Validity Arguments for Alternate Assessment Systems

Scott Marion, Center for Assessment

Reidy Interactive Lecture Series

Portsmouth, NH

September 25-26, 2008

Acknowledgments

- I appreciate the opportunity to learn from the states, evaluators, expert panelists, and my colleagues participating in the NAAC-GSEG and the NH-EAG projects.
- Much of what I am presenting today is drawn from collaborative work with Marianne Perie and from many conversations with Brian Gong.

Documenting Technical Quality

- Our interest in validity arguments for alternate assessment grew out of the need to provide defensible documentation of the technical quality of these assessments.
- We argue that the purpose of the technical documentation is to provide data to support or refute the validity of the inferences from the alternate assessments at both the student and program level.

Challenges of Alternate Assessments

- Small, very heterogeneous group of students
  - Creates difficulties for statistical analyses
- Flexibility in:
  - Assessment targets
  - Assessment events
  - Administration
- All create psychometric challenges, especially for comparability

Is it Psychometrics or Social Justice?

- Initially, we focused on consequences—yes they are part of validity!—because the intended consequences were a major rationale for including all students in standards-based education.
- However, we realized even before writing the NH EAG proposal that we needed to more completely evaluate the technical quality of alternate assessments.

Many Possibilities

- Elsewhere we (Marion & Perie, 2008) have presented examples of evaluation questions and potential studies within familiar (e.g., joint standards) and less familiar (e.g., *Knowing What Students Know*, Ryan, 2002) frameworks for structuring *legitimate* validity evaluations.
- Further, the work of the NHEAG and NAAC have demonstrated that many familiar forms of analyses are possible even if they require some different thinking.

Why a Validity Argument

- Serves to organize studies
- Provides a framework for analysis and synthesis
- Uses a falsification orientation
- Forces critical evaluation of claims
  - Basically requires the user, developer, and/or evaluator to search for all the reasons why the intended inferences are NOT supported
  - In practice we cannot search for ALL the reasons, so we need to prioritize studies.

Kane’s argument-based framework

- “…assumes that the proposed interpretations and uses will be explicitly stated as an argument, or network of inferences and supporting assumptions, leading from observations to the conclusions and decisions. Validation involves an appraisal of the coherence of this argument and of the plausibility of its inferences and assumptions” (Kane, 2006, p. 17).

Two Types of Arguments

- *An interpretative argument specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances*
- *The validity argument provides an evaluation of the interpretative argument* (Kane, 2006)

The Interpretative Argument

- Essentially a mini-theory—the interpretative argument provides a framework for interpretation and use of test scores
- Like theory, the interpretative argument guides the data collection and methods and most importantly, theories are falsifiable as we critically evaluate the evidence and arguments

More Simply…

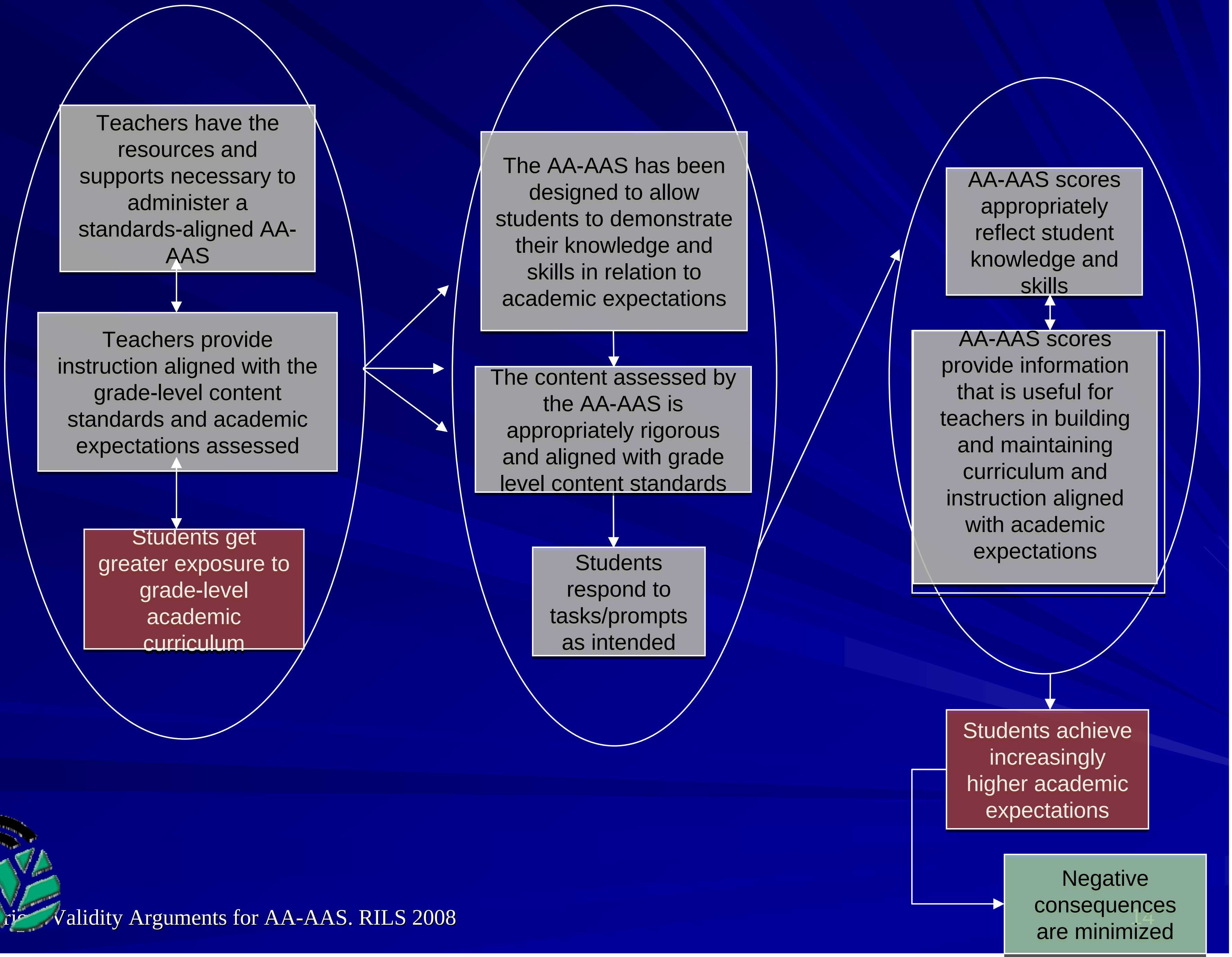
- Test validation is basically the process of offering assertions (propositions) about a test or a testing program and then collecting data and posing logical arguments to refute those assertions
  - If the assertions cannot be refuted, we can say that they are tentatively supported (and that's the best we can do!)
- A simple organizational scheme for the propositions
  - What does the testing practice claim to do;
  - What are the arguments for and against the intended aims of the test; and
  - What does the test do in the system other than what it claims, for good or bad? (Shepard, 1993, p. 429)

