JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

Journal of Educational Measurement
Spring 2013, Vol. 50, No. 1, pp. 99–104

Agreeing on Validity Arguments

Stephen G. Sireci
University of Massachusetts Amherst

Kane (this issue) presents a comprehensive review of validity theory and reminds us that the focus of validation is on test score interpretations and use. In reacting to his article, I support the argument-based approach to validity and all of the major points regarding validation made by Dr. Kane. In addition, I call for a simpler, three-step method for developing validity arguments, one that focuses on explicit testing purposes, as suggested by the Standards for Educational and Psychological Testing. *If testing purposes are appropriately articulated, the process of developing an interpretive argument becomes unnecessary and validation can directly address intended interpretations and uses.*

I am honored to have the opportunity to comment on Michael Kane's newest contribution to the validity literature—"Validating the Interpretations and Uses of Test Scores" (Kane, this issue). I have long been a fan of Dr. Kane's writing on validity (e.g., Kane, 1992, 2006) and consider him to be one of the greatest validity theorists of our time. The "argument-based approach" to validity he articulated in Kane (1992) was essentially incorporated into the most recent version of the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999). As the *Standards* described, "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" (p. 17).

The *Standards*' characterization of a comprehensive validation endeavor as a validity argument is significant because the *Standards* represent a "consensus" in the validity literature. As Kane pointed out, the validity literature is over 100 years old and is rich with important ideas and some contentious debate. The *Standards* do not represent the ideas of a single validity theorist, but rather the consensus understanding of three organizations that have promoted guidelines on fair and appropriate testing practices for over 50 years. The definition of validity provided by the *Standards* is consistent with Kane (this issue): "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA et al., 1999, p. 9).

From this definition, and from the previous description of a validity argument, it can be seen that validity is *not* a property of a test, but rather it refers to the use of a test for a particular purpose. It refers to the degree to which evidence exists to support explicit uses. In his earlier writings on validation, Kane (1992, 2006) steered the measurement community away from the various "types" of validity and toward a systematic "approach" to validity—an approach that is focused on determining whether a sufficient body of evidence exists to justify the use of a test for a particular purpose. This body of evidence is called a validity argument.

The argument-based approach to validity articulated by Kane (1992, 2006) involves first developing an interpretive argument and then developing a validity argument. As reiterated by Kane (this issue) the interpretive argument specifies the intended interpretations and uses of test scores, and the validity argument is a comprehensive analysis of the evidence gathered to evaluate the interpretive argument. What separates Kane's current writing from his earlier works is that Kane (this issue) distinguishes between validity arguments for test interpretation and those for test use. Essentially, the interpretive argument is extended to an "interpretation/use argument" (p. 8). He also provides a concise summary of some fundamental tenets of validity theory (i.e., his eight "general points") with which I wholeheartedly agree. However, rather than spend my remaining allotted space praising Dr. Kane's most recent article, I would like to spend some time describing how I think we can simplify an argument-based approach to validity. It is not my intent to lose any of the important ideas put forth by Dr. Kane or by the *Standards*. Rather, I hope to retain these ideas in a way that is more straightforward for practitioners.

Articulating Testing Purposes

I believe an argument-based approach to validity can be simplified, and hence made more accessible to practitioners, by eliminating the need to develop an interpretation/use argument. If test developers and testing agencies properly articulate the intended purposes of a test, there should be no need for an interpretive argument. Clearly articulated statements of testing purpose should provide the necessary focus for validation. Again, borrowing from the *Standards*, "Validation logically begins with an explicit statement of the proposed interpretations of test scores, along with a rationale for the relevance of the interpretation to the proposed use" (AERA et al., 1999, p. 9). I agree that an explicit statement of testing purposes is the logical beginning of validation, but I go one step further—it is the logical beginning of *developing* a test. That is, tests are developed to fulfill one or more intended purposes. It is incumbent upon us as psychometricians to help those who commission these tests to articulate the intended purposes. Once these purposes are articulated, we know what we need to validate. We also know what it is we need to measure!

Note that I have used one word—"purposes"—to refer to the two terms that Kane distinguishes in his article—test interpretation and test use. To me, a clear statement of explicit testing purposes defines both the intended interpretations and uses.

