World Journal of Educational Research
ISSN 2375-9771 (Print) ISSN 2333-5998 (Online)
Vol. 8, No. 1, 2021
www.scholink.org/ojs/index.php/wjer

Original Paper

Examining a Multisemiotic Approach to Measuring Challenging Content for English Learners and Others: Results from the ONPAR Elementary and Middle School Science Study

Rebecca J. Kopriva[1*], Laura Wright[1], Robert Triscari[2] & Lynn Shafer Willner[1]

[1] College of Education, Wisconsin Center for Education Research, University of Wisconsin, Madison, Wisconsin, USA

[2] College of Education, Florida Gulf Coast University, Fort Myers, Florida, USA

[*] Rebecca Kopriva, University of Wisconsin, Madison, Wisconsin, USA

Received: November 16, 2020 Accepted: November 25, 2020 Online Published: December 8, 2020

doi:10.22158/wjer.v8n1p1 URL: http://dx.doi.org/10.22158/wjer.v8n1p1

Abstract

This study contributes to the empirical research base on the effectiveness of ONPAR, a promising multisemiotic test item development process. ONPAR uses a variety of multisemiotic performance techniques to present and measure challenging concepts and skills of students, including low English proficient English Learners (ELs) and non-ELs. Experimental trials were used to investigate how 648 ELs at three English proficiency levels and native English speaking non-ELs performed on randomly assigned 4th and 8th grade traditional items and equivalent ONPAR items of challenging science content. General linear modeling using a covariate variable of classroom performance and bi- and multi-nomial regressions found differential boost across both grades. That is, findings showed that lower English proficient ELs perform better on ONPAR vs traditional forms in both grades, with $p < .05$ in favor of ONPAR in grade 8, whereas there were no significant differences between the two forms for non-ELs. The results also underscore the viability of the assessment methodology where students often demonstrate their response by showing their knowledge and skills. Item level results indicate that the ONPAR approach is useful at mitigating the effect of group.

Keywords

ONPAR, assessment, technology enhanced items, test item development, science, English learners, multisemiotic

1. Introduction

The Next Generation Science Standards (NGSS) in K-12 education demand rigorous teaching and learning of challenging content. Yet, the sophisticated language demands associated with complex content are generally beyond the reach of most English Learners (ELs). Do ELs have to wait until they obtain a corresponding level of academic language to fully participate in learning and to demonstrate their mastery of challenging concepts and skills such as multi-step reasoning and complex problem solving? Is it possible to teach and fluidly assess the more complex content Knowledge, Skills and Abilities (KSAs) of lower as well as higher English proficient students?

Many linguistic and EL experts, teachers, and others believe it is. They suggest that ELs do not necessarily lack at least some level of challenging KSAs, but rather that most current assessments of rigorous content lack the facility to communicate with many ELs about their knowledge, reasoning, and strategies. As Walqui and Heritage (2009), Williams, Tang, and Won (2019), and Walqui (personal communication, March 20, 2018) have commented, these less English proficient students are learning, developing unconventional mental schemas and relying on other resources, using their own multimodal meaning representations to acquire and demonstrate more challenging concepts and skills than they can typically read or adequately express to others in English text or orally.

1.1 What Is ONPAR?

The performance-based ONPAR design methodology uses multiple modalities to both present item questions and provide an interactive response environment for students. The ONPAR approach presents stimuli nestled within multi-layered situations, and allows students to respond in such a way that the multisemiotic techniques of animating, frequent interactive sequences, building, drawing, simulating, demonstrating, and reasoning and concept building based on open statement frames, carry a substantial portion of the meaning load to and/or from students. Text (Note 1) is used for precision and as a complementary communication method along with these other modalities. The ONPAR method was designed to primarily measure challenging content using the various types of multi-modal methods to support or substitute for non-construct relevant text, and many of the items studied here are measuring more challenging concepts and skills as well as some more basic questions that require a heavier language load.

The theoretical underpinnings and empirical support for the ONPAR assessment design methodology come particularly from the fields of semiotics, cognitive science, and gaming (e.g., see Myers, 2015). ONPAR's novel, performance-based assessment methodology has strong theoretical and empirical underpinnings. Its item assembly methods have been supported by data obtained from cognitive labs (Kopriva & Wright, 2017; Wright, Staehr-Fenner, Moxley, Kopriva, & Carr, 2013) and experimental studies on elementary and middle school math and high school biology and chemistry (Kopriva, Gabel, & Cameron, 2009; Kopriva, Winter, Triscari, Carr, Cameron, & Gabel, 2013).

