Comment: Making an Argument for Design Validity Before Interpretive Validity

Derek C. Briggs
School of Education
University of Colorado

Michael Kane (this issue), in his thought-provoking attempt to mold an argument-based structure onto the seemingly amorphous notion of test validity, reminds me that test development should not fall outside the purview of scientifically based research. Recently, the National Research Council (NRC) published a report describing six principles thought to underlie all scientific inquiry in educational research (National Research Council, 2002). It is worth recapping these principles here:

1. Pose significant questions that can be investigated empirically.
2. Link research to relevant theory.
3. Use methods that permit direct investigation of the question.
4. Provide an explicit and coherent chain of reasoning.
5. Replicate and generalize across studies.
6. Disclose research to encourage professional scrutiny and critique.

It seems to me that elements of what Kane (this issue) describes as a test’s interpretive argument fit within these principles rather nicely. In particular, I think there is a strong connection between validating steps of an interpretive argument and validating the links in a chain of causal reasoning, as in the fourth (NRC) principle.

Requests for reprints should be sent to Derek C. Briggs, University of Colorado School of Education, Room 211, 249 UCB, Boulder, Colorado 80309. E-mail: derek.briggs@colorado.edu
Only if the links are made apparent can their strength be properly evaluated. As Kane writes: “If it is not possible to come up with a coherent interpretive argument that gets us from the observed performance to the proposed interpretation and uses, there is no point in proceeding further.” We can think about this as a statement about causal inference by equating observed performance with the word cause, proposed interpretation with the words causal mechanism, and uses with the word effect.

It seems clear that test validation must be seen as a comparative venture when it comes to evaluating causal inferences. We can conceive of taking a particular test as an experimental treatment, not taking a test or taking a different test as an experimental control, and the decision reached as the experimental outcome. Unfortunately, this is precisely the sort of evidence that is most difficult, if not impossible, to obtain before a test has been administered for its intended purpose. The implication of this is clear: No test, when first administered to its target population, can have the full proof of validity behind it. At best, it may have what Kane has described as an interpretive argument, much of which remains to be proven. It seems to me that, for rather obvious reasons, large-scale test developers will be loathe to make an interpretive argument explicit (a threat to NRC Principle 6) if doing so makes it clear that elements of the argument have yet to be validated. To do so would open the door for any number of lawsuits in contexts where tests are used for high-stakes purposes. This seems to point to a paradoxical problem in the test validation enterprise suggested by Kane’s framework: If a test has not been fully validated, than using it to make high-stakes decisions becomes questionable. But if it has not been administered to a sample from the target population for the expressed purpose of the test, it becomes impossible to fully validate it!

If measurement specialists wish to have an impact on the design and use of high-stakes tests, it seems to me that we need to embed Kane’s vision of an argument-based approach to validity within a somewhat more pragmatic paradigm, such that we have some hope of seeing our theoretical visions of test validity fulfilled in practice. We might consider separating, and as I illustrate, somewhat expanding, Kane’s interpretive argument into two stages of test validation.

1. Develop an interpretive argument for a test that can be justified with respect to the test’s design before it has been administered for its intended purpose. This would be termed “Design Validity.”
2. Evaluate the interpretive argument for the test’s use after it has been administered for its intended purpose. This would be termed “Interpretive Validity.”

Kane (this issue) has illustrated his argument-based approach to test validation using the example of a certification exam. I briefly push his idea a little further within the context of a high-stakes standardized achievement test. With the advent of the No
Child Left Behind (NCLB) legislation (NCLB, 2001; U.S. Department of Education, 2004), it has become increasingly common for states to use high-stakes tests to determine whether some threshold of students in public schools show evidence of mastering state-mandated content standards. An implicit assumption seems to be that students who meet or exceed state standards for learning will be those who are most successful in a range of post-high school outcomes. How should we proceed in the validation enterprise for high-stakes tests to be used as part of NCLB?

In Figure 1, I show what I call the expanded two-stage argument for validity. The first four steps are an expansion of Kane’s first step in the interpretive argument for his certification test example. These steps of the argument make more explicit the sort of work that would be expected in the design of any high-stakes test:

<table>
<thead>
<tr>
<th>The Expanded Two Stage Argument for Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Theory for how students learn and develop along the construct to be measured (i.e., state standard)</td>
</tr>
<tr>
<td>2. Development of items that represent the intended construct</td>
</tr>
<tr>
<td>3. Rules for going from an observed performance on an item to an observed score on an item</td>
</tr>
<tr>
<td>4. A model for relating item scores back to the construct of interest</td>
</tr>
<tr>
<td>5. Generalization of the observed score to the expected score</td>
</tr>
<tr>
<td>6. Extrapolation from the test to state standard(s)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>7. Extrapolation from state standards to post high school outcomes</td>
</tr>
<tr>
<td>8. Evaluation of post high school outcomes as a function of test results</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

**FIGURE 1** Validating a high-stakes test used for No Child Left Behind.
starting with a theory for how students learn and develop along the construct of interest, crafting items that elicit evidence of development on the construct, creating rules that assign scores to observed item responses, and formalizing a model that relates item scores back to the construct of interest. Steps 1 to 4 are encapsulated by what have been called progress or construct maps (cf. Wilson & Sloane, 2000; National Research Council, 2001, pp. 115–119). Steps 1 to 6 constitute what might be reasonably labeled as design validity. These are the steps we should expect any test developer to make and justify before administering a high-stakes test to the target population. For a test to be fully validated, we would have to make the case for interpretive validity by justifying Steps 7 and 8. With some exceptions, these are steps that cannot be justified until the test has actually been administered. In particular, the evaluation of post-high school outcomes (e.g., college performance, salaries, quality-of-life indicators) as a function of test results would probably require the analysis of student data 5 to 10 years after testing.

Splitting Kane’s (this issue) interpretive argument framework into two stages codifies our sequential expectations for test validity. Note that within the design validity stage we find evidence of the traditional content, criterion, and construct validity indicators. These are further enhanced by embedding them within a context in which measurement follows a theory under which students learn and develop. The interpretive validity stage marks the point at which all inferences based on test use must be evaluated with an explicit experimental or quasi-experimental design. Hence, design validity can be viewed as the necessary, but not sufficient, condition for test validity. Without establishing interpretive validity, the validation enterprise may crumble like a house of cards. I promote the idea that design validity is a reachable starting point for those developing high-stakes tests. It should fall on test contractors to fund research to evaluate interpretive validity once a test with design validity has been administered. Kane’s article makes abundantly clear the daunting enormity to the test validation enterprise. It falls on measurement professionals to chart a course for this enterprise that can be navigated in practice.

REFERENCES


