



Building Comparable Computer-Based Science Items for English Learners:

Results and Insights from the ONPAR Project

**National Conference on Student Assessment
Los Angeles, CA
June 23, 2009**

Rebecca Kopriva, University of Wisconsin
Therese Gleason Carr, Center for Applied Linguistics



The ONPAR Studies:

Addressing the Needs of Low English Proficient Students and Others Using Novel Item Types Suitable for Large-Scale Science and Mathematics Testing

Rebecca Kopriva
Principal Investigator
University of Wisconsin

Tim Boals
Project Director
WIDA, University of
Wisconsin

Jim Bauman
Program Manager
Center for Applied
Linguistics



The Problem: ELLs and Controls

Balanced sample distractor analyses for 58 items,
(PBI) ELLs and controls

	Subject	# items	%problem*		# items	%problem
	ELA	6	33		4	100
3-5	Math	6	67	6-8	16	63
	Science	8	63		7	86
	SS	4	100		7	57
	Mean %		63%			71%

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



The Problem: LDs and Controls

Balanced sample distractor analyses for 58 items,
LD students and controls

	Subject	# items	%problem*		# items	%problem
	ELA	6	33		4	25
3-5	Math	6	50	6-8	16	69
	Science	8	50		7	71
	SS	4	50		7	29
	Mean %		46%			56%

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



The Problem: Deaf/HH and Controls

Balanced sample distractor analyses for 58 items,
Deaf and hard of hearing students and controls

Subject	# items	%problem*	# items	%problem
ELA	6	50	4	50
3-5 Math	6	33	6-8 16	63
Science	8	63	7	71
SS	4	50	7	29
Mean %		50%		56%

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



The Problem: Criterion Validity Study

		<i>Multiple Choice</i>			<i>Constructed Response</i>		
		B	p	Rsq	B	p	Rsq
3	Beg	.41	.29	.02	1.41	.00	.21*
	Int	.87	.00	.05	1.15	.00	.06
	Adv	1.11	.07	.05	1.74	.02	.08
	Exit	2.73	.00	.20	3.68	.00	.20
	Nat E	2.67	.00	.27	3.57	.00	.23
5	Beg	-.32	.50	.01	1.32	.02	.13*
	Int	.78	.01	.04	1.58	.00	.06
	Adv	1.61	.01	.13*	2.91	.00	.20*
	Exit	1.87	.00	.19	3.43	.00	.26
	Nat E	2.25	.00	.23	3.20	.00	.22

* NO significant difference with exited and native English speaking students



Science and Mathematics ONPAR Projects

- The goal of the projects is to build large-scale operational tests in science and mathematics that can measure what our students know.
- The goal of the research is to build and investigate the feasibility of prototype items in two subject areas, science and mathematics, which would be appropriate for students with language challenges, especially low English proficient students, and possibly some students with disabilities.
- The research will investigate the prototype items relative to traditional multiple choice and constructed response items that measure the same content at the same level of cognitive complexity.



General Project Questions

- Can we use the dynamic capabilities of the computer to substantially minimize the required language and develop defensible large-scale test cores in science and mathematics for low English proficient ELLs and others with language challenges?
- What are the item elements and procedures that should underpin defensible dynamic and interactive large-scale testing in general?
- Because of the nature of the language challenges, novel item types will be investigated. What elements associated with these new kinds of items need to be addressed?
- How do the focal groups of students respond to the prototype items relative to the control groups, and relative to how they perform on the traditional items?
- Are there ways that the project has extended how all students might be able to demonstrate knowledge and skills beyond what is done in large-scale assessment to date?



Additional Science Questions

- Two different ONPAR item versions were built at fourth grade and eighth grade: a low language (LL) and a very low language (VL) version.
 - Four groups of students were tested: two focal, one control and one exploratory group
 - ELLs with English proficiency at levels 1 and 2 (focal)
 - ELLs at level 3 (focal)
 - Non-ELLs (control)
 - ELLs at levels greater than level 3 (exploratory)
1. When controlled for ability, how does the performance of each group on the LL and VL forms compare to performance on the traditional form?
 2. What item characteristics appear to be effective and not effective in measuring the targeted content?



Additional Mathematics Questions

- One ONPAR and one traditional form were developed at grade 4 & 7.
 - Four groups of students will be tested: two focal, one control and two exploratory groups
 - ELLs with low English proficiency (focal)
 - Students with learning disabilities and other SD students with language challenges (focal)
 - Non-ELLs (control)
 - Non-ELLs with low reading proficiency (exploratory)
1. When controlled for ability, how do the performances of each of the focal groups on the ONPAR forms compare to performances of the control group on both forms?
 2. What are the factor structures of the ONPAR and traditional forms for the focal and control groups?
 3. How do state reading and mathematics scores compare to results?