1.2 Research Hypotheses

This study was conducted by the University of Wisconsin, in collaboration with the Center for Applied Linguistics and the partner state of Rhode Island. It contributes to the evidence base supporting the validity claims of multisemiotic assessment items, and their relevance for a wide range of students. In this particular investigation, the research team focused principally on measuring more challenging science abilities of elementary and middle school ELs with low proficiency in English. The assumptions underlying the study are, first, that low English proficient students are learning challenging content, not just basic content more commensurate with their current understanding of English. Second, these students are learning more complex content by working with their teachers to take advantage of other modes of representations that, along with some English, are conveying meaning from the teacher to the student and from the student back to the teacher (Kress, Jewitt, Ogborn, & Tsatsarelis, 2001; Lemke, 1998). These modes may include, but often are not limited to their home language for a variety of reasons.

From these assumptions the ONPAR team posited the following hypotheses:

1) Lower English proficient ELs will be able to demonstrate what they know in science significantly better using ONPAR assessment techniques than they could in a traditional test measuring similar content at similar cognitive complexity levels.

2) High English ELs and native English speakers will be able to demonstrate what they know using ONPAR as well as traditional testing formats. The scores of mid-level English proficient ELs on both forms will generally sit between those of their lower and higher English proficient peers.

3) Between groups and within each form, there will be significant differences between low ELs and non-ELs on the traditional form, but less difference on the ONPAR form.

4) At the item level, ONPAR tasks will show less differences between EL and non-EL groups than traditional items measuring the same content.

2. Background

The ONPAR design methodology was developed to address meet a series of ongoing challenges in the assessment of more challenging content, and in the assessment of ELs and others with diverse learning schemas. As the field of assessment design has transitioned from paper-based to computer-based formats, a door has been opened for the innovative methodological approaches proposed by ONPAR.

2.1 Ongoing Challenges for ELs

The number of U.S. K-12 students identified as English learners remains high compared to previous decades, escalating the already enormous pressure on school districts and states to provide not only viable schooling for their ELs that includes not only the basics but the readiness for college and to enter our more sophisticated workplaces (Batalova & Zong, 2016; Garcia, Jensen, & Scribner, 2009). The sophisticated language demands associated with complex content are generally beyond the reach of most ELs unless schooling and assessment involves multi-modal opportunities for students to engage

with (Kim & Herman, 2009). ELs consistently perform below grade level in all content areas, as reflected in significant achievement gaps reported in accountability measures. (For example, the 2015 NAEP average scale score gaps between ELs and their non-EL counterparts were 38 points (U.S. Department of Education/National Center for Educational Statistics, 2016)). Further, ELs are nearly twice as likely as their native English-speaking peers to drop out of high school (Callahan, 2013; Silver, Saunders, & Zarate, 2008).

One contributing factor that begins early is often that standardized or local academic assessments used in whole or part to place students are typically language-intensive, leading to a false impression that the students have little knowledge or—worse yet—that they "can't learn" (Escamilla, 2015; Rumberger & Gándara, 2004). Because ELs are then typically placed, again and again, in classes that are remedial or teaching only more basic content without recourse to challenging content their English-speaking peers are receiving, a large percentage of ELs fall further and further behind native English-speaking peers with the same academic capacity. Thus, for ELs to have equal access to deeper learning opportunities in the classroom, they require access to assessment opportunities reflecting the full range of their cognitive skills.

Solano-Flores (2014) and others argue that even when lower English proficient ELs are learning more challenging cognitive and content, it is often not measured accurately. Many mid-level to higher English proficient ELs also often struggle with comprehending English text and responding to it, especially in middle and high school. This is particularly true in navigating the truncated language of test items (Brown, Donovan, & Wild, 2019; Solano-Flores & Li, 2009). Kopriva et al. (2013) found that highly English proficient ELs students in high school Biology benefited from multisemiotic supported language significantly more than non-ELs. Gebhard (2019) and National Academies of Science, Engineering, and Medicine (2018) among others (e.g., Schleppegrell, 2004; Echevarria, Vogt, & Short, 2004; Walqui & Heritage 2009), argue that students lacking facility in comprehending and producing this distinctive academic language have been shown to benefit from systematic approaches that address language demands using alternative semiotic representations. A body of linguistic research has examined the use of representations other than language in conveying meaning (for instance, see Cope & Kalantzis, 2009a, 2009b, Kress, 2010; Kress & van Leeuwen, 2001, 2006; Choi & Yi, 2016; Gee, 2008; Iedema, 2004). However, till now, little has been done to build this type of methodology centrally into assessments for students whose language-based communication abilities have not yet caught up with the sophistication of their content knowledge.

