JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME measurement in education

Journal of Educational Measurement
Spring 2013, Vol. 50, No. 1, pp. 115–122

Validation as a Pragmatic, Scientific Activity

Michael T. Kane
Educational Testing Service

This response to the comments contains three main sections, each addressing a subset of the comments. In the first section, I will respond to the comments by Brennan, Haertel, and Moss. All of these comments suggest ways in which my presentation could be extended or improved; I generally agree with their suggestions, so my response to their comments is brief. In the second section, I will respond to suggestions by Newton and Sireci that my framework be simplified by employing only one kind of argument, a validity argument, and dropping the interpretation/use argument (IUA); I am sympathetic to their desire for greater simplicity, but I see considerable value in keeping the IUA as a framework for the validation effort and will argue for keeping both the IUA and the validity argument. In the third section, I will respond to Borsboom and Markus, who raise a fundamental objection to my approach to validation, suggesting that I give too much attention to justification and too little to truth as a criterion for validity; I don't accept their proposed conception of validity, and I will indicate why.

Brennan, Haertel, and Moss

Brennan (this issue), Haertel (this issue), and Moss (this issue) expand on issues raised in my paper, and in doing so add a lot to this discussion. I am in general agreement with what they have to say, and therefore, my comments on these three commentaries will be quite brief.

Brennan provides an overview of the development of the argument-based approach to validity and a summary of its essential structure that is probably clearer than mine; it certainly is more concise. He goes on to provide a brief, but relatively thorough, explication of the structure of scoring inferences, and he discusses some ways in which generalization and extrapolation inferences are entwined. He concludes by making a strong case for the role of consequences in validity, and he examines the distribution of responsibility for evaluating the consequences of test uses. Brennan's discussion of the validity of some proposed uses of value-added models for high-stakes decisions raises a number of important issues—including aspects of test development, scaling, causal assumptions, and consequences—that are questionable in the context of value-added-model-based decision models and that deserve a lot more attention.

Haertel discusses a range of consequences of testing programs, including intended direct effects, intended indirect effects, and unintended effects, and he makes the point that unintended effects can be hard to identify and evaluate, especially if they develop slowly. He suggests that involving researchers from other disciplines (e.g., social psychology) could be of significant help in understanding the immediate and more distal consequences of test programs. As Haertel concludes, the measurement community does not have to assume responsibility for all of the

potential consequences of testing programs, but research on consequences is likely to enhance the value of testing programs.

Moss extends the discussion of consequences to the real world of day-to-day, week-to-week, and month-to month decision making in educational contexts, and in doing so she makes an important contribution to this issue. I was not familiar with much of the literature that she cites on how data is actually used in schools, and I suspect that many readers of JEM share this blind spot. As Moss points out, these issues go beyond validity theory *per se*, but validation has a role in helping to facilitate the effective use of score-based evidence in education. I had given a lot of attention to the role of validation in avoiding inappropriate interpretations and uses of test scores, and Moss reminds us that we have an affirmative obligation to make test scores and other kinds of information a positive force in schooling.

Newton and Sireci

Newton and Sireci both suggest that an argument-based approach to validation can be simplified by employing only one argument, preferably the validity argument; they do not see a need for a separate interpretation/use argument (IUA). Newton suggests that the IUA can be viewed as a preliminary and less well-developed version of the validity argument. Sireci suggests that, if test developers articulate the intended purpose of a test well enough, there would be no need for a separate IUA.

If a validator states the claims inherent in the proposed interpretation and use of the scores clearly and then evaluates these claims, I would have no complaints. I am not too particular about how the claims inherent in an interpretation and use are identified and documented; my concern is that it be done thoughtfully and in some detail, but exactly how it is to be done is not so important. Much good work in validation has been done over the last 100 years without mentioning any kind of "argument" (e.g., Frederiksen, 1984). My main concern is that we be clear about what is actually being claimed and that we avoid overstating the claims being made (and thereby setting up straw men that are too easily knocked over) or understating the claims (and thereby begging various questions). I see the specification of an IUA as a useful safeguard that makes a misspecification of the interpretations and uses to be validated less likely.

Confusion about the claims being made is especially likely when evaluating new tests for new applications, especially if the new application is similar in some ways to a familiar application. For example, if the task is to validate an end-of-course achievement test, it is natural to think of focusing on examining how well the test content represents the course content (as well as the appropriateness of the scoring procedures and the generalizability of the results); in many cases (e.g., classrooms with a well-defined curriculum), this makes perfectly good sense. However, Frederiksen (1984) recognized that in developing tests for courses designed to prepare students for some kind of performance, the validation effort should focus on the performances and not on the content of a curriculum. Similarly, Shepard (1993) recognized the important differences between the goals of a kindergarten readiness

testing program and those of standard academic selection programs and I (Kane, 1986) suggested a model for licensure and certification examinations that is substantially different from the standard, criterion-based model for employment testing. In each of these cases, it was necessary to specify the inferences and assumptions inherent in the application at hand and to distinguish it from a well-established model.