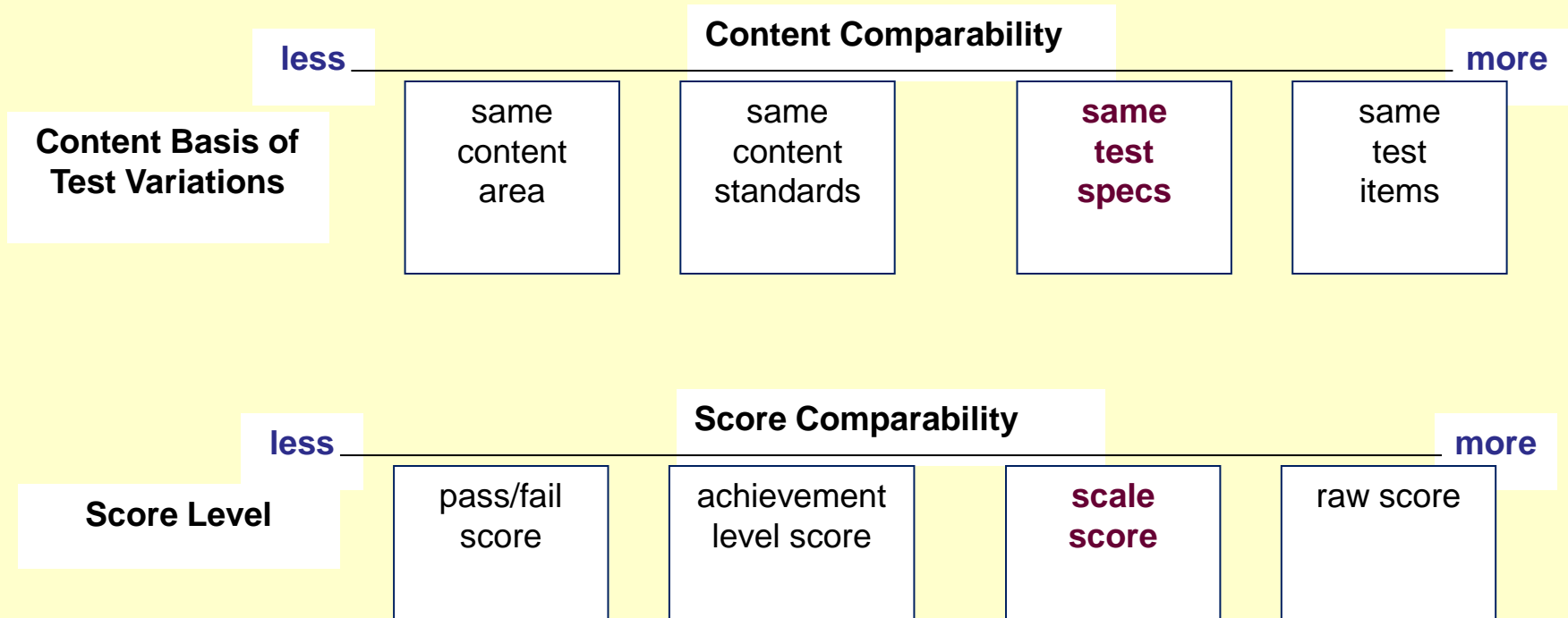
Underpinnings: Use of Evidenced Centered Design

Used Mislevy's Evidence Centered Design principles:

- Began by establishing an inference claim about what each item is supposed to measure
- Developed items with elements known to be accessible for focal and control groups
- Used standardized development and review procedures
- Tested items in iterative series of cognitive labs



