I’m a Psychometrician and I’m Here to Help (and Learn)

Scott Marion
smarion@nciea.org
Center for Assessment
CCSSO-Large Scale Assessment Conference
San Francisco, CA
June 25-28, 2006

Overview of Talk

- The power of interdisciplinary collaboration
- Our focus and approach for evaluating technical adequacy (a brief review from last year's talk)
- Conceptualizing our major challenges
  - Flexibility-Standardization
- The Technical Manual
- Some Strategies for Section III

Multiple disciplines/Multiple perspectives

- This project could not have made the inroads that it has without the multiple and diverse perspectives brought together
- We suspect that projects with similar goals that do not rely on this type of strength will fall short of its desired outcomes

Expert Panel Roles & Responsibilities

- Function like a TAC-brainstorm, debate, advice
  - State assessment leaders need to help provide context and guide the panel members through the state contexts
  - Measurement experts need to think flexibly about how traditional psychometric understandings of technical criteria can be applied to AA-AAS
  - Curriculum experts need to help keep the measurement discussions grounded in what grade level content means for students with significant cognitive disabilities
  - Special education experts have to ensure that the measurement discussion doesn't inadvertently drift away from what is possible but that reflects high expectations for this group of students

The Problem of Technical Documentation

- Most psychometricians would likely rate validity as the most important technical criterion
- Yet, most technical manuals include only a superficial treatment of validity
- In fact, a recent call for the standardization of assessment technical reports, Becker and Camilli (2004) include validity as part of the required information, but it clearly appears secondary to reliability and other statistical concerns
  - By their own admission, Becker and Camilli were focusing only on the “nuts and bolts” and expected more would be added to the state’s technical manual

Validity Should be Central

- We argue that the purpose of the technical manual is to provide data to support or refute the validity of the inferences from the alternate assessments at both the student and program level.
- But, it is not so easy…

Expanding Technical Quality

- Following Linn, et al. (1991), we support the need to expand our conception of technical quality to better evaluate alternate assessment programs.
- Drawing on the work of Cronbach, Messick, and Shepard, the proposed evaluation of technical quality is built around a unified conception of validity.
  - For example, if the assessment program leads to positive instructional improvements for the state's students, it can be argued that these consequences support the validity of the program.

Validity framework

- Linn, et al. (1991) pointed out that we already have the theoretical tools for expanding validity investigations, but in practice validity is usually viewed too narrowly.
  - *Content frameworks are described, and specifications for the selection of items are provided for standardized achievement tests. Correlations with other tests and sometimes with teacher assessments of achievement may also be presented. Such information is relevant to judgments of validity but does not do justice to the concept* (p. 16).

Shepard (1993)

- Shepard (1993) advocated a straightforward means to prioritize validity questions. Using an evaluation framework, she proposed that validity studies be organized in response to the questions:
  - 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).
- The questions are directed to concerns about the construct, relevance, interpretation, and social consequences respectively.

