



Overview of ONPAR Test Development Activities

Jim Bauman
Center for Applied Linguistics
Presentation: ISBE April 18, 2008

The Problem Today for Students with Language Challenges: Case 1

- Operational statewide test
- Evaluated the distractor distributions of select multiple choice items
- Item writers well versed in item writing principles of universal design
- Analysis compared scores
 - Prefunctional, Beginner, Intermediate (PBI) ELLs,
 - LD students,
 - Hearing Impaired students
 - Native English speakers without IEPs as controls
- Question: Are the distributions for the ELL, LD, and Hearing Impaired students generally the same as those of the control students?

Case 1: Analysis of Operational Test Data: Treatment Groups vs. Controls

■ Balanced sample distractor analyses for 58 items

Subject	Grades 3-5				Grades 6-8			
	# Items	% Problems*			# Items	% Problems*		
		PBI ELL	LD	Hear Imp		PBI ELL	LD	Hear Imp
ELA	6	33	33	50	4	100	25	50
Math	6	67	50	33	16	63	69	63
Science	8	63	50	63	7	86	71	71
Social Studies	4	100	50	50	7	57	29	29
Mean %		63%	46%	50%		71%	56%	56%

* Chi-square results for treatment group are significantly different from those of control students

Problems with Multiple Choice Items: Case 2

- Design
 1. A large-scale mathematics test with both multiple choice and constructed response items
 2. Tested ELLs at 3 levels of proficiency: Beginning, Intermediate, and Advanced and at 2 grades: 3 & 5
 3. Controls were Exited ELLs and Native English speakers
 4. Investigated the relationship between scores of ELL and scores of controls
- Desired outcome: no significant differences between each ELL group and the controls
- Criterion validity study included teacher ratings of student's skills in specific mathematics elements covered by the test
- Analyses focused on regression results (B and R squared)

Question: How do the validity of the score inferences for ELL students measure up to those of non-ELLs?

Case 2: Analysis of Criterion Validity Study

Multiple Choice

Constructed Response

		B	p	R ²
Grade 3	Beg	.41	.29	.02
	Int	.87	.00	.05
	Adv	1.11	.07	.05
	Exit	2.73	.00	.20
	Nat E	2.67	.00	.27

		B	p	R ²
Grade 3	Beg	1.41	.00	.21*
	Int	1.15	.00	.06
	Adv	1.74	.02	.08
	Exit	3.68	.00	.20
	Nat E	3.57	.00	.23

		B	p	R ²
Grade 5	Beg	-.32	.50	.01
	Int	.78	.01	.04
	Adv	1.61	.01	.13*
	Exit	1.87	.00	.19
	Nat E	2.25	.00	.23

		B	p	R ²
Grade 5	Beg	1.32	.02	.13*
	Int	1.58	.00	.06
	Adv	2.91	.00	.20*
	Exit	3.43	.00	.26
	Nat E	3.20	.00	.22

* NO significant difference with exited and native English speaking students

- Movement to dynamic computer-based testing
- New item formats possible with online technology
- More accessible to ELLs
- Possibly more accessible to students with language or literacy challenges

Four Central Issues for Dynamic Computer-Based Testing

- Relevance to mainstream testing
- Provision of testing accommodations for ELLs and students with literacy needs
- Comparability of paper-and-pencil and dynamic testing systems
- Level of reduction in test item language
 - What construct irrelevant information is lost
 - What construct irrelevant information is substituted

