing on techniques to accomplish a task leads to "mechanistic" thinking, as Moss indicates in her quotation from the National Research Council (p. 473). It is true that techniques popular for a time are often improved or replaced by new techniques, but we think that assessment professionals can accept the challenge to keep current in exchange for clear presentations of the options and explicit statements of the value of each technique for test validation. It is hard not to provide lists of such procedures if one is concerned with encouraging best practice in assessment.

Ironically, in trying to create a new, simpler validity vocabulary, we used terminology that was confusing to one discussant

and possibly to others as well. Sireci raises a good point about our selection of the term *internal* to describe a form of validity. We did not think about the notion of threats to internal validity in the research design area as a source of confusion. A better term could and probably should be found. However, we remain convinced of the value of the conceptualization we outlined in our article contrasting internal and external focuses, as well as our use of highly conceptual (dare we say, theoretical?) approaches rather than pragmatic and practical ones. We hope that such a system remains helpful in furthering the understanding of validity. We also believe that our two tables serve as a worthwhile introduction to some of the important questions and to the many approaches that are available for studying validity. We understand the confusion that we caused Gorin by discussing Campbell and Fiske's (1959) approach as an example of external work and discussing convergent and divergent analysis of different parts of the same test as an example of internal analyses that are useful in understanding internal test behavior.

## The Content-Versus-Construct Debate

We tried to explain that focusing on construct definition as part of the test development phase is important. For us, working toward content validity is a very important activity (and, for many purposes of testing, it is the most important activity). The test development phase has always been considered critical to the successful completion of a test construction or assessment activity. We greatly admire Mislevy's work concerning the structure of evidence, the creation of arguments, and "validity by design." Because of his emphasis on the structural analysis of test construction, we see his work as essentially relating to content validity. It is an excellent example of what we wish to see more of in testing. Please note that we like constructs and believe that thinking about meaning and definition is a good thing to do as one develops a test. However, we must reiterate that believing that you are working with constructs is not the same thing as engaging in construct validation. We further stress that it is not the subsequent correlation of results from a test-taking enterprise with some other test data that justifies the creation process, except where one is explicitly concerned with utility. An illustration of this confusion is the suggestion by Gorin that if a test constructed to measure mathematics and a test constructed to measure English are highly correlated, perhaps they are measuring the same thing. The resolution of her concern will be greatly supported by an examination of the domains of both tests and of the test blueprints for each, regardless of the general effect of IQ on test performance.

Sireci states that "there is often a thin distinction between a construct and a content domain" (p. 478). We agree, although we think the distinction may be even thinner than thin. We disagree, however, with Gorin's comment that, "historically, the use of operational definitions as indicators of score meaning has been tried and discarded" (p. 457). Looking at the content of a test and its psychometric model is not a discarded approach in efforts to understand the domain (or as Sireci might agree to say, its construct). We do not see the analysis of the test content and properties as ever being separable from the understanding of the construct (or the domain), and we believe that characterizing the operationalization of a construct as in demise is premature.

Sireci also notes that "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” (p. 477). The same point is made by Moss in her quote from Cronbach (1971): “A single instrument is used in many different ways” (p. 473). Our article is an attempt to show that this is not generally true, although in the area of testing for utility, relevant evidence must be marshaled regarding the relation of the test to the criterion (i.e., the purposes or uses of the instrument). To return briefly to the example of the ruler, it seems that Sireci is suggesting that one cannot define length as measured by a ruler unless one knows the purpose for measuring the length of the object in question. We remain convinced otherwise, whether the measurement occurs in physical science or social science, including education research. Finally, we ask whether one can consider the validity of a test that has never been administered (in other words, for which there are no test data), and our answer is *yes*. We certainly grant that additional insights may be obtained when test data become available, but the test construction process that leads to an operational test form holds a wealth of information prior to administration.

### Conclusion

Sireci further notes that “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” (p. 481). We agree that both sorts of internal evidence should be examined. One impetus to our proposing a new vocabulary was to make the concept of validity more accessible, understandable, usable, and supportable. We strongly believe that change is needed and that the quality and content of communication to the states and others involved in assessment need to be improved. States are much involved in the process of validating their tests for purposes of No Child Left Behind. Testing companies are providing services to the states and to testing professionals, who need to be able to explain to the lay public and to each other what they are doing and why they are doing it. We very much hope that our article and the excellent comments by all of the authors who contributed to the Features section in this issue of *Educational Researcher* will serve to further the quality and usefulness of discussion in the challenging area of validity.

It is clear that our effort to change the terminology, or at least the definitions, regarding validity will be a great challenge and that we will have plenty of help from members of the profession. We hope that the excellent, relatively new work by psychometricians on the development phase of assessments will bring appropriate prestige to this effort. We also hope that any new concepts and any improvements in the understanding of traditional concepts will offer useful guidance for practitioners and satisfy the intellectual standards of theoreticians, rather than focus solely on

the latter. We are optimistic about the commitment of the profession to meet the challenges of this very important area.

### REFERENCES

- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- 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., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Embretson, S. E. (Whitely). (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93(1), 179–197.
- Embretson S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher* 36, 449–455.
- Gorin, Joanna S. (2007). Reconsidering issues in validity theory. *Educational Researcher* 36, 456–462.
- Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 119–158). Washington, DC: American Council on Education.
- Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher* 36, 437–448.
- Mislevy, Robert J. (2007). Validity by design. *Educational Researcher* 36, 463–469.
- Moss, Pamela A. (2007). Reconstructing validity. *Educational Researcher* 36, 470–476.
- Sireci, Stephen G. (2007). On validity theory and test validation. *Educational Researcher* 36, 477–481.

### AUTHORS

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

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

Manuscript received September 27, 2007

Accepted October 4, 2007