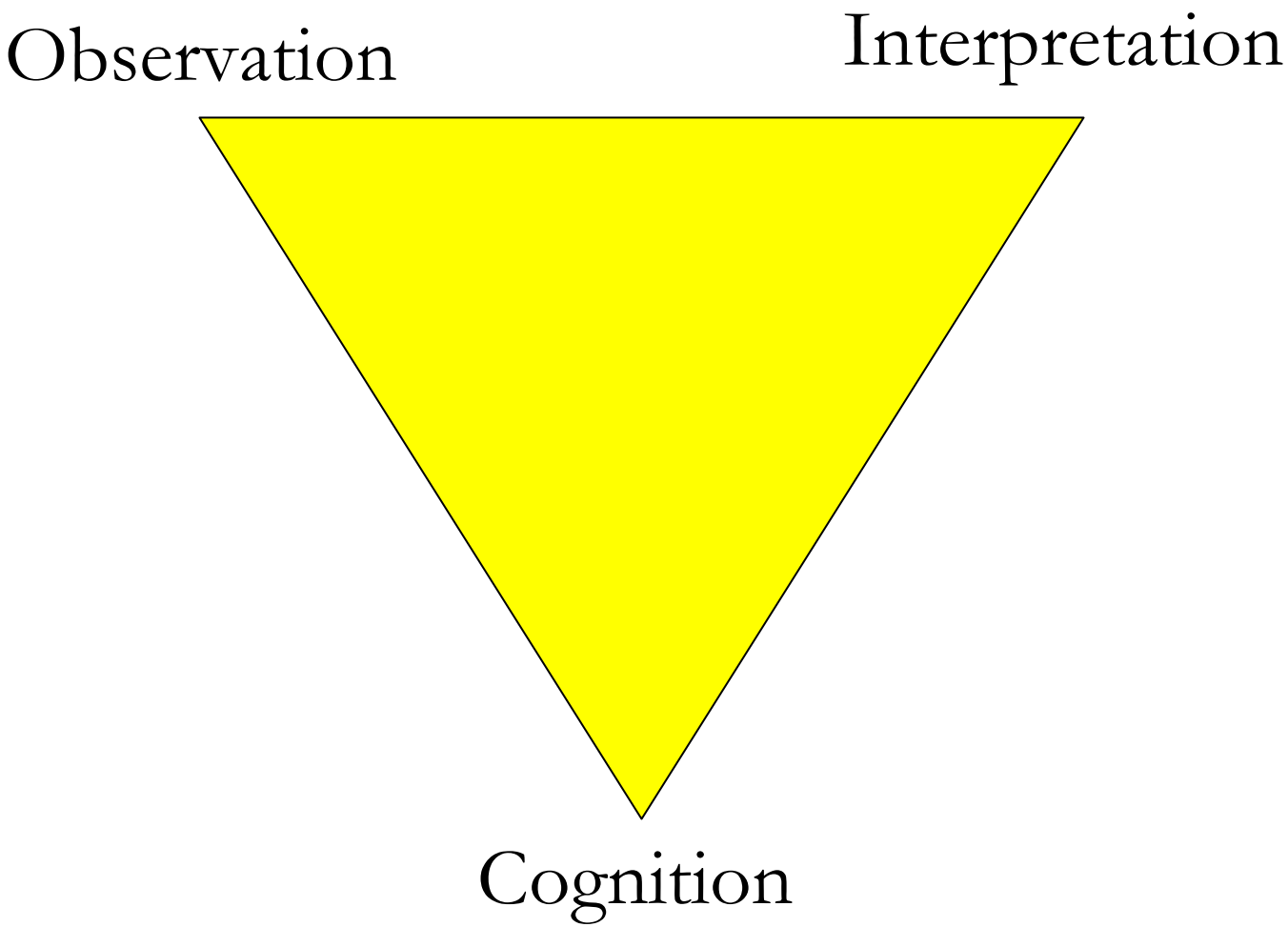
Haertel (1999)

- Typical technical manuals include chapters that are written to stand alone. The technical manual proposed here will be designed to weave the various chapters together through the use of the validity argument.
- Haertel (1999) reminded us that the individual pieces of evidence (presented in separate chapters) do not make the assessment system valid or not, it is only by weaving these pieces of evidence together into a coherent argument can we judge the validity of the assessment program. Therefore, this technical manual will be structured to facilitate such evaluations.

KWSK

- Fortunately, we are undertaking this work after the publication of *Knowing What Students Know*: *The science and design of educational assessment* (NRC, 2001), which synthesized a tremendous body of learning and measurement research and set an ambitious direction for the development of more valid assessments. *Knowing What Students Know* (KWSK) builds off of Mislevy's (1996) notion of assessment as a "process of reasoning from evidence" (p. 39).

The Assessment Triangle



A Heuristic

- We are using the assessment triangle as a heuristic to organize the validity evaluation.
- The triangle immediately reveals an important piece of missing information:
  - When we started this project, we were lacking models of cognition that could be applied generally to students with the most severe cognitive disabilities
  - However, work by Kleinert, Kearns, Browder, and others have started to provide some important insights in this realm

The Challenge of Alternate Assessments

- Documenting the technical qualities of alternate assessments is very difficult for many reasons:
  - heterogeneity of the group of students being assessed
  - relatively small numbers of students/tests
  - the measurement field has been slow to move away from traditional correlational indicators
  - often “flexible” assessment experiences

Flexibility and Standardization

- Gong, B. & Marion, S. F. (2006). Dealing with flexibility in assessments for students with significant cognitive disabilities. Minneapolis, MN: University of Minnesota, National Center for Educational Outcomes Synthesis Report No. 60. http://education.umn.edu/nceo/OnlinePubs/Synthesis60.html.