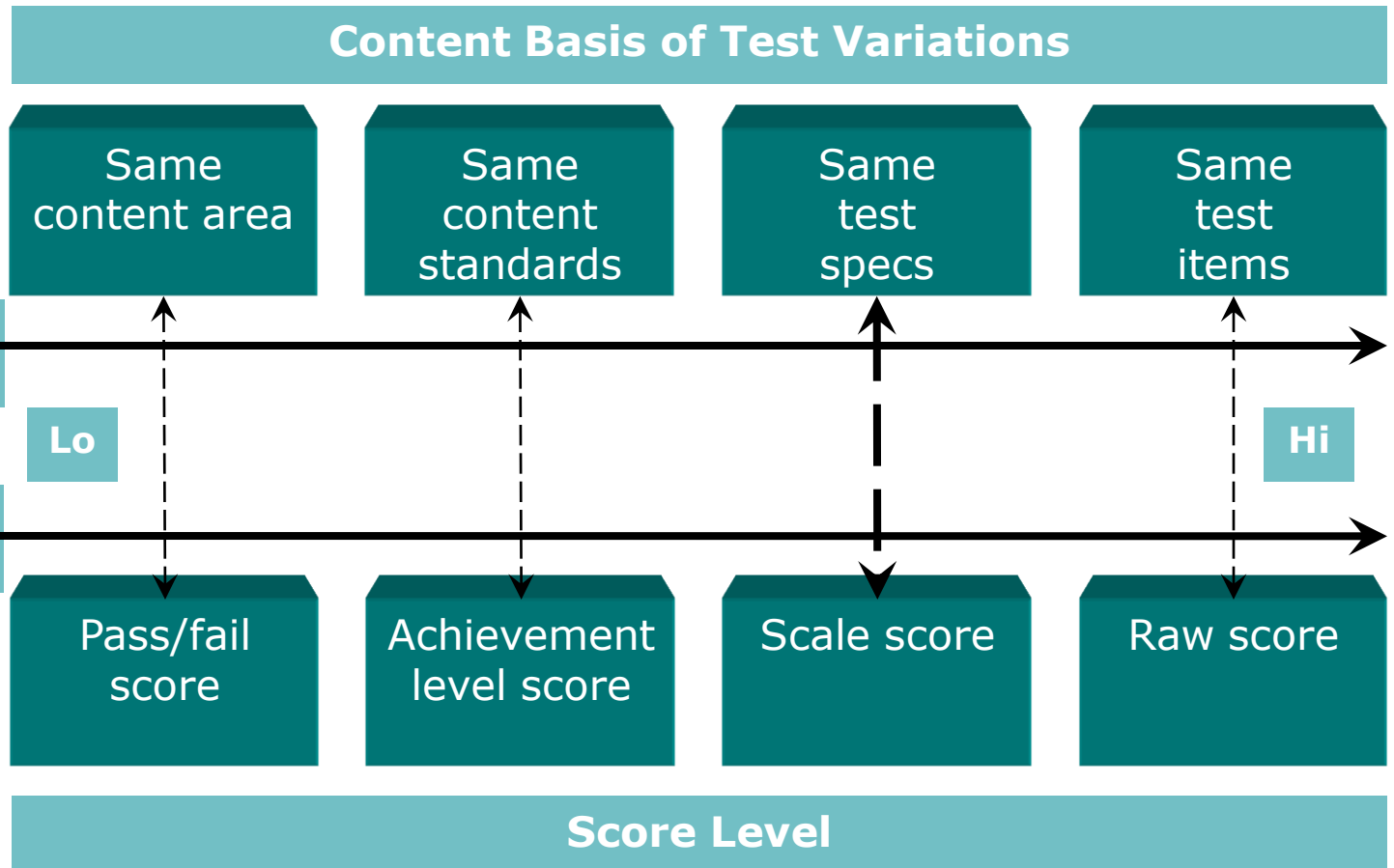
- The goal is to build and investigate the feasibility of prototype items for low English proficient students in two subject areas, science and mathematics.
- The science ONPAR items will be based on NECAP multiple choice and constructed response items.
- The math items will be based on released items of standard tests congruent with standards in common across a wide variety of states.
- The intent of the projects is to measure content at the full range of cognitive complexity consistent with what is currently being measured on today's large-scale tests.

