Rebecca Kopriva, University of Wisconsin
David Gabel, Center for Applied Linguistics
Cathy Cameron, Center for Applied Linguistics

The first ONPAR project, funded in 2006 through U.S. Department of Education's Enhanced Assessment Grant program, focused on developing and investigating a novel, dynamic, multi-semiotic approach for testing the science achievement of fourth- and eighth-grade low- English ELs. Besides testing out the viability of selected techniques, the research was primarily interested in identifying a method that would be useful for low English proficient students who often do not have the language to show what they know on traditional tests. As such, the project sought to answer these questions:

- How does the performance of low-English ELs in each grade compare to performance of control non-ELs on test of prototype ONPAR assessment tasks in science, and how do the performance of the two groups compare on tests of traditional items designed to measure the same science content at the same levels of cognitive complexity?

- How does the difference in performance of low-English ELs on the prototype ONPAR test and the traditional test in science in each grade compare with that of their non-EL peers?

Additionally, for each grade, the project was interested in understanding how the performance of mid-English ELs and high-English ELs compare to the performance of low-English ELs and non-ELs on both tests. The project was also interested in item by item comparisons.

ONPAR items were built from released state items and checked for comparability in content and cognitive demands by an independent group of science and assessment experts.

The project asked participating teachers to rate their students achievement in the classroom on the same objectives covered by the traditional and ONPAR items. This rating was used as an independent measure of science achievement in order to control for variation in science ability in each group.

Cognitive Labs

The first year and a half of the grant was spent developing draft tasks and testing them in a series of cognitive labs conducted in three school districts with 58 students. Results from these labs were used to refine the ONPAR approach. Additionally, the findings indicated that once functional help icons were installed in the items and a 10-minute interactive tutorial was built to precede testing, all students, even new arrivals, had no problem accessing the varied methods in different tasks, even when item types were novel and varied across screens. Think-alouds also found that ELs at various levels of English proficiency and non-ELs were generally able to access the range of ONPAR presentation and response techniques used in this study.

Experimental Study

An experimental study with equivalent groups randomly assigned fourth- and eighth-grade students to one of three science test forms: (a) a *traditional form* consisting of existing state items from which the ONPAR items had been built; (b) a *low language form* (*ONPAR_LL*) that used complete sentences, onscreen support, and audio translation of text in the student's native language (Spanish and Korean for this study); and (c) a *very low language form* (*ONPAR_VL*) that used single keywords or short phrases, without translations or onscreen support. Both ONPAR versions used dynamic simulated contexts and asked students to demonstrate what they knew by manipulating stimuli in novel ways. At each grade, the three

forms consisted of 11 items. The traditional fourth-grade form included 10 multiple-choice and 1 constructed-response items, while the traditional eighth-grade form included 9 multiple-choice and 2 constructed-response items. Students were assigned to one of four groups for testing based on their level of English language proficiency as measured by the WIDA-developed ACCESS for ELLs® assessment: (a) *low- proficiency ELs* (Level 1 or 2 out of 5), (b) *mid-proficiency ELs* (Level 3), (c) *high-proficiency ELs* (Level 4 or 5), and (d) *non-EL controls*. A total of 513 fourth-grade and 468 eighth-grade students from 26 schools in five states took part in the study (Table 2).

Table 2. *Sample by Form*

Test type	Grade 4 *n*					Grade 8 *n*				
	Low	Mid	High	Non-EL	Total	Low	Mid	High	Non-EL	Total
Trad	19	19	28	78	144	54	37	15	41	147
ONPAR_LL	21	28	30	109	188	55	35	16	45	151
ONPAR_VL	19	27	40	95	181	67	42	17	44	170

Fourth- and eighth-grade interactive tutorials were developed to orient students to ONPAR. Students could take the tutorial in either English or their first language. Teachers were asked to complete a questionnaire providing demographic information and a 3-point rating of each student's science knowledge for each of the eight constructs covered on the tests. Adapted from the method used by Schmidt et al. (2001), this science rating was used to define an ability index, which was used, in turn, as a covariate in the analyses. The eighth-grade forms were administered in the spring and fall of 2008 to eighth- and ninth-grade students, respectively; fifth-grade students took the fourth-grade forms in fall 2008. Testing periods, including the tutorial preceding the items, lasted about 45 minutes.

During the randomized trials, the ONPAR_VL items seemed to cause frustration with

non-ELs, and results were consistently depressed over the two grades for all students. Based on these results, the ONPAR_LL form was adopted as the standard, and analyses compared data from the ONPAR_LL form and the traditional form only. At this time, the Rasch model for the dichotomous response items, and one of its polytomous extensions, the Partial Credit Model (PCM), for the non-dichotomous items, were used to calibrate, equate, and scale the science ONPAR_LL and traditional test forms. To obtain parameter estimates, the Rasch or PCM models were calibrated for each test form at a grade level. The parameter estimates for each form were placed on a common metric by fixing the person ability measures to a mean of zero, and subsequently, a linear transformation of the scores was used, fixing the mean at 500 and the standard deviation at 100. Scale score means by group from the traditional and ONPAR forms were computed, and distributions of scores by group and form can be found in Table 3.