2.2 Learning and Assessing ELs

Content experts have argued for years that the deep learning that prepares students for challenging high school coursework, postsecondary education and careers involves exposure at all grade levels to activities explicitly designed to elicit complex reasoning, inquiry, and metacognitive skills (e.g., Pellegrino, Chudowsky, & Glaser, 2001; Duschl, Schweingruber, & Shouse, 2007). For students to learn effectively, these experts argue that assessment opportunities reflecting the full range of cognitive

skills must be part of testing. It stands to reason that this means that ELs, as well as others, need to have access to these types of assessment tasks at most grade levels, if this type of challenging coursework is to be taught to them as well as to non-ELs.

Text editing practices associated with Universal Design for Learning (UDL) principles and simplified language and visuals, has been shown to be effective for ELs when items are measuring more basic knowledge or skills (Carr, 2008; Emick & Kopriva, 2007). However, this type of UDL editing and simplified language is frequently not adequate with the more nuanced and abstract language, language structures, and heavier language load used with more challenging content. To-date, these more challenging test requirements and response environments typically still primarily require nuanced reading skills or almost grade level writing skills, carryinga heavy receptive and expressive linguistic load (e.g., American Association for the Advancement of Science (AAAS), 2007; Hansen & Zapata-Rivera, 2010).

Further, technology-rich environments such as those first used on the NAEP 2009 science tests tend to use large amounts of text in developing the problems, nuanced language in the selected response item types, and writing in English (Quellmalz & Silberglitt, 2010). Recently, some drag and drop response types have been added, but by themselves are insufficient for measuring the true mastery of lower English proficient ELs.

2.3 Advances in the Theoretical Design and Interpretations of Assessments

Within the field of educational measurement, the traditional argument for common inferences had been made on procedural grounds: common content in items and a common approach for synthesizing and summarizing items and response data over items. The latter part of this argument required standardized conditions of observation as a key aspect of synthesizing item data. However, based on the foundational work of Mislevy et al.'s (2004) Evidence Centered Design (ECD), Mislevy's extension of this design into gaming (2011), and Kane's (2013) theoretical and procedural advances in providing defensible evidence of test scores using traditional and novel item types, the conceptual argument has become prominent. This argument relies on evidencing appropriate relationships between target inferences, the knowledge and skills of interest, necessary observations, the properties of tasks or items designed to elicit the observations, and the assessment situations where students interact with the assessment requests. It shifts the focus to being able to collect commensurate but not necessarily identical types of evidence and allows for some flexibility in how data are collected as long as validation criteria are critically evaluated.

For some ELs, the rate of English development will be steep; for many, it may take four or more years to achieve the proficiency needed to handle more challenging content text unaided (Goldenberg, 2013; Greenberg Motamedi, 2015), even though these students should concurrently be learning more complex content through other means. The challenge for large-scale, standardized tests and classroom-based assessment is to produce instruments, protocols, and other guidance that can help measure what these students are learning regardless of their ability to articulate these concepts,

reasoning, and skills using primarily text-based methods. As such, this study provides a possible approach.

3. Methodology

To examine the hypotheses discussed in this article, a randomized experimental study was conducted. The first three hypotheses focused at the test level. It examined how ELs at different levels of English proficiency and the control groups would perform on the ONPAR and traditional forms, and if clear differences of form-group combinations could be identified. The fourth hypothesis focused at the item level to quantitatively examine the viability of the ONPAR methodology as compared to the traditional methods.

3.1 Instruments

The main instruments used in this study consisted of (1) released state and federal test science multiple choice and constructed response items selected to measure science standards in the two grades, and (2) ONPAR items measuring the same science targets at the item level. Two sets of supportive instruments were also developed for the study, (3) online tutorials and (4) teacher-rating questionnaires.