I recognize the logic and elegance in an interpretive argument, and I have seen at least two excellent examples described by Forte (2012). However, interpretive arguments can be complex to develop and sometimes overwhelming, which may prevent validity practitioners from following the argument-based approach. In developing the interpretation/use argument, Kane (this issue) suggested specifying the "network of inferences and assumptions leading from the test performances to the conclusions to be drawn and to any decisions based on these conclusions" (p. 8), and that this network of inferences address *scoring inferences*, *generalization inferences*, and *extrapolation inferences*. It is at this juncture I think we may lose validity practitioners. Although I agree with everything Kane says about these inferences, I realize it represents (somewhat) new vocabulary and a new validation paradigm. I think we already

have a validation paradigm described in the AERA et al. (1999) *Standards*. I want to avoid new terms and strategies and encourage practitioners to follow the argument-based approach to validation articulated in the *Standards*, which I describe next.

A Validity Argument Based on Testing Purposes and Sources of Validity Evidence

The argument-based approach I advocate here involves three steps—(a) clear articulation of testing purposes, (b) considerations of potential test misuse, and (c) crossing test purposes and potential misuses with the five sources of validity evidence listed in the AERA et al. (1999) *Standards*. These three steps can be used to develop a *validation plan* that is used to build the validity argument. Given that validation must focus on explicit testing purposes, the first step should come as no surprise. The second step is part of responsible test use and must be considered in judging "whether the intended consequences are likely to be achieved and the potential for negative consequences is indeed small" (Kane, this issue, p. 56). The third step is where the fun for the validator begins—deciding which sources of validity evidence, and which specific studies, will best help to "accumulate a preponderance of evidence for or against the proposed interpretation or use" (Messick, 1989, p. 50).

Articulating Testing Purposes

Ideally, articulating the testing purposes should not be the role of the validator because it is the first step in test development (Downing, 2006, p. 6). How can we develop a good test if its purpose isn't clearly articulated? In reality, however, testing goals are not always clearly articulated by testing agencies, and in such cases the task falls to those conducting the validation. This step is increasingly important because many contemporary testing programs strive to accomplish multiple purposes. Some, such as the Smarter Balanced Assessment Consortium and the Partnership for the Assessment of Readiness for College and Careers, involve comprehensive "theories of action" that stipulate many goals—some specifically related to the assessments and some that speak to broader educational goals. A validation plan associated with such a broad testing *program* will need to incorporate the theory of action into the validation (Bennett, 2010), which will involve deriving explicit testing purposes from the theory.

Considering Potential Test Misuse

In the first step we addressed what the testing agency intends to do. The second step is consideration of potential test misuse. In this step, we confront how test scores may be misinterpreted or how the testing program may lead to unintended results. This may not always be a separate step, because many statements of testing purpose include cautions against misuse. For example, the purpose statement for the Tests for General Educational Development (GED Tests) states:

> The GED Tests have been designed to provide an opportunity for adults who did not complete a formal high school program to certify their attainment of high school–level academic knowledge and skills and earn their jurisdictions' high school–level

> equivalency credential, diploma, or certificate. Thus, the intended use of the GED credential is similar to that of a high school diploma—to qualify for jobs and job promotions, to enable further education and training, and to enhance an adult's personal satisfaction. (GED Testing Service, 2009, p. 10)

However, the purpose statements also prohibit what the testing agency considers test misuse:

> GED test scores should not be used to make inferences regarding any non-cognitive aspects often developed by attending high school, such as creativity, team work, planning and organization, ethics, leadership, self-discipline, and socialization. In addition, ACE policy clearly states that the GED Tests should not be used to validate high school diplomas and does not permit the tests to be administered to high school students still enrolled in school or high school graduates, except under special circumstances. Employers and postsecondary institutions are explicitly forbidden to use the GED Tests to verify the achievement level of high school graduates. (GED Testing Service, 2009, p. 10)

From these statements of testing purpose and potential misuse we begin to envision the types of evidence that could be used to develop a validity argument for the GED Tests. Another way to identify potential misuse is to listen to common criticisms of tests (e.g., narrowing the curriculum) and determine whether they should be considered as sources of invalidity to be investigated.