The problems associated with not specifying the proposed interpretation and use in some detail before getting into the validation effort can be especially acute if multiple players are involved. In my experience, it is not unusual for the different players associated with a testing program to be in agreement about the proposed interpretation as long as it is stated at a fairly general level (e.g., to measure achievement in mathematics at the end of high school), but the agreement tends to fray as we get into the details. Years ago, I was in a meeting in which we were developing plans for a testing program to assess the work-related skills of high-school students. The primary purpose of the program was to help students who were not going on to higher education to evaluate their readiness for the world of work. Early in the meeting I asked whether the test would be used for employment decisions, because if so, the validation effort would have to address particular legal requirements associated with employment testing. I was assured that it would not be used for employment decisions. About an hour later, I asked who was going to use the tests (and, as a practical matter, pay for the use) and why they would want to spend their limited resources in this way; I was told that employers would adopt the test as an efficient way of deciding who to hire. Within an hour or so a single committee had expressed two diametrically opposed views about a critical aspect of how the test was to be used. This problem obviously gets worse if we have multiple organizations with different agendas (as is usually the case in high-stakes educational testing programs). The development of an explicit statement of the proposed interpretation and use (e.g., an IUA) can help to identify such inconsistencies and ambiguities before they cause too much trouble.

In a sense, the idea of developing the IUA as a template for validation is to make critical evaluations of what is being claimed a routine component in validation. After developing the IUA, we may conclude that some more-or-less standard approach to test development or validation is appropriate for the case at hand, but even in these cases, it is useful to lay out the proposed interpretation and use in enough detail that we fully understand the claims to be evaluated.

The IUA and the validity argument are different kinds of argument. The IUA is to provide a clear statement of the claims based on scores. It is descriptive rather than evaluative. It is to specify the inferences and assumptions inherent in the proposed interpretation and use. It is not primarily an argument for the proposed interpretation/use, but rather a specification of the interpretation and use.

The IUA is evaluative only in the sense that, if one cannot put forward a coherent and plausible IUA indicating how one might get from the scores to the conclusions and decisions based on the scores, there would not be much point in going any further. However, the fact that we can specify a plausible IUA does not provide much assurance that the inferences actually hold up for this test, in this context, for this population.

The validity argument is to provide a critical evaluation of the proposed interpretation and use. The goal is to "kick the tires" by evaluating the inferences and assumptions inherent in the proposed interpretation/use. A carefully prepared IUA facilitates this process by specifying what is being claimed (and just as important, what is not being claimed); the IUA provides a framework for the validity argument. If the IUA is judged to be coherent and complete and its inferences and assumptions are supported by all relevant evidence, the proposed interpretation/use can be considered valid.

I think that a separate IUA is helpful, because it makes it less likely that some key assumption will be passed over without thought or that some unnecessary assumptions will be implicitly adopted (e.g., see Frederiksen, 1984; Kane, 1986; Shepard, 1993). As noted above, it certainly is possible to develop a fine validity argument without formally specifying an IUA, especially if the validator is experienced and thoughtful and the interpretation/use is not too complicated, but I think that it is unnecessarily risky to skip this step. The development of the IUA should not require a lot of time or resources. It does not require the collection or analysis of any data, although it can be informed by data and experience. It mainly involves some serious thought about what is being claimed (i.e., the inferences and assumptions inherent in the proposed interpretation and use) and about the *a priori* plausibility of these claims.

The argument-based approach is contingent in that the evidence needed for validation depends on the inferences and assumptions inherent in the proposed interpretation/use, and these inferences and assumptions have to be specified in some way (e.g., as an IUA) for this approach to work well. Validity theories face difficulties in identifying any particular kind of evidence as essential, or as irrelevant, because test-score interpretations and uses are so varied. If we identify some specific kind of evidence as essential for validity, we generally can come up with some cases in which it is not necessary; if we identify some kind of evidence as irrelevant, we generally can come up with multiple cases where it is necessary, or at least highly relevant.