First, in identifying the traditional items for the study, the research team selected a set of multiple choice and constructed response items and scoring rubrics measuring more challenging KSAs from a set of released large-scale New England Common Assessment Program (NECAP) tests. In 3 cases released constructed response items from recent NAEP science tests were also used. These items were used to begin the development of the ONPAR items, and a subset of them and their ONPAR 'peers' were selected for the final test forms.

Second, after the development of the ONPAR framework criteria that would guide the construction of the ONPAR items, and the selection of the released items, a set of ONPAR items were built to specifically measure the same intended item targets as each of the traditional items, at the same levels of cognitive complexity and content demand. The ONPAR items were designed to utilize dynamic simulated contexts in computer environments, and to also use a variety of novel response environments that asked students to demonstrate what they knew by manipulating stimuli on the computer in various ways. The test items were assembled carefully using the ONPAR framework to ensure accessibility and their equivalency with target intent of the traditional items. For instance, the science team first built and then revised item storyboards that visually display how the item questions, contextual stimuli, response elements, and response space sit on one or more screens, and how dynamic aspects of the items such as animations would unfold. Using iterative feedback sessions with other staff, science ed and state stakeholders, the ONPAR item designs went through multiple iterations. As the final set of ONPAR items were nearing completion, a group of technology designers and programmers worked beside the ONPAR science team to build the ONPAR items electronically, and to score them in real time using algorithms. All the ONPAR items used novel performance techniques and could not be characterized as either multiple choice or constructed response. A decision was made to constrain the ONPAR scoring

to 0-1 when the traditional versions were dichotomous, and to constrain the polytomously scored items to use measure the same points in each ONPAR item and its traditional counterpart. Individual rubrics for ONPAR were created to mirror the NECAP rubrics used for the open-ended traditional items. Voice-overs of the text questions which students could access as needed were developed in English, Spanish, and Korean and uploaded electronically in each ONPAR item.

As reported elsewhere (Kopriva & Wright, 2017), the team conducted 58 cognitive labs with elementary and middle school students during the ONPAR item development to evaluate the items before they were finalized. The cognitive lab feedback was essential in subsequently refining and finalizing the item design, screen layouts, and assembly processes, evaluating if the multi-modal process was measuring the intended science, and if the designs were accessible to students with a range of challenges. Results in terms of access and measuring the target constructs were generally very positive.

To independently confirm that the ONPAR and traditional items were measuring the same assessment targets at the same level of cognitive complexity the research team consulted with an independent group of assessment and cognitive science experts (Kopriva et al., 2009). Based on their recommendations, the final set of ONPAR/traditional item pairs were selected.

In all, 14 ONPAR/traditional item pairs were selected for the 4th grade ONPAR and traditional test forms, and 13 for the 8th grade forms. Each of the forms were electronically assembled to be delivered by computer. The traditional 4th grade form included 5 constructed response and 9 multiple choice items, and the traditional 8th grade form included 2 constructed response and 11 multiple choice items. In both grades 4 and 8 the traditional form also included an ONPAR item that was not scored, but were included at the end of the traditional forms at the request of the participating teachers. The four forms (two per grade) were assembled electronically and placed in secure online locations accessible during the administrations of the study.

Third, the instruments the team developed were tutorials to be used immediately prior to the test administration. For students taking the ONPAR form in each grade, the tutorial oriented students to the screen layout and the interactive components of the ONPAR items. For students taking the traditional form (with an ONPAR item at the end), the tutorial was typical to classic instructions for the multiple choice and constructed response items, and then a shortened version of the ONPAR tutorial explaining how the ONPAR item is laid out and its interactive components.

Fourth, the researchers developed a teacher-rating questionnaire for each grade using a 3-point rating scale. The questionnaires asked teachers to rate each student on how well the student demonstrated mastery in the classroom regarding particular content objectives aligned to each of the item pairs. Teachers rated students on a scale ranging from consistently below grade level, sometimes below grade level, meets or exceeds grade level, or not covered yet. The rating instruments designed to capture a mid-range grain size of data that Schmidt et al. (2001) found teacher raters could use to differentiate with relatively little guidance.