1. What are the item elements and procedures that should underpin defensible dynamic and interactive large-scale testing in general?
2. Can we use the dynamic capabilities of the computer to develop technically defensible large-scale test cores in science and mathematics for low English proficient ELLs?
3. Because of the nature of the language challenges for low English proficient ELLs, expanded item types will be investigated. What elements associated with the expanded item types need to be addressed?
4. How does the data collected from these types of items relate to static data collected from other students? What kinds of comparability evidence is necessary to defend the scores from dynamic testing for these students?

- Use computer capabilities to:
 1. Radically reduce the language load in the items.
 2. Provide compensatory avenues in which to present contextual information and target relevant details.
 3. Provide additional avenues for students to demonstrate what they know.

- Project approach
 1. For standard items, develop ONPAR item versions that measure the same content at the same level of cognitive complexity.
 2. Build items that are comparable at the scale score level.

Continuum of Comparability



- Item approaches use computer functionalities for both receptive and expressive purposes:
 - Receptive and Expressive
 - Point and click
 - Drag and drop
 - Stimulus manipulation (using menu or slider bar)
 - Stimulus rearrangement and assembly
 - Animation (using menu or slider bar)
 - Expressive only
 - Drawing freehand or from toolbox
 - Keyboarding
 - Receptive only
 - Roll-overs presenting audio or written glosses and translations, illustrations, highlighting or framing
 - Video display

ONPAR Item Types Identified to Date

Traditional Item Types

- Multiple Choice*
- Free/Open Response

Complex Computer Dynamic Item Types

- Select and Classify
- Matching and Connecting
- Ordering Stimuli
- Stimulus Manipulation: Simulation of context and target stimuli to present item requirements; students manipulate stimuli to address item requirements.
- Statement Construction (Simulated Constructed Response): Students construct syntactically ordered statements using visual and linguistic elements; statements may include Boolean or conditional logic.
- Representational Modeling: Modeling using toolkit elements; incorporates assembling and drawing.
- Graphical Modeling: Modeling of quantitative relations by manipulating a graphic representation.