Values and Consequences

- Evaluating a decision procedure requires an evaluation of values and consequences
- “To evaluate a testing program as an instrument of policy [e.g., AA-AAS under NCLB], it is necessary to evaluate its consequences” (Kane, 2006, p.53)
- Therefore, values inherent in the testing program must be made explicit and the consequences of the decisions as a result of test scores must be evaluated.
  - Yet one more authority in the long line of validity theorists (Cronbach, Messick, Shepard, Linn, Haertel, Moss) making it quite clear that consequences are an integral part of test validation

Getting Started

- Katherine Ryan (2002) and others have suggested that laying out a more general “theory of action” is a useful starting point for developing a more complete validity argument
- Marianne Perie, and I created the following EXAMPLE theory of action for an alternate assessment system…



Teachers have the resources and supports necessary to administer a standards-aligned AA-AAS

Teachers provide instruction aligned with the grade-level content standards and academic expectations assessed

Students get greater exposure to grade-level academic curriculum

The AA-AAS has been designed to allow students to demonstrate their knowledge and skills in relation to academic expectations

The content assessed by the AA-AAS is appropriately rigorous and aligned with grade level content standards

Students respond to tasks/prompts as intended

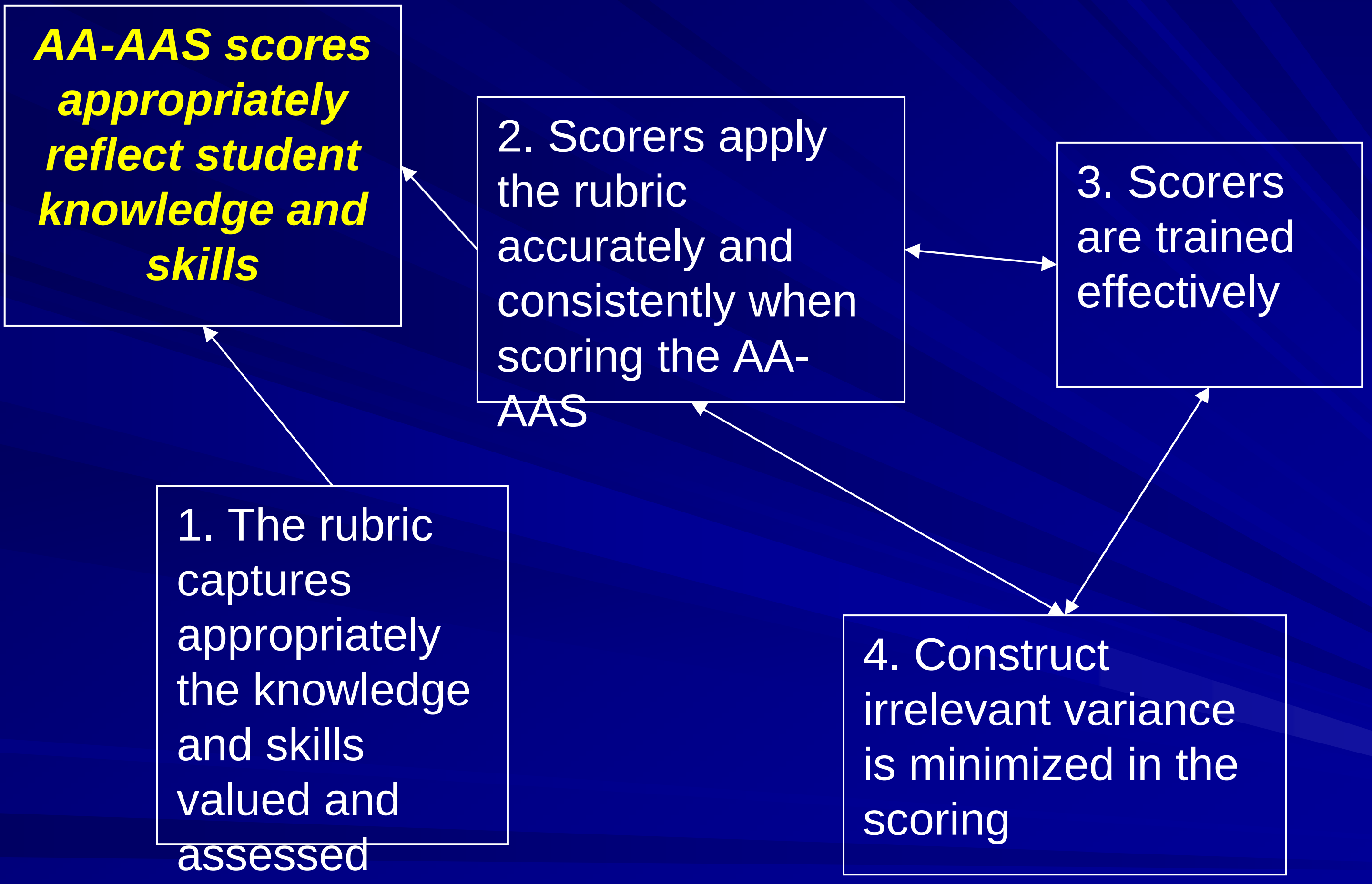
AA-AAS scores appropriately reflect student knowledge and skills

AA-AAS scores provide information that is useful for teachers in building and maintaining curriculum and instruction aligned with academic expectations

Students achieve increasingly higher academic expectations

Negative consequences are minimized

Propositions underlying a single claim



AA-AAS scores appropriately reflect student knowledge and skills

2. Scorers apply the rubric accurately and consistently when scoring the AA-AAS

3. Scorers are trained effectively

1. The rubric captures appropriately the knowledge and skills valued and assessed

4. Construct irrelevant variance is minimized in the scoring

- One of the most effective challenges to interpretative arguments (or scientific theories) is to propose and substantiate an alternative argument that is more plausible
  - With AA-AAS we have to seriously consider and challenge ourselves with competing alternative explanations for test scores, for example…