Flexibility and Standardization

- Nominal categories are NOT often useful for characterizing the technical aspects of the assessment
- This paper was written to explore the flexibility (variability)-standardization of the various components of alternate assessment
- There is no question that the evaluation of technical adequacy will interact with the types of alternate assessments being employed
- It was also designed to assist states consider where they might want to increase or decrease standardization

Assessment System Component	General	General w/ std. accommodations	Alternate on AAS
1. Flexibility in the curricular goals among students at a point in time and over time (e.g., grade-level curriculum)	Low	Low	High (individual)
2. Flexibility in the instruction (learning experiences)	Moderate	Moderate-High	High
3. Flexibility in the content standards chosen to be assessed for specific students (e.g., the standards used to guide the development of the specific grade-level assessment)	Low	Low – moderate	Low-high
4. Flexibility in the methods/items used to assess	Low	Low – moderate	Low-high
5. Flexibility in how the tests is administered including administration conditions	Low	Low – moderate	Moderate-High
6. Flexibility in the scoring	Low	Low	Low-high
7. Flexibility in the performance standards (evaluative criteria)	Low	Low – moderate	Low-Moderate
8. Flexibility in interpretation and reporting	Low – moderate	Low – high	Moderate – high
9. Flexibility in how handled for student accountability	Low	Low – high	High
10. Flexibility in how handled for school accountability	Low	Low – moderate	Low

Draft Technical Manual TOC

- **Section I—Overview, Background, and Key Components of the Validity Evaluation**
- **Section II—Test Development, Administration, Scoring, and Reporting**
- **Section III—Technical Criteria**
- **Section IV—Consequential aspects of the assessment system**
- **Section V—The Validity Evaluation**

Section I—Overview, Background, and Key Components of the Validity Evaluation

- Overview of the Assessment System
- What is the content?
- Who are the students?
- Introduction of the Validity Framework and Argument

Section II—Test Development, Administration, Scoring, and Reporting

- Test Development
- Administration & Training
- Scoring
- Reporting

Section III—Technical Criteria

- Alignment
- Item Analysis and DIF/bias
- Characterizing & quantifying error
  - Decision consistency and accuracy
- Scaling and Equating
- Standard Setting

Section IV—Consequential aspects of the assessment system

- Effects on students learning opportunities
- Effects on teacher professional growth
- Programmatic effects on schools and districts

Section V—The Validity Evaluation

- Revisiting the validity evaluation questions
- Synthesizing and weighing the various sources of evidence
- An overall judgment of the validity of the AA-AAS system

Responsibility

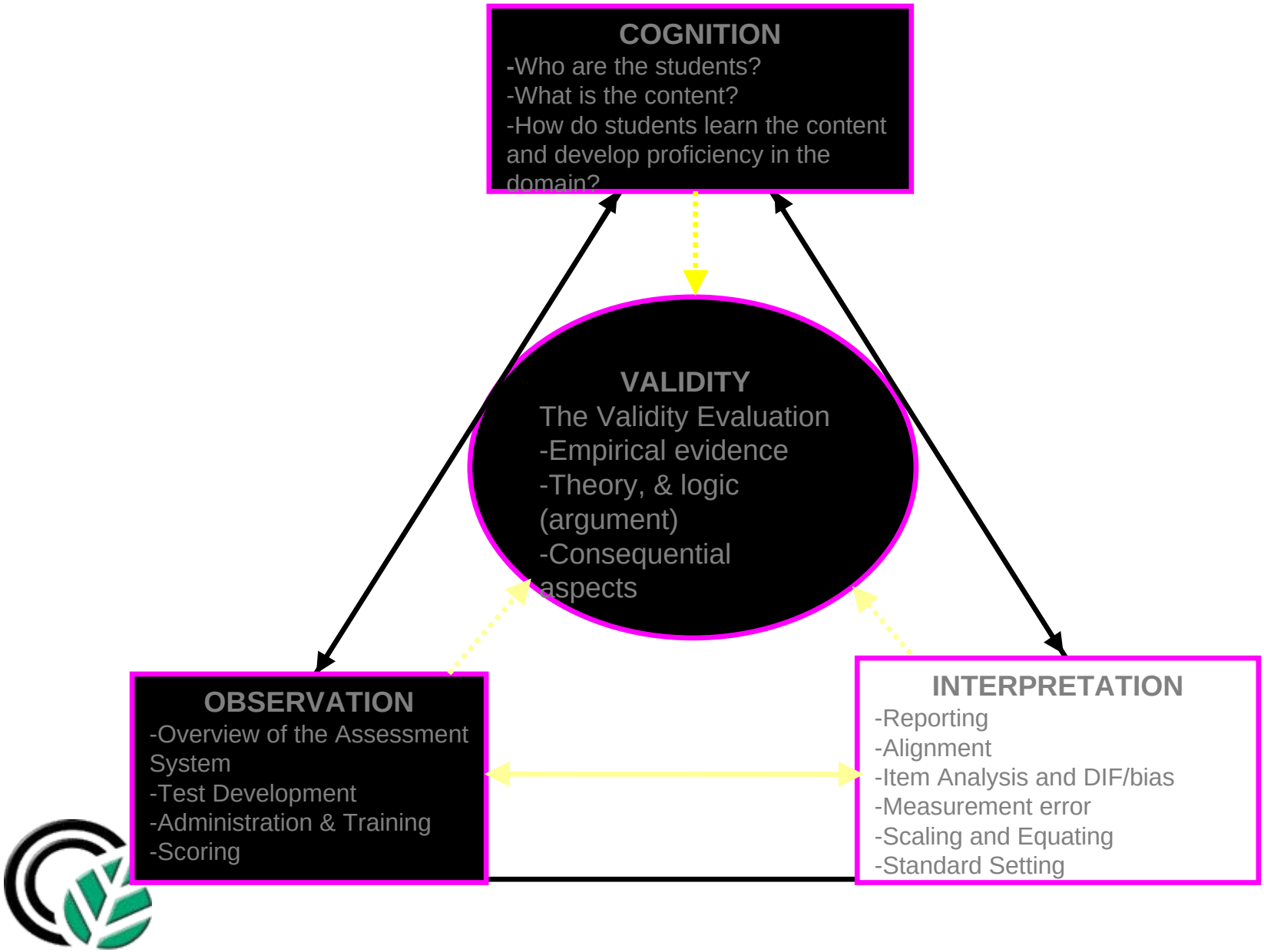
- The responsibility for collecting and analyzing these data does not rest solely with the contractor (Kevin loves it when I say that!)
- It should be a joint effort between the state, the contractor, and others (e.g., university partners)
- There is no expectation that the full manual be produced each year, but it is crucial that there be a plan for systematic data collection