3.2 Sampling

Four groups of elementary and middle school students (native English-speaking students and low, mid and high English proficient ELs) from 26 schools across five states (Georgia, Illinois, Pennsylvania, Virginia, and Wisconsin) participated in the study. ELs were assigned to one of the groups based on their level of English language proficiency reading score as measured by the WIDA-developed ACCESS for ELLs® assessment: (a) levels 1 or 2 out of 5 (low)), (b) level 3 (mid), or (c) level 4 or 5 (high). In all, 648 students completed the assessments where students were randomly administered ONPAR or traditional forms measuring the same content at the same level of cognitive complexity. The 338 4th grade students came from 16 different schools, and the 310 8th graders were from 10 different schools. School sites overall represented a mixture of urban and suburban areas in Pennsylvania, Virginia, South Carolina, Illinois and Wisconsin, and included students ranging from low to high socio-economic status.

3.3 Procedures

After the development and selection of the items, forms and support instruments, the administration and analyses of the cognitive lab results, and the recruitment of participating schools and teachers, the experimental trials occurred over adjacent spring and fall semesters. In each case, the 4th and 8th grade students (5th and 9th in the fall administration) were randomly assigned one of the test forms for their grade, the traditional item form, or the ONPAR item form.

In conformance with IRB requirements, the research team obtained the necessary participation permissions and signed agreements and conducted online webinar training and question/answer sessions with participating teachers, school administrators, and IT staff.

Before the test administrations, participating teachers received and returned the teacher rating ability questionnaire to be completed for each of their students. Unique student identifier tags for each student per classroom, and instructions to the teacher to prepare the students for the test administrations were sent to each teacher before the test dates.

For testing, students were randomly assigned to forms. This was accomplished using the following process. The unique student identification tags sent to the teachers prior to the administration were developed from the latest classroom lists sent from the participating schools. Going alphabetically down the lists, the student tags included a 2-digit unique number assigned each student per classroom, along with classroom, teacher, and school identifying numbers, and a computer identification number. A copy of these numbers per list of students and educators were kept securely by the research team at the development site, and the researchers at each of the test administrations had a copy in case the student tags were misplaced.

The laptop computers used to administer the tests were assigned a number commensurate with the computer numbers on each student tag, and were brought to the sites by research staff for the test administrations at each school. Each of these computers were preloaded with either the traditional or ONPAR form as well as the associated pre-test tutorials. Matching the computer number on each

student's tag, participants were assigned to this laptop to complete their test form.

The entire testing period lasted about 45 minutes. All data collected via the laptops were subsequently downloaded and placed in a central database by student identifier and type of form for cleaning and analyses.

3.4 Data Analyses

After the data were cleaned and data sets were compiled, the scoring of the test items were completed as well as calibration and scaling of the forms. Descriptive and inferential analyses of total test scores were then computed, followed by item level analyses.

3.4.1 Scoring

After the data were cleaned and data sets assembled to conduct the analyses, each of the dichotomous ONPAR-traditional pairs, and the polytomously-scored ONPAR items (except for one) were scored electronically. NECAP rubrics were used for the five 4th grade and two 8th polytomously scored traditional and ONPAR items where most of the items were scored on a 0-3 scale except for two 4th grade items scored 0-2. Four independent raters (2 per grade) were trained by project staff using industry methods to score the traditional constructed response items and the remaining ONPAR item. Each grade-level rater scored each of the constructed response items, with a 10% read-behind by a third rater from the other grade to ensure reliable results. In the case of adjacent scores, an average score was computed. Project staff mediated in the event that the 10% read-behinds raised concerns or when rater results were not exact or adjacent. Interrater correlations were subsequently computed with results between the two grade-level raters ranging between from .83 to .99. The exact agreement between scores varied from 67% to 93% across items.

3.4.2 Calibration and Scaling

The Rasch model for the dichotomous response items (Rasch, 1980), and one of its polytomous extensions—the Partial Credit Model (PCM, Masters, 1982) for the non-dichotomous items—were used to calibrate, equate, and scale the science ONPAR and traditional test forms. To obtain parameter estimates, the Rasch or PCM models were calibrated for each test form at each grade level.