    - “higher scores on our state’s AA-AAS reflects greater learning of the content frameworks” OR
    - “higher scores on our state’s AA-AAS reflects higher levels of student functioning” OR
    - “higher scores on our state’s AA-AAS reflect greater understanding by the teachers on how to gather evidence or administer the test”

Evaluating the Validity Argument

- Haertel (1999) noted that the individual pieces of evidence do not make the assessment system valid or not, it is only by synthesizing this evidence in order to evaluate the interpretative argument can we judge the validity of the assessment program.
- But, it is hard to find “rules” in educational measurement to help us to this synthesis and evaluation.
  - – However, much can be learned and borrowed from program and policy evaluation.

Synthesis & Evaluation

- Synthesizing all of this information to arrive at a judgment about the testing program is an intellectual challenge in part because we’re working along continua of evidence and arguments.
- The interpretative argument is used to structure the evaluation.
  - The propositions should be written in such a way so that we can judge whether the evidence supports or does not support this particular claim.

Evaluating the Propositions

- We have to critically evaluate the evidence and logic that support or refute the specific propositions.
- In the context of states’ large assessment systems, we do not have the luxury of concluding, “that’s not working, let’s start over.”
  - In cases when the findings refute the propositions, we need to look for ways to improve the system .

Dynamic Evaluation

- In almost all cases when evaluating the validity of state assessment systems, the studies are completed over a long time span.
- We are rarely in the position of having all the evidence in front of us to make a conclusive judgment.
  - Therefore, we must engage in an ongoing, dynamic evaluation as new evidence is produced .

Back to Argument

- Basing the validity evaluation on a well-founded argument enables us to structure our dynamic evaluation so that we can build a comprehensive case for or against the assessment system.
  - It is possible—and it has happened recently—that the evidence suggests starting over
- Without the structure of an argument, we just have a bunch of studies and little guidance for how to weigh the different results.

For more information

- Scott Marion…smarion@nciea.org
- www.nciea.org
- www.naacpartners.org