The Triangle Revisted

- So how does this proposed design for technical document mesh with where we started with the assessment triangle?

- Funny you should ask…

Relationship between the Assessment Triangle & the Technical Manual TOC



Some hints for Section III

- **Alignment**
- **Item Analysis and DIF/bias**
- **Characterizing & quantifying error**
  - **Decision consistency and accuracy**
- **Scaling and Equating**
- **Standard Setting**

Alignment

- Browder & Flowers are leading this effort
- General assessments need to deal with 2-way alignment, but AA-AAS need to deal with at least 3-way alignment
  - "Indicators"-to-content standards
  - Items/tasks-to-indicators
  - Item/tasks-to-content standards
- All of these go in both directions

Item analysis/DIF

- Traditional methods of examining the statistical properties of items can be used, but we have to be careful of interpretations, e.g., discrimination parameters might only be separating functional levels of students
- Does DIF make sense? Perhaps, but what are the appropriate focal and referent groups?
- The bigger issue is item/test bias—we argue this can be evaluated only through carefully designed judgmental methods

Characterizing error

- This is one of the most challenging components to evaluate fairly
- Typical statistics such as interrater reliability are useful, but not nearly sufficient
- To quantify the error associated with AA-AAS we need to account for sources of error not usually done for general assessments, especially errors associated with administration.

Scaling and Equating

- Scaling and equating are focused on facilitating that similar inferences result from similar performances—i.e., comparability
- Scaling decisions are similar to those of general assessments, but need to attend to issues of multidimensionality and score composite considerations
- Equating challenges are reduced when the assessment does NOT change each year
  - However, maintaining the same assessments year after year does not preclude comparability concerns
  - When tasks change—due to flexibility or other reasons—formal equating is almost impossible. Must use judgmental or rubric-based methods to make judgments about comparability

A Few Ways to Establish “Comparability”

- Establish construct comparability based on similar content – for example, one assessment item taps the same construct as another assessment item. This may be based on a content and/or cognitive analysis.
- Establish comparability based on similar or compensatory functionality – distributional requirements often specify profiles of performance will be treated as comparable; total scores based on a compensatory system do similarly.
- Establish comparability based on judgments of relatedness or comparability – disciplined judgments may be made to compare almost anything in terms of specified criteria (e.g., is this bottle as good a holder of liquid as this glass is?). Decision-support tools and a common universe of discourse undergird such judgments.

Standard Setting

- You should have been here yesterday…
- See the presentations by the following people at the websites listed below:
- Marion-- www.nciea.org
- Kearns-- http://www.naacpartners.org/
- Quenemoen-- http://education.umn.edu/nceo/

Document and Defend

- Many of the methods we are suggesting are based on traditional methods that have been extended to work in this context
- However, we are not interested in banging square pegs into round holes
  - Approach this like good program evaluators—use the methods that best address the questions. If judgmental methods are the best you have, you must document your methods and defend your approach
- If this was easy, somebody would have done it already!

- Kevin will now talk about the trials and tribulations of having to do this for real before he had the benefit of the expert panel as well as how he might do things differently now that he’s had a chance to work with the expert panel