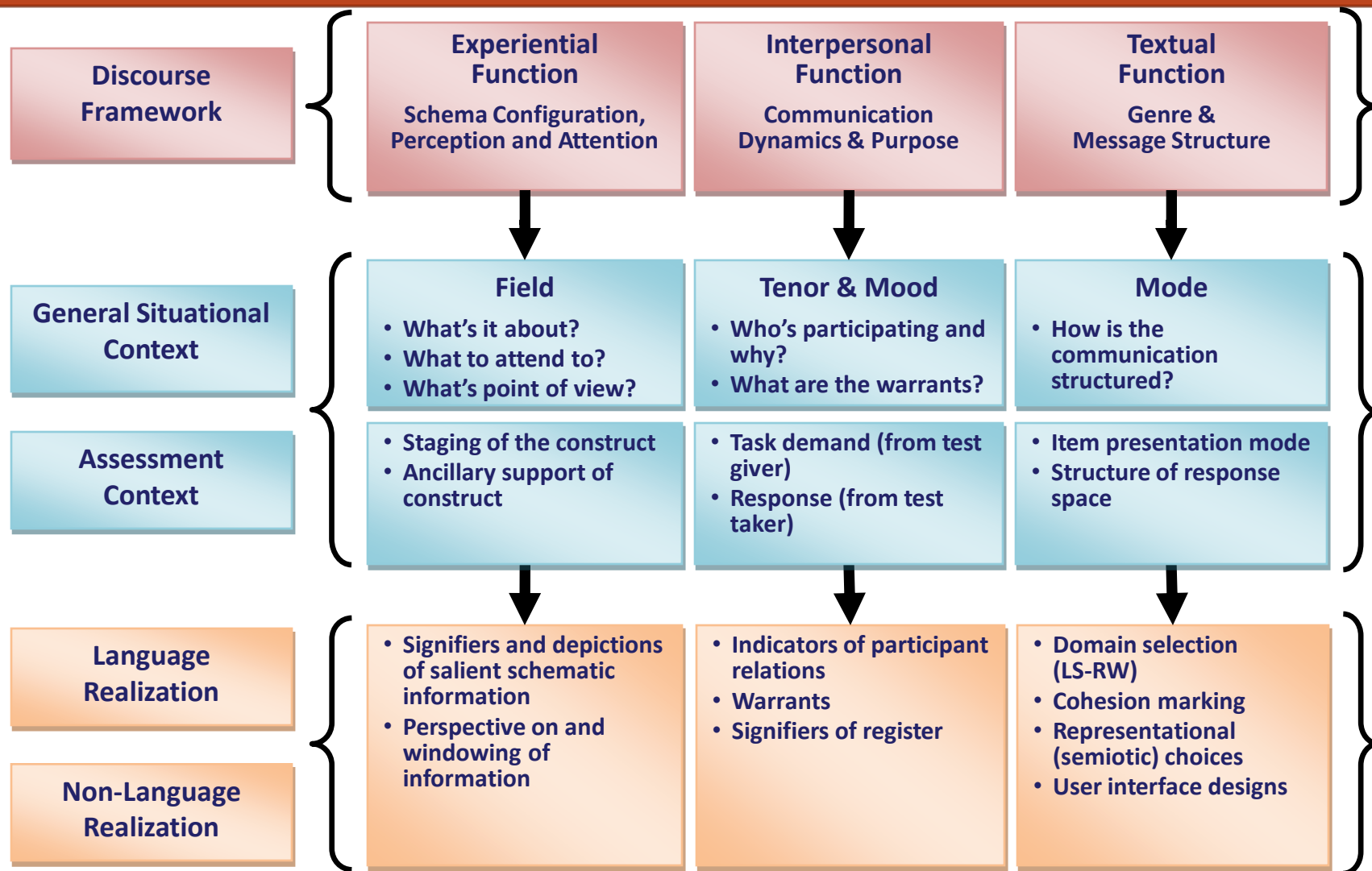
Composite Item Types (multi-tasked items with tasks representing different item types)

- Thematic Probing: Loosely integrated tasks each leading to a discrete result
- Problem Solving Vignette: Tightly integrated, temporally sequenced tasks converging to a single result

Research Agenda: Science Grant & Math Grant

1. Cognitive labs with:
 - ELLs at different levels of proficiency
 - Native English speakers
2. Independent judgment review of cognitive demands in both ONPAR and standard items to determine:
 - Science and math targets
 - Other cognitive demands in items that are non-essential to the science and math target constructs
3. Controlled studies with science items using standard and ONPAR items
4. Large scale study with mathematics items investigating the convergent and discriminant validity of items vs. other salient indicators.

Linguistic Framework for Describing Active Elements



Controlled Trials on Science Items Study Questions

- How do students interact with ONPAR items as compared to traditional test items?
 - Test versions (NECAP and ONPAR) built to be equivalent at the scale score level
 - Three groups of students:
 - Low English proficient students
 - Other ELLs
 - Native English speakers
 - Test forms randomized over students
- How low can you go? How little language can be in items and still have task demands clear to the students?
 - Low language version with translation versus
 - Very low language version with no translation
 - Two versions of ONPAR items randomized over students

- Keep language constructions simple
 - Avoid phrasal and sentence level syntax
 - Avoid morphologically complex words
- Avoid all L1 translations
 - Translation can have unpredictable or difficult to control effects on cognitive complexity and item difficulty
 - Operationally unwieldy to maintain multiple L1 translations
- Make the task demand inferentially clear
 - Set the task demand with a distinctive command icon
 - Provide ancillary supports to make inferences clear
- Keep information redundancy low
 - Redundant components, particularly if not comprehensible, may be distracting, rather than facilitating
 - Redundancy may increase burden on working memory and cognitive load