Prior to scaling, any ONPAR item with a p value below 0.05, or with a theta beyond +/- 4.0 were eliminated from the analysis, along with the corresponding traditional item. This elimination process resulted in removing three pairs of ONPAR/traditional items from the 4th grade test and removing two pairs of corresponding items from the 8th grade test. The final eleven items pairs were scaled for each grade, with five constructed response and six multiple choice items on the final 4th grade traditional form and their corresponding ONPAR items on the ONPAR form for this grade. Likewise, two constructed response and nine multiple choice items made up the traditional form for 8th grade with their corresponding ONPAR items on the ONPAR form. Parameter estimates for the 22 final traditional and ONPAR item pairs at each grade were placed on a common metric for scaling. The randomly equivalent groups design was used and the person ability measures were fixed to a mean of zero,

providing equated estimates across the groups/forms. A linear transformation of the scores was then completed, fixing the mean at 500 and the standard deviation at 100.

3.4.3 Descriptive Analysis

Using the SAS statistical package, frequency data about the sample groups, test items, teacher questionnaire results, and item level data were compiled. Correlations were computed between each of these variables per form and grade.

3.4.4 Inferential Analyses

The General Linear Model program within SAS was used to calculate the omnibus 2 independent variables (group and form) analysis of covariance (ANCOVA) analyses for grades 4 and 8. The covariates were the total scores of the students' data from the teacher rating questionnaires, and the dependent variables were test scores. The ANCOVA results showed main effects as well as general interaction effects. Covariate contrasts were conducted for each cell of the group by form interactions. These analyses, rather than the typical contrast analyses, were calculated using a one independent variable interaction-only ANCOVA to determine if form/group differences could be considered when the main effects were not in the model. Frequency data of the means adjusted for the covariate were also computed.

3.4.5 Item Regressions

To address the viability of the ONPAR forms relative to the traditional method a series of binomial logistic and multinomial/rank order regressions computed in SAS analyzed if the independent variables of group status and classroom accomplishment of the students operationalized by the teachers' ratings at the question level were significant predictors of ONPAR and traditional item scores. Dichotomous items were analyzed using the binomial logistic regressions, and the multinomial/rank order regressions were used for the polychromous items. The methods were used to provide a greater degree of stability of the item coefficients as compared to using traditional regression techniques.

4. Results

This experimental study examined four hypotheses that looked at if and how ELs at different levels of English proficiency and native English speakers might perform on the ONPAR and traditional forms and whether there would be clear differences among form-group combinations. In particular, the study was interested in if ONPAR techniques could minimize systemic biases in typical testing, especially in lower English proficient students, by reducing literacy demands while retaining similar cognitive complexity in science items. To assess the differential impact of form and group differences at the test level descriptive analyses, analysis of covariance, and a pre-defined set of focal and exploratory contrasts. At the item level, a final set of analyses focused on viability of ONPAR methods as compared to traditional item methods.

4.1 Descriptive Analyses

First, frequency counts of the participating sample by grade, form, and group were computed and can be seen in Table 1. Scale score means by grade, form and group can be found in Table 2. Reliability estimates using the standard Cronbach Alpha graded scale statistic was computed by form per grade (4^{th} grade: 0.68 for traditional forms and 0.44 for ONPAR forms; 8^{th} grade: 0.58 for traditional forms and 0.51 for ONPAR forms. The results are low but not surprising given the limited number of items per testlet and the non-standard quality of the ONPAR tasks.

Table 1. Sample by Grade, Form, and Group

Group	Traditional Form		ONPAR Form	
	Grade 4	Grade 8	Grade 4	Grade 8
Low	19	56	21	58
Mid	21	38	29	36
High	28	16	31	16
Non-EL	78	43	110	47
Total	147	153	191	157

Table 2. Scale Test Score Descriptive Data

Grade	Test Form	Low EL			Mid EL			High EL			Non-EL		
		M	SE	N	M	SE	n	M	SE	n	M	SE	n
4	Traditional	424	20	19	460	20	21	506	17	28	527	11	78
	ONPAR	451	17	21	465	18	29	501	19	31	519	10	110
8	Traditional	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

Teacher ratings of individual student performance of topics covered in this study and exhibited in the classroom were used for two purposes: To suggest the student opportunity to learn the topics addressed in test items and to serve as covariates in ANCOVA calculations. The aggregated scores of the topics of the ratings over topics by classroom can be found in Table 3. As expected, means increased in the groups as students learned more English, but the variation in ratings and variance remained similar across groups. These findings are encouraging as they suggest that teachers were able to differentiate students with a similar spread of scores across all groups whether students were ELs or not, and across the three levels of English proficiency. A similar analysis also found that teachers were able to similarly differentiate topics among each group. These results suggest that teachers had a working knowledge of the topics and that students received at least some instruction about each of them. They also provided credibility to the researchers that the ratings could be used as an independent indicator of

student mastery, making it a viable covariate to standardize test score results across groups in the ANCOVAs.

