




The role of policy goals in validity research

Sue Rigney
U.S. Department of Education


CCSSO National Conference on Student Assessment
June 2008






Validity – Whose Model?

● Traditional


- Content
 - Criterion
 - Construct “no criterion or universe of content is accepted as entirely adequate to define the quality to be measured”
(Cronbach and Meehl, 1955)
- 



Validity – Whose Model?

Messick (1989)

“Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness of inferences and actions* based on test scores or other modes of assessment.”






Facets of Test Validity

	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct Validity + Relevance/Utility
Consequential Basis	Value Implications	Social Consequences

Messick (1988)





Validity – Whose Model?

- Mislevy's Evidence Centered Design
 - Design validity in
 - Warrants and arguments
 - Detailed specifications

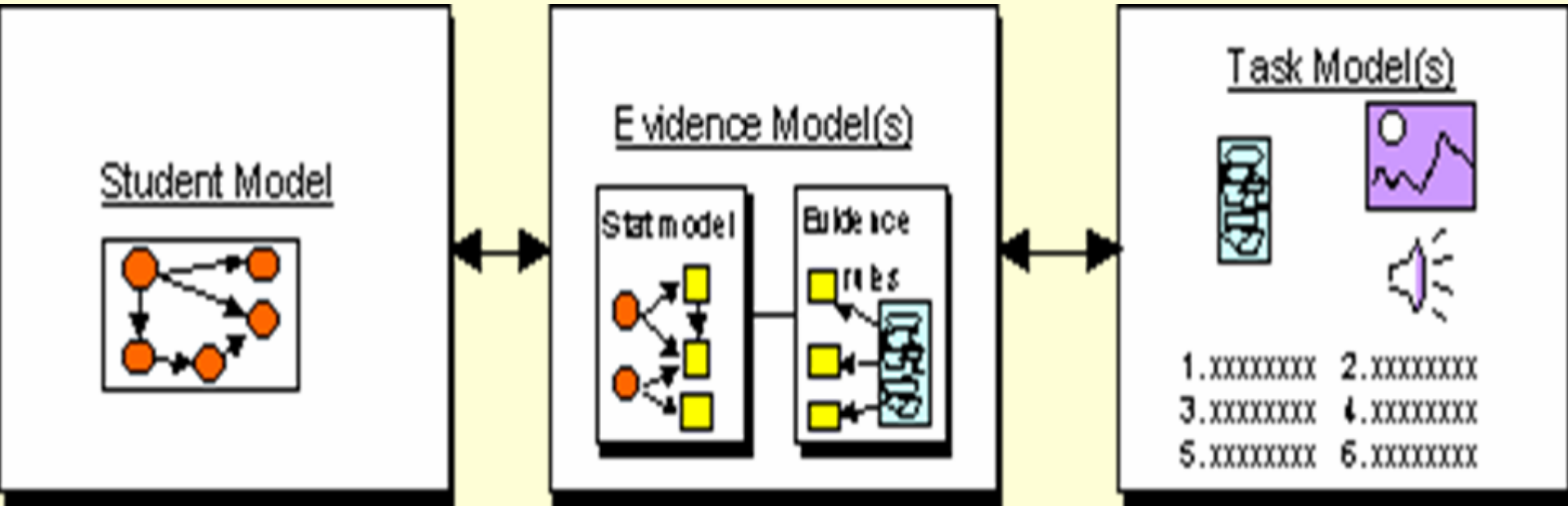
“Designing assessment products in such a framework ensures that the way in which evidence is gathered and interpreted bears on the underlying knowledge and purposes the assessment is intended to address.”


(Mislevy, Sternberg & Almond, 2000)



Figure 2 is a high-level schematic of the three central models in the CAF, and objects they contain. The CAF contains the core of the evidentiary-reasoning argument, from task design to observations to scoring to inferences about students.

The Three Central Models of the Conceptual Assessment Framework






Validity – Whose Model?