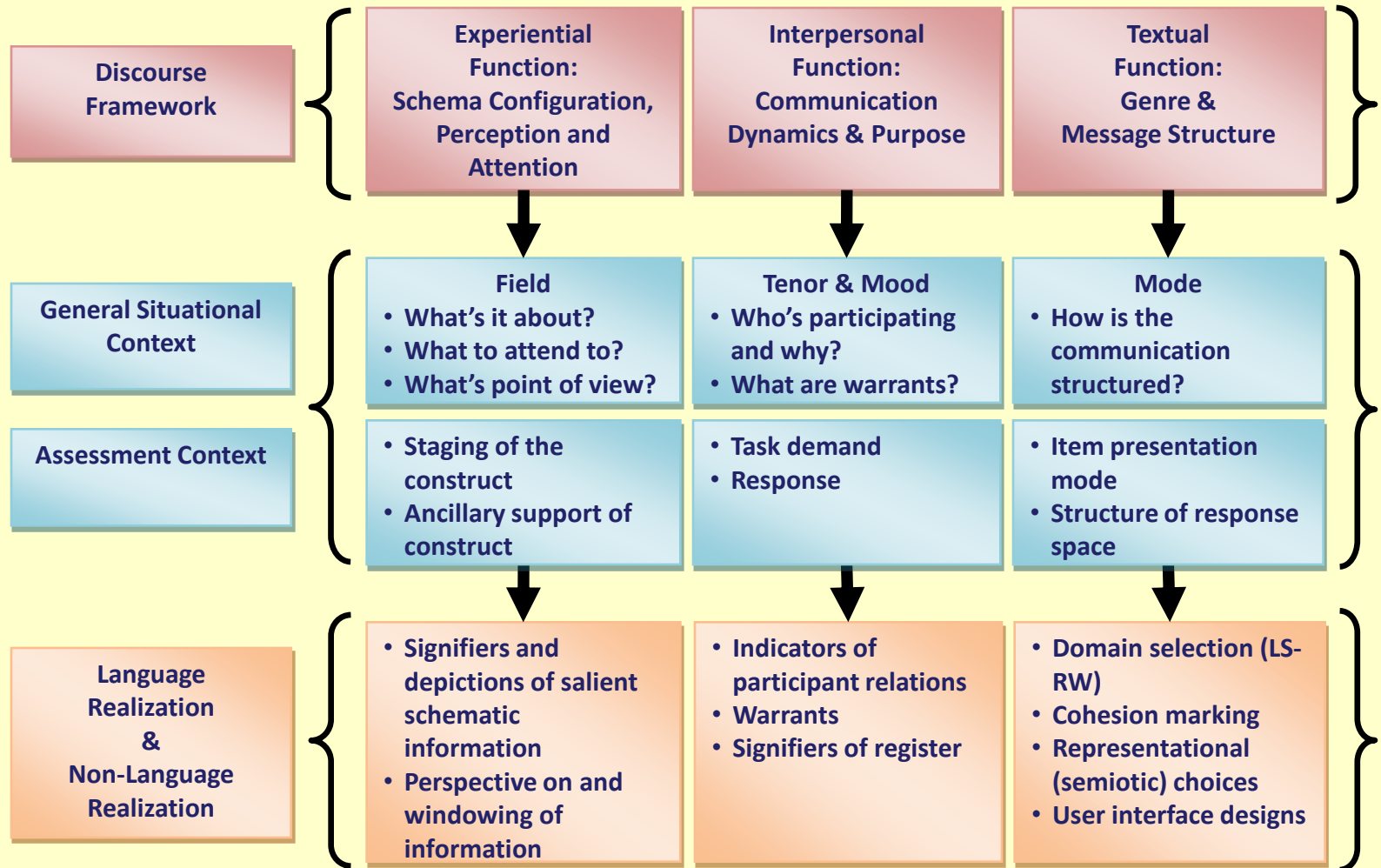
Underpinnings: *Comparability Focus



* Graphic developed by Phoebe Winter



Underpinnings: Linguistic Framework





Interactive Functions of ONPAR Items

Several types of interactions allow students to demonstrate their knowledge and skills. Among these are:

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



Standardized Aspects of ONPAR Items: Unrelated to the Assessment Target

- Placement of item requirement statements, contextual real estate and function icons
- Language and non-language element communication strategies, functions and amounts
- Timing, length and presentation approach of primary and secondary contextual elements
- Identification of interactive response avenues
- Animated procedural icons
- Navigation buttons and rollover explanations

The screenshot displays an interactive assessment interface. At the top, a question asks: "What will happen to the water level?" with a speaker icon on the left. Below the question is a green button with a hand icon and the text "Estimate". The main area contains three beakers, each with a sphere above it. The first two beakers have a grey sphere, and the third has a brown sphere. Each beaker has a dashed horizontal line representing the water level, with a vertical double-headed arrow and a question mark next to it, indicating the water level is to be estimated. At the bottom, there is a navigation bar with buttons for back, refresh, GO, forward, and checkmark. The text "Question: 1 of 1" is visible in the bottom right corner.



Items Presentation



Results To Date

- Cognitive labs in science and in mathematics have yielded important item development information related to
 - the dynamic interface,
 - the efficacy of different interactive item types, and
 - how stimuli are understood by the students.
- An independent judgment panel of experts found that
 - most of the science ONPAR items were measuring the same content as their traditional counterparts, and
 - the overall cognitive complexity of the target science was similar.