Table 3. Teacher Rating of Students' Science Mastery Level Exhibited in the Classroom

Grade	Test Form	Low EL			Mid EL			High EL			Non-EL		
		M	SE	n	M	SE	n	M	SE	n	M	SE	n
4	Traditional	1.75	0.62	19	1.75	0.70	19	2.01	0.64	28	2.58	0.61	78
	ONPAR	1.76	0.56	21	1.66	0.57	28	2.18	0.62	30	2.58	0.61	109
8	Traditional	2.17	0.62	54	2.21	0.56	37	2.02	0.61	15	2.17	0.69	41
	ONPAR	2.03	0.56	55	2.15	0.52	35	2.04	0.66	16	2.41	0.51	45

4.3 Analysis of Covariance

To address the first three hypotheses for the study, ANCOVA tests and subsequent contrasts were completed for both grades. These analyses examined the differences within between groups and within and between forms.

4.3.1 ANCOVAs, Grade 4

Frequency data, adjusted for the influence of the covariate, are shown in Table 4. In reviewing the results of the omnibus ANCOVA test for grade 4 results indicate that, not surprisingly, there are significant differences by group ($F= 3.92$, $p< .009$) but not form. The interaction was non-significant even though the adjusted means suggest there may be form differences by group especially in the mean differences between forms for low ELs versus other groups (albeit with small n's for low ELs on both the traditional and ONPAR forms), and, within forms where there were greater differences for low and mid ELs versus non-ELs. Because of the magnitude of these differences, we decided to examine the viability of the interaction hypotheses using a one independent variable interaction-only ANCOVA to determine if form/group differences could be considered when the main effects were not in the model. The F (5.66) was $p<.0001$, with a correlation coefficient of almost .4 ($r = .39$) suggesting the differential relationship over cells may be worth investigating.

Table 4. Grade 4 ANCOVA Descriptive Data

Test Form	EL Group	Adjusted Mean	SE	n
Traditional	Low	443.29	21.84	19
	Mid	471.64	21.83	19
	Hi	516.14	17.76	28
	Non-EL	518.02	10.87	78
ONPAR	Low	469.78	20.80	21

	Mid	487.75	18.35	28
	Hi	498.93	17.03	30
	Non-EL	507.96	9.31	109

Table 5 reports the results of a set of ANCOVA contrasts most pertinent to this study. Inspecting the results within groups by form no significant differences were found. That the low ELs in particular did not score significantly differently was disappointing but power for this contrast was very low. Differences in adjusted means within groups for low, high and non-ELs decreased as expected (27, 18, and 10 respectively). Across groups by form, low ELs scored significantly lower than non-ELs on the traditional form, with close but non-significant results ($p<.06$) for mid ELs versus non-ELs, and no differences between high ELs and non-ELs. For the ONPAR form there were no differences between any EL groups and non-ELs. In inspecting the magnitude of the mean differences from the traditional form, adjusted mean score differences of 65 can be seen for low versus non-ELs, 46 score point difference for mid ELs versus non-ELs, and a two-point difference for high ELs as compared to their non-EL peers. In contrast, the ONPAR score differences for low, mid and high ELs versus non-ELs are 38, 20 and 9, respectively. Across groups, there seems to be a trend towards more proficient ELs scoring more similarly to non-ELs on both forms, while the performance of less proficient ELs favored the ONPAR forms. As expected, differences between EL group and non-ELs were substantially smaller for the ONPAR compared to the traditional form suggesting that ONPAR methodology may be useful in reducing disparity in scores across groups due to issues of test format and presentation.