The argument-based approach gets around this problem by making validation requirements contingent on the claims being made. If the proposed interpretation/use claims or assumes X, then evidence supporting X is necessary for validation; if the proposed interpretation/use does not claim X, explicitly or implicitly, then evidence supporting X is largely irrelevant to validation. As a result, we can impose serious, contingent requirements and make them stick. We can say that if you are going to claim that the scores can be used to predict performance in some context, we expect (require) evidence that these predictions are accurate enough for the case at hand. If you propose the use of the scores in a particular way, we expect you to justify this use by showing that the consequences of such use are likely to be positive. The requirements are contingent but they are clearly reasonable, and they are clearly reasonable because they are contingent.

To carry out this kind of contingent analysis, one does not need to employ the terminology which I have suggested but one does have to implement two conceptually distinct steps: (1) state what is being claimed and (2) evaluate the plausibility of these claims. I associate step 1 with the IUA and step 2 with the validity argument.

Borsboom and Markus

Borsboom and Markus focus on a single issue: the relationship between truth and validity. Using their terminology, which is based on a traditional philosophical definition of "knowledge" as "justified true belief" (Musgrave, 1993), they are concerned primarily with "true belief" while I am more concerned with "justified belief." Borsboom and Markus recognize that justification is important but they see it as distinct from truth, and their notion of validity focuses on Truth with a capital "T." I focus on plausibility and justification and leave Truth in the background. I assume that a careful specification of what is being claimed and an evaluation of the resulting IUA in terms of its coherence and its consistency with relevant empirical results generally will get us to a more accurate, useful, and appropriate interpretation of test scores than we would achieve otherwise, but I do not expect to capture Truth.

I accept Borsboom and Markus's contention that the best available evidence could lead to the acceptance of an interpretation that subsequently would be shown to be incomplete, inaccurate, or otherwise inadequate. Sound methodology should make it more likely that we will reach accurate conclusions, but even the most careful methods cannot provide certainty. Science deals with observations, theories, evidence, justification, and probability, but absolute, eternal Truth is beyond its grasp.

In many testing situations (including most high-stakes contexts), talk of Truth seems hollow. For example, consider an achievement test in an academic area such as world history. If we asked a group of subject-matter experts whether the proposed interpretation of scores in terms of achievement in the area were true or not, we probably would get blank stares and questions about what we mean by "true." If we ask whether the interpretation is appropriate or reasonable, we probably would get a lively discussion about format (emphasis on objective items or essays), about content (coverage of different periods, different geographic areas, and different kinds of events and trends), and about cognitive demands (recall or analysis). It is not simply that questions about Truth are hard to answer in these contexts; rather, it is not clear that they are even meaningful. In some cases, we might have serious disputes about the appropriateness of test content (especially in areas like history), but such disputes are more likely to be about values or points of view than about Truth.[1]

Borsboom and Markus express concern about emphasizing justified belief over true belief. They fear that one might develop a validity argument for an interpretation without developing a better understanding of how well the test is functioning. This is a legitimate concern, but the argument-based approach to validation is intended to make inadequate analyses less likely by requiring that the IUA specify how the scores are to be interpreted and used, that the claims being made be evaluated analytically and empirically, and that plausible alternative interpretations/uses be considered.

Borsboom and Markus ask how I would evaluate their phlogiston example (see Borsboom and Markus, this issue). They pose the question as a three-option multiple-choice item, and I of course pick a fourth option, "none of the above," as my preferred answer. I would say that within the context of 18th century science, the interpretation of certain measures as indicators of the amount of phlogiston in a body was reasonable in light of prevailing theory and all available evidence and therefore was valid in that context. In the context of current theory, interpretations in terms of phlogiston do

not make sense and would not be considered valid. The phlogiston theory has been overturned, and interpretations based on the discredited theory have been discarded. Scientific theories and interpretations based on these theories are fallible and subject to revision.

This bothers Borsboom and Markus because they seek Truth rather than justified belief. I am more pragmatic (with a small "p"). I am concerned about what can reasonably be claimed on the basis of test scores in the current context. I am less interested in what "rational scientists would arrive at, should they continue their investigations for an infinity of time" (Borsboom and Markus, this issue, p. 113) than I am in what a scientist can reasonably (and provisionally) conclude here and now.

As a thought experiment, it is useful to consider how the Borsboom and Markus model might have been applied to phlogiston measurements when the phlogiston theory was in its prime. As Borsboom and Markus posit, interpretations in terms of phlogiston once were supported by "a quite impressive theory on the nature of burning as phlogiston emission" (this issue, p. 112). Furthermore, the theory assumed that phlogiston existed and caused the empirical outcomes used to estimate it. It was a textbook example of the kind of validation analysis advocated by Borsboom, Mellenbergh, and Van Heerden (2004).

As I understand Borsboom and Markus, they would consider the interpretation in terms of phlogiston to be justified in the 18th century but not to be valid now or in the 18th century (because the theory is and was untrue). They define validity in terms of the Truth of the interpretation, and they take Truth to be eternal.