Crossing Testing Purposes With Sources of Validity Evidence

In this step, we bring in the validity framework implied in the AERA et al. (1999) *Standards*, which stipulates five sources of validity evidence. Explanation of all five sources is beyond the scope of this article and I assume readers are sufficiently aware of each source (if not—go read the *Standards*!). They are validity evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations with other variables, and (e) testing consequences. Table 1 presents some testing purposes and misuses associated with a fictitious testing program, and I cross them with the *Standards'* five sources of validity evidence. The fictitious program is a statewide mathematics test for Federal accountability under NCLB. The check marks in Table 1 illustrate the sources of evidence that likely should be used to support the use of the test (i.e., develop the validity argument) for each specific purpose (or guard against potential misuse).

Table 1 does not provide the details of the types of studies upon which the validity argument would be built, but most students of validity and most testing practitioners will be able to envision the types of studies to be conducted (e.g., alignment studies to confirm tests measure curricula, decision accuracy and consistency studies to confirm reliable achievement level classifications, etc.). The purpose of Table 1 is merely to illustrate how the three-step process can be used to develop a validity argument. Two points about Table 1 are worth noting. First, like Kane (this issue) stated, "The kinds of evidence required for validation are determined by the claims being made, and more-ambitious claims require more evidence" (p. 3). Thus, claims regarding student proficiency, or accountability for the state, schools, or teachers, obviously require more validity evidence. Second, some purposes and criticisms, such as the effect of

Table 1
Proposed Validity Framework for Fictitious Statewide Mathematics Test

Testing Purpose (Potential Misuse)	Source of Validity Evidence				
	Content	Internal Structure	Relations With External Variables	Response Processes	Testing Consequences
Measure students' math proficiency with respect to the state curriculum frameworks	✓			✓	
Determine whether students are at the basic, proficient, or advanced math achievement level for their grade	✓	✓	✓	✓	
Provide information regarding students' math proficiency that can be used for state, school, and teacher accountability	✓		✓		✓
Provide information that can be used to improve instruction at the classroom, school, district, and state level	✓	✓			✓
(Teachers teach to test rather than to curriculum frameworks)				✓	✓
(Students drop out of school to avoid taking test)					✓

the test on instruction or dropout, can only be evaluated by examining evidence of testing consequences. This is another point emphasized by Kane (this issue).

Gains, Losses, and Limitations

In my reaction to Kane (this issue), I praised Dr. Kane's current and prior work but suggested a simplification of his argument-based approach. Although I believe that the approach I proposed is simpler, I admit that by dropping the interpretation/use argument we lose the specific links that tie test-based interpretations to underlying assumptions regarding scoring, generalization, and extrapolation. The simpler three-step model I proposed gives the validation team more leeway to come up with specific studies to address specific purposes, and it provides a framework based on the

Standards' five sources of validity evidence to help them determine which types of studies are most relevant to each purpose. However, this leeway requires responsible validators who will be comprehensive in completing each of the three steps, will understand the underlying assumptions to be tested, and will ensure that sufficient evidence is collected to evaluate each purpose.

In closing, I thank Dr. Kane for once again moving validity theory and test validation forward by providing us with (a) a comprehensive understanding of how validity theory evolved, (b) an understanding of how the argument-based approach was influenced by construct validity theory, and (c) a reminder that, for test use to be valid, the positive consequences associated with a testing program must outweigh the negative consequences. If I had more space, I would comment further on all of the good advice Kane provides for validation as well as the many cautions and fallacies regarding score misinterpretation. Rather than do that, I will just recommend you read it and absorb as much as you can! I hope the approach I outlined here will add to the good advice provided by Kane and prove useful to those who are assigned the task of evaluating the use of a test for particular purposes.

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.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement*, *8*(2–3), 70–91.

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3–26). Mahwah, NJ: Lawrence Erlbaum.

Forte, E. (2012, January). *A next generation of validity evaluation and technical documentation*. Paper presented at the meeting of the Northeastern Educational Research Association, Rocky Hill, CT.

GED Testing Service (2009). *Technical manual: 2002 series GED tests*. Washington, DC: American Council on Education.

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

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–100). Washington, DC: American Council on Education.

Author

STEPHEN G. SIRECI is Professor of Educational Policy, Research, and Administration and Director of the Center for Educational Assessment, University of Massachusetts Amherst, School of Education, 156 Hills South, Amherst, MA 01003; sireci@acad.umass.edu. His primary research interests include test development, test validation, cross-lingual assessment, educational assessment policy, and fairness issues in testing.