Science Controlled Trials

- Approximately 500 students tested in each of two grades (4/5 and 8/9) in 6 districts.
- Per student, teacher ratings of science ability (by objective) were also collected.
- Preliminary results indicate that, when controlled for ability, low English proficient ELLs did not score significantly different from non-ELLs on the ONPAR tests, while there were significant differences on the traditional tests. This occurred for both grades.
- Overall, the LL ONPAR items worked better; however in some cases, the VL were more effective. Additional analyses are underway.



ANCOVA Results with Ability as Covariate: 4th Grade

GROUP/FORM Dyads		ELLGroup =Low	ELLGroup =NonELL
		Trad	Trad
ELLGroup =Low	ONPAR_LL	NS p=.37	Sig p=.03
	ONPAR_VL	NS p=.19	NS p=.13
ELLGroup =NonELL	ONPAR_LL		NS p=.47
	ONPAR_VL		NS p=.96

When controlling for ability, Low ELL group performs significantly lower than Non ELL group on traditional form, but performs similarly to the Non ELL group on the ONPAR LL form.

Substantial increase in scores for low ELL group on ONPAR LL form vs. the traditional form but not statistically significant. This was probably due to small sample sizes: N= ?? ONPAR; N= ?? traditional.

Non ELL group performs similarly on both forms (no significant difference). N= ?? ONPAR; N = ?? traditional.

Still a significant difference between non-ELL on traditional and low ELL on ONPAR.

Overall, findings support hypothesis that ONPAR LL form is better measure of science ability for Low ELL group



ANCOVA results with Ability as Covariate: 8th Grade

GROUP/FORM Dyads		ELLGroup =Low	ELLGroup =NonELL
		Trad	Trad
ELLGroup =Low	ONPAR_LL	Sig p=.05	Sig p<.01
	ONPAR_VL	NS p=.30	NS p<.01
ELLGroup =NonELL	ONPAR_LL		NS p=.09
	ONPAR_VL		NS p=.73

When controlling for ability, Low ELL group performs significantly lower than Non ELL group on traditional form, but performs similarly to the Non ELL group on the ONPAR LL form.

Non ELL group performs similarly on both forms (no significant differences). N= ?? ONPAR; N= ?? traditional.

??Statistically significant increase in scores for low ELL group on ONPAR LL form vs. ?? N= ?? ONPAR; N= ?? traditional.

Still a significant difference between non-ELL on traditional and low ELL on ONPAR.

Overall, findings support hypothesis that ONPAR LL form is better measure of science ability for Low ELL group



Discourse Analysis Results

- An item level measure of discourse coherence was conceived to inform the issue of item accessibility
 - The measure was drawn from ten indicators, gauging level of lexical support, syntactic cohesion, consistency of discourse marking, and inferencing and evidential load.
 - A composite measure was constructed for each item in its traditional and ONPAR forms.
- Composite measure is predictive of misperforming items
- For performing items, composite measure is correlated with item performance on ONPAR items, though not on traditional items



Science Operational Plans

- A field test is scheduled in winter/spring of 2010 for ONPAR science tests at the elementary and middle school grade clusters: 4-5 and 7-8.
- The field test will focus on building an operational core of approximately 30 items. It is expected that this core will be augmented by states as part of their statewide testing system in science for students with language challenges.
- The field test will make use of a Web based test delivery platform developed and deployed by WIDA
 - Participating districts will use their own computer infrastructure to support the testing
 - Districts will be reimbursed for their participation and results will be shared



Implications for Large-Scale Testing

- The projects to-date have developed novel ways of measuring what students know, especially in how students demonstrate their knowledge and skills in ways other than through traditional multiple choice or constructed response.
- While the dynamic items seem to be measuring the same content at the same overall cognitive complexity as their static counterparts, there are clearly differences. These include
 - Directness to the latent construct underlying the content target
 - Density of the cognitive demands on test takers
 - Construct relevant: Scope and range of construct that item engages
 - Construct irrelevant: Besides language proficiency, includes level of cultural awareness presumed by visual representations, level of computer facility, and range of item types on test.
 - How target cognitive schemas are engaged
- What are the implications when the probability of correct response changes dramatically at the item level and does this matter at the test level?



Implications for Formative and Benchmark Testing

- If properly designed, many the dynamic items and tasks can evaluate the sophistication of conceptual understanding and procedural strategies by tracking how students interact with stimuli and move across screens.
- Patterns of conceptual sophistication and procedures can be confirmed over items using different content.
- This approach can provide incremental formative and summative information to students and teachers.



For More Information:

Rebecca Kopriva rkopriva@wisc.edu

Therese Carr tcarr@cal.org

Jim Bauman jbauman@cal.org