Table 3: *Scaled Test Score Descriptive Data*

Gr	Test	Low EL			Mid EL			High EL			Non-EL		
		M	*SE*	*n*	*M*	*SE*	*n*	*M*	*SE*	*n*	*M*	*SE*	*n*
4	Trad	424	20	19	460	20	21	506	17	28	527	11	79
	ONPAR	451	17	21	465	18	29	501	19	31	519	10	110
8	Trad	445	12	56	508	14	38	529	14	16	557	15	43
	ONPAR	484	12	58	487	16	36	518	21	16	524	17	47

To address the research questions, four a priori contrasts per grade were analyzed using analysis of covariance (ANCOVA). With the science ability rating as the covariate in each case, two of the contrasts compared the two focal groups (low-English EL vs. control non-EL) on each form, and two contrasts compared the two groups within each form (Table 4).

Table 4. *A Priori ANCOVA Contrasts*

Gr size		Source of variation	DF	MS	F	p-value	Effect
4	W/in forms	Trad low vs. trad non-EL	1	91440	10.46	<.001	1.84
		Error	498	8744			
		ONPAR low vs. ONPAR non-EL	1	13546	1.55	.21	0.71
		error	498	8744			
	B/w forms	Trad low vs. ONPAR low	1	613	0.07	.79	0.34
		Error	498	8744			
		Trad non-EL vs. ONPAR non-EL	1	17612	2.01	.16	-.10
		Error	498	8744			
8	W/in forms	Trad low vs. trad non-EL	1	285268	31.42	<.001	1.48
		Error	455	9078			
		ONPAR low vs. ONPAR non-EL	1	33881	3.73	.07	.37
		Error	455	9078			
	B/w forms	Trad low vs. ONPAR low	1	35015	3.86	.05	.40
		Error	455	9078			
		Trad non-EL vs. ONPAR non-EL	1	25677	2.83	.09	-.33
		Error	455	9078			

For the eighth-grade form, findings show that, when controlling for science ability, low- English ELs performed significantly better on the ONPAR form than on the traditional form. Further, the performance of low-English ELs on the ONPAR form did not differ significantly from that of their non-EL peers, who performed similarly on both forms. These results indicate that the ONPAR items are more effective in allowing these low-English ELs to demonstrate their content knowledge. Additionally, because no significant differences were found between the low-English ELs and non-ELs on ONPAR, and because the traditional and ONPAR items measure the content similarly for non-ELs, the results suggest that ONPAR is useful in bridging the measurement gap between low-English ELs and non-ELs with similar levels of science ability.

The results were identical for fourth grade except that a small sample size in the low- proficiency EL group (n = 19) substantially reduced the power of the statistical procedure to detect significant differences across forms. So, although low-proficiency ELs scored higher on the ONPAR_LL as compared with the traditional form (means of 451 vs.

424, respectively), the result is not significant. Nonetheless, the findings suggest that the ONPAR strategy is likely to be useful in bridging the measurement gap at this grade level as well.

For both grades, effect sizes are largest between low-English EL and non-EL groups on the traditional forms, and smallest for the non-ELs across forms.

Additional analyses using logistic regression models to identify predictors for scores on individual items indicated that six of the traditional items, across all depth-of-knowledge levels, showed English language proficiency level to be a significant predictor of achievement, compared to only one ONPAR item. Ten of the ONPAR items were significant predictors of student achievement, compared to only five of the traditional items. Of those five traditional items, three also had English language level as a predictor, compared to none of the ONPAR items. This pattern held more strongly for eighth grade than for fourth, but the trend was similar for both grades.

These results indicate that the ONPAR items appeared to do a better job of mitigating the effects of a student's English language proficiency level and to be more effective at demonstrating science achievement. Two implications stand out:

1. ONPAR scores for low-English ELs might be considered interchangeable with scores for the general population derived traditional multiple-choice and constructed-response items. ONPAR items appear to measure the content more validly and effectively for low-English ELs. Moreover, since non-ELs perform similarly on both formats and relative to the same content covariate, the ONPAR and traditional forms seem to be measuring similar constructs.

2. The ONPAR items measure the same content and cognitive complexity as traditional

items, but much more richly and often more directly. In particular, the performance elicited by ONPAR items is more closely tied to the intended latent construct than performance on traditional items, which typically have students either differentiate among choices or explain in writing what they otherwise can demonstrate "live" with ONPAR. Thus, ONPAR assessments appear to measure more fully, and measure a broader range of skills, than traditional assessments.

Piloting in a second ONPAR study in mathematics will be completed in 2011. The study is using a randomized design to investigate how learning disabled (LD) and native English speaking poor readers with no IEPs (poor readers) in Grades 4 and 7 perform on ONPAR as compared to the traditional test measuring the same content. Data are currently available only from a very small sample (n's of 102 and 104 good readers and 37 and 34 low readers on the traditional and ONPAR tests, respectively). Very early analysis of the 4th grade math pilot results show that when adjusted for student math ability, the interaction between Test Type and Reading Level is significant ($F= 6.32$; $p<.01$) with low readers (including both LD students and poor readers) increasing their scores by 41 points from the traditional test to the ONPAR compared to good readers whose scores increased by 15. A third ONPAR study, funded fall of 2009, is developing and researching ONPAR tasks in high school biology and chemistry.