This view would seem to have some serious implications for the validation of current measurements. If a proposed interpretation is incoherent or inconsistent with empirical results, it is likely to be considered invalid (at least by me, and I think by Borsboom and Markus). If the proposed interpretation is coherent and consistent with well-established theoretical frameworks and a wide range of empirical evidence, I would consider it valid (provisionally). Borsboom and Markus might consider it plausible but probably would not want to say that it is valid because the current evidence might be overturned or reinterpreted in the future; for example, the phlogiston interpretation was very strongly supported in the 18th century, but according to their model it was not valid on this account. So even if we have what appears to be overwhelming evidence for an interpretation, the Borsboom and Markus framework would make us reluctant to make claims for its validity (i.e., Truth).

To go a bit further, take any interpretation that might be proposed now. Is it likely to be accepted in essentially the same form in 100 years, in 200 years, in 500 years? According to Borsboom and Markus, if the interpretation is rejected at any time in the future, it is invalid for all time. There are very few scientific theories or explanations from 500 years ago that are still taken seriously. Newtonian mechanics reigned for a few hundred years, but it was eventually overturned. Even Euclidean geometry, which was taken to be self-evident for over two millennia, has been replaced as a framework for describing the characteristics of physical space.

I think that very few, if any, of our current models in the social sciences will survive unchanged for 200 years (never mind into eternity), and therefore according to Borsboom and Markus, very few, if any, of our current interpretations are valid. And if a few of our current interpretations happen to be valid in the sense that they

are True, we have no way of identifying these gems. Under these circumstances, the most sensible strategy would be to reject the validity of any proposed test-score interpretation; to do otherwise would be a triumph of optimism over experience. How useful is a validity framework that leads us to reject all test-score interpretations as invalid?

I think that the Borsboom and Markus model is not helpful in addressing questions about the appropriateness of test-score interpretations or uses; rather, it points us toward an almost certainly fruitless quest for the "holy grail" of absolute, eternal Truth (Musgrave, 1993). It is hard to see how we can evaluate the accuracy of a proposed interpretation other than by evaluating its coherence and the evidence relevant to its claims and assumptions. We have no direct access to Truth. We can assume that sound methodology will promote what Popper (1962) calls "verisimilitude" and Lakatos (1970) refers to as a "progressive research program," but making Truth the centerpiece in our conception of validity is not likely to be helpful.

Concluding Remarks

The argument-based approach does not provide an algorithm for validation, but it does provide a framework for designing and implementing validation efforts that address the claims based on test scores. It requires that the inferences and assumptions inherent in the proposed interpretation and use be specified (the IUA) and that these inferences and assumptions be critically evaluated (the validity argument).

Note

[1]Note that I am not claiming that the blessing of an expert committee is, in itself, adequate for the validation of an achievement-based interpretation, but I would expect an achievement test to pass this kind of challenge.

Acknowledgments

I would like to thank Brian Clauser and the *Journal of Educational Measurement* for the opportunity to present my views on validity, and I want to thank my colleagues for taking the time to comment on my paper. As a long-term fan of Karl Popper (1962) and Imre Lakatos (1970), I greatly appreciate thoughtful criticism of my work. This does not, of course, mean that I agree with all of the points made in the comments, but whether I agreed or not, I found them all interesting and stimulating.

References

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

Frederickson, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, *39*, 193–202.

Kane, M. (1986). The future of testing for licensure and certification examinations. In B. Plake & J. Will (Eds.), *The future of testing* (pp. 145–181). Hillsdale, NJ: Lawrence Erlbaum.

Kane, M. T. (1994). Validating interpretive arguments for licensure and certification examinations. *Evaluation and the Health Professions*, *17*, 133–159.

Lakatos, I. (1970). Falsification and the methodology of scientific research programs. In I. Lakatos and A. Musgrave (Eds.), *Criticism and the growth of knowledge*. Cambridge, UK: Cambridge University Press.

Musgrave, A. (1993). *Common sense, science and skepticism: A historical introduction to the theory of knowledge*. Cambridge, UK: Cambridge University Press.

Popper, K. R. (1962). *Conjecture and refutation: The growth of scientific knowledge*. New York, NY: Basic Books.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education*, Vol. 19 (pp. 405–450). Washington, DC: American Educational Research Association.

Author

MICHAEL T. KANE serves as the Samuel J. Messick Chair in Validity at Educational Testing Service, Rosedale Road, Princeton, NJ 08541; mkane@ets.org. His primary research interests include validity theory, generalizability theory, licensure and certification testing, and standard setting.