



Inclusive Innovations:

*Considering the Essential Characteristics of
Task Design in Determining the Technical
Defensibility for Students with Language and
other Challenges*

Rebecca Kopriva
Wisconsin Center for Education Research



The Problem

- Some students with normal intelligence can't show what they know using language as the primary communication vehicle.
- These include
 - lower English proficient ELs
 - some LDs
 - some others, such as DHH, ED, ADD, Speech, and some G & T
 - some poor readers who are not otherwise identified
- At the same time, these students should be exposed to challenging content.
- The conundrum is cognitively complex content is often intertwined with sophisticated language.
- However these students ARE still learning the complex content because they and their teachers have learned to convey meaning using other semiotic representations as their primary communication methods.



What Might This Type of Assessment Look Like?

- For many of these students using accommodations with traditional or most innovative tasks is not sufficient.
- Two aspects are essential:
 - conveying meaning to the student from the test maker
 - conveying meaning to the test maker from the student.
- One approach mitigates language by using deliberate computer-interactive, multi-semiotic strategies for
 - building up and communicating the problem environment and target questions
 - designing response spaces that allow students to successfully demonstrate their knowledge and skills using methods outside the strictures of traditional item types.



Defining an Approach

- In order to defensibly integrate these types of tasks into general assessment systems at all levels, evidence about what these tasks measure can be defined by a framework spelling out what meaning is intended and how the tasks intend to convey it.
- While the framework would fit within a more general system's approach, it is needed because the approach to communicating meaning is distinct. As such, it provides the basis for a content validation crosswalk to other assessment elements in the system.
- Adequate documentation in each section of the framework also provides the basis for interpreting other technical evidence.

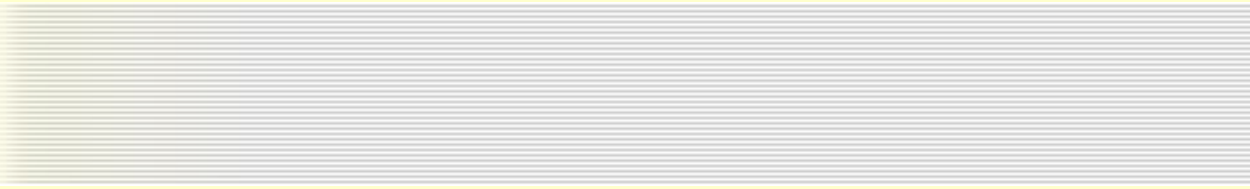


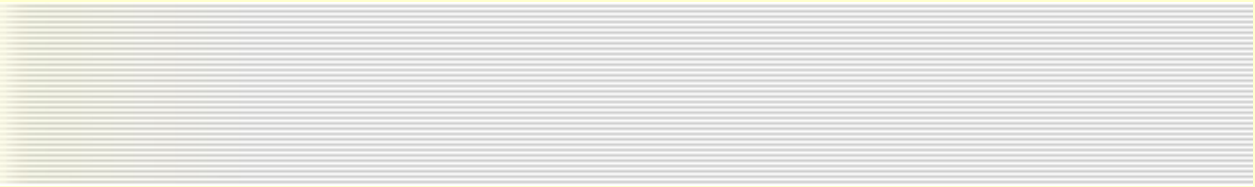
ONPAR Framework

- The ONPAR framework is divided into six sections, each responsible for defining evidence associated with it's scope:
 - Specifying the Intended Claim(s) of the Task
 - Specifying the Architectural Loads of the Task
 - Defining the Context Environment
 - Constructing the Problem Environment
 - Defining the Target Question(s) or Statement(s)
 - Specifying the Response Environment(s)
- Each environment focuses on activating particular cognitive processes to convey meaning using relevant effective representations at specified junctures.
- By designing tasks where meaning is explicitly and strategically introduced, supported, and distributed, complex concepts and skills can be assessed and conveyed to and from the test taker with little language.



- **Demonstration slide**







What Kind of Comparability Evidence is Needed?

Essential Components:

1. What do we want when we want score comparability?
 - A) Identification of targeted inferential construct claims, at the test level and at the item or task level
2. What do we mean when we say comparability?
 - B) Identification of grain size
3. How can we evaluate comparability (in other words, what evidence do we need)?
 - Construct equivalence geared to (A)
 - Score equivalence geared to (B)



Construct Equivalence

Construct equivalence: Evidence of similar *meaning*.

- This may be demonstrated by:
 - similar internal structure
 - same standards coverage
 - similar criteria for inclusion (as in portfolios)
 - similar judgments about relevant cognitive demands



Construct Equivalence

- In addressing construct equivalence about relevant cognitive demands the issue is particularly salient when
 - the variation is more divergent from the general test
 - test methods within the system differ in the amount/quality of information they are obtaining about targeted constructs. In this case a decision needs to be made about if the general test is the reference instrument, and then how should a different quality of information be handled across versions?
 - changes in kinds of information collected occurs for all students over years. In this case how are the test score inferences over years impacted?



Score Equivalence

Score equivalence: Evidence of scores '*behaving the same way*' for students with similar abilities.

- At the test level this may be demonstrated by:
 - similar proficiency percentages
 - similar score distributions
 - similar rank order
- At the item level this may be demonstrated by
 - similar distractor distributions (considering content ability)
 - similar DIF*
 - similar p-values



Given What We Know Today

Five recommendations about what can be done right now:

1. Operationalize the ECD principles:
 - a. Specify inferential claims first, and making sure these claims guide *all* format, form, and item development decisions in the system.
 - b. Identify the range of students (and perhaps testing situations). For some students, when the score inferences are questionable, identify defined profiles and *specifically* what they need to demonstrate their knowledge and skills about the targeted content.
 - c. Build a comprehensive system that concurrently plans for and considers design and evidence elements of all variations as well as the general test.



Given What We Know Today

2. Make explicit and put in place what needs to be standardized in variations during development, implementation, and scoring (e.g. Rigney & Pettit, 1996; Barton, 2008).
3. Test for construct equivalence at the item/task levels as well as at the total score level using controls with similar content abilities. Bootstrapping is recommended as needed for defensibility.



Given What We Know Today

4. Define suitable variables (e.g. which cognitive demands are the most salient) and relationships (e.g. conjunctive, compensatory) of construct and/or score equivalence evidence when the “this or better” factor is operating.
5. Identify the proper levels for determining equivalence: items, task ‘clumps’ or types of portfolio entries, etc. Produce defensible crosswalks between general test and variation when levels are different.



Looking Ahead

Two research/development recommendations for the near future:

1. Unpack the cognitive demands and required evidence in interim benchmark tests meant to support learning (as compared to verifying it).
 - We need to understand how to design assessments to support information about what comes next in learning for individual students.
 - We need to identify suitable types of evidence for documenting the ongoing integrity of the constructs within the environment of differential development and diverse student challenges.



Looking Ahead

2. Taking advantage of the interactive computer capabilities in short and extended items for students who cannot *demonstrate* what they know on the general test.

For both of these approaches, comparability needs to be understood when variations reflect distinct differences from the general test in

- *Directness* to the latent construct underlying the content target
- *Response* opportunities
- *Density* of the cognitive demands (target relevant and irrelevant)
- *How* target cognitive schemas are engaged