Lissitz & Samuelsen (2007)

“...thinking clearly about the content of the assessment is the first step and the most basic step.”

“...criterion validity (what we call utility) is a useful concern, for very limited purposes where a specific criterion can be specified and a theory does not exist.”

“...theory is important but...for most applications of psychometrics in education there is no external theory being developed and hence no opportunity or necessity to think hard about that issue.”






Perspective

		Theoretical	Practical
Investigative Focus	Internal	Latent Process	Content and Reliability
	External	Nomological Network	Utility and Impact



- 
- ...the type of validation that is most important depends on the inferences to be drawn from the test scores. Debates about which type of validity evidence is most fundamental to test score interpretation may actually serve to obscure this point.” (Crocker & Algina, 1986)
-



Can the desired inferences be found in
the policy goals?

If so, the policy goals provide a framework
for validity studies.





Policy

- Expressed (relies on data about)
 - Student scores and demographics
 - Other indicators related to achievement
 - Implied (relies on accountability machinery and assumptions)
 - Steady improvement over time (individual/group)
 - Lowest performing can/will improve the most
 - Assumed
 - Assessments sensitive enough to reflect instructional change
 - Assessments accurate and consistent enough to reflect change over time
 - Assumptions of the psychometric model are met by the measures employed
-



Federal Practice

Separated Assessment and Accountability
for practical reasons

2002-03 review of state accountability workbooks

2005-06 review of state assessment systems

2006-07 growth model pilot






Peer Review Outcome

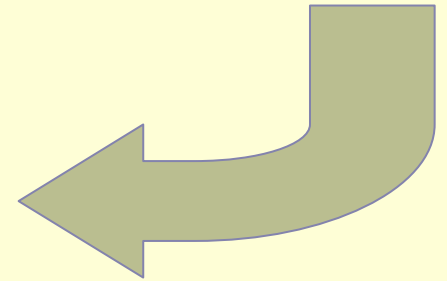
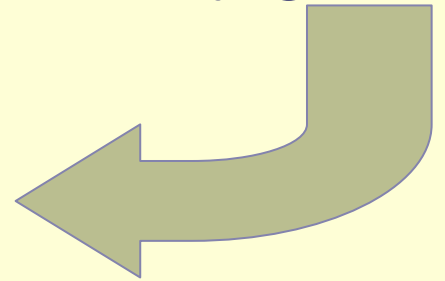
- **“Direct evidence regarding instruction in the schools needs to be obtained to evaluate the implicit assumption that schools with different student achievement outcomes differ in terms of the quality of their instructional offerings. This is needed regardless of whether the school accountability relies on measures of current status or on measures of student growth.”**
- **“Although states have made substantial progress in accumulating evidence that is relevant to the evaluation of the validity of the uses and interpretations of their assessments, questions about the consequences of the uses of assessment results have been given little, if any, attention despite the fact the Peer Review Guidance requires that evidence be provided regarding the consequences of the uses and interpretations of state assessments.”**

(Linn, 2007)




A practical validity agenda

- Validity questions = desired inference/policy goal
 - Student learning
 - School quality
- Data collection & analysis
 - Data accuracy (e.g. scores, demographics, graduation rate formula)
 - Assumptions of psychometric models
- Cost & feasibility
 - Time & resources available
 - Sampling plan sufficient for generalization
 - Classroom observation intrusive





Reconnect assessment and accountability

- Validate the assessment as a measure of reading skill, science knowledge, etc.
 - Consider the integrity of the statistical models employed
 - Alignment
 - Vertical scales
 - Performance standards
 - Longitudinal growth models
- 




Expanded validity agenda should include questions about...

Statistical validity

- Value-added estimates can be quite sensitive to choices made in the development of a vertical scale. (Briggs, Weeks, Wiley)

Policy validity

- Is this assessment having the desired effect?
- 



Challenges for States

- Clear statement of policy goals
 - Prioritize validity studies
 - Provide adequate resources
 - Sustain a long term validity agenda
 - Validate the use of assessment results
- 