Table 5. Grade 4 ANCOVA Interaction Contrasts

Contrast		df	Contrast SS	Mean Square	F	Sig.
Within Group						
ONPAR Low	Trad Low	1	6997.14	6997.14	0.81	.37
ONPAR Mid	Trad Mid	1	2934.66	2934.66	0.34	.56
ONPAR Hi	Trad Hi	1	4274.95	4274.95	0.49	.48
ONPAR Non-EL	Trad Non-EL	1	4594.79	4594.79	0.53	.47
Between Groups						
Low EL ONPAR	Non-EL ONPAR	1	23375.08	23375.08	2.69	.10
Mid EL ONPAR	Non-EL ONPAR	1	7886.30	7886.30	0.91	.34
Hi EL ONPAR	Non-EL ONPAR	1	1861.94	1861.94	0.21	.64
Low EL Trad	Non-EL Trad	1	78432.85	78432.85	9.03	.003
Mid EL Trad	Non-EL Trad	1	30270.01	30270.01	3.48	.06
Hi EL Trad	Non-EL Trad	1	68.55	68.55	0.01	.93

4.3.2 ANCOVAs, Grade 8

In similar fashion to grade 4, adjusted descriptive data for grade 8 can be found in Table 6, followed by results of the contrasts in Table 7. Overall, there is variation in the spread of the adjusted means for different groups and forms. The main effects of the omnibus ANCOVA are similar to those in grade 4 (significantly different for the group variable ($F=10.28$, $p < .00$), not different for form), but that the interaction effect ($F= 2.45$) is close to significance at .06. The interaction-only ANCOVA was also conducted for grade 8 with results similar to grade 4 ($F= 5.42$ with $p< .0001$). In reviewing the subsequent pair wise contrasts, a number of relevant interactions are apparent. In part, this may be due to the fact that the 8th grade samples for the low and mid ELs are higher than in grade 4.

Table 6. Grade 8 ANCOVA Descriptive Data

Test Form	EL Group	Adjusted Mean	SE	n
Traditional	Low	448.12	13.02	54
	Mid	504.77	15.73	37
	Hi	527.94	24.75	15
	Non-EL	558.75	14.94	41
ONPAR	Low	484.34	12.98	55
	Mid	487.32	16.17	35
	Hi	518.16	23.95	16
	Non-EL	523.51	14.42	45

Table 7. Grade 8 ANCOVA Interaction Contrasts

Contrast		df	Contrast SS	Mean Square	F	Sig.
Within Group						
ONPAR Low	Trad Low	1	35548.76	35548.76	3.89	.05
ONPAR Mid	Trad Mid	1	5471.29	5471.29	0.60	.44
ONPAR Hi	Trad Hi	1	735.49	735.49	0.08	.78
ONPAR Non-EL	Trad Non-EL	1	26319.57	26319.57	2.88	.09
Between Groups						
Low EL ONPAR	Non-EL ONPAR	1	36659.66	36659.66	4.01	.05
Mid EL ONPAR	Non-EL ONPAR	1	25440.35	25440.35	2.78	.10
Hi EL ONPAR	Non-EL ONPAR	1	337.85	337.85	0.04	.85
Low EL Trad	Non-EL Trad	1	285258.32	285258.32	31.18	<.0001
Mid EL Trad	Non-EL Trad	1	56670.18	56670.18	6.19	.01

Contrast		df	Contrast SS	Mean Square	F	Sig.
Hi EL Trad	Non-EL Trad	1	10397.50	10397.50	1.14	.29

As shown in Table 7, within groups, low ELs showed a significant preference for the ONPAR form. While none of the other groups appeared to favor a form, it is interesting to note that the mid EL means adjusted for the covariate were higher on the traditional form than the ONPAR. When the scores for each EL group were contrasted by form with those from the native English non-EL group, low and mid ELs scored significantly lower than their non-EL peers on the traditional form. Again, no significant differences between high and non-ELs were identified, although the small sample size of high ELs are low (n's of 16 per form). For the ONPAR form the low ELs scored significantly differently from non-ELs ($p<.05$), and otherwise findings were non-significant. Similar to what was found in grade 4, the trend in the average score gaps between non-ELs and low and mid EL groups lessened when the ONPAR form was administered as compared to the traditional form. Score differences of 40 versus 111 for low ELs on the ONPAR as compared to the traditional form, and 37 versus 54 for mid ELs on these respective forms can be seen. The spread of score differences for high ELs was 6 on ONPAR and 31 on the traditional.

4.4. Item Level Analyses

In order to quantitatively examine the viability of the ONPAR methodology as compared to the traditional methods, a series of item level regression analyses on the ONPAR and traditional items at both grade levels were conducted to address the 4th hypothesis. Eleven item pairs were analyzed in 4th grade (Table 8) and 11 in 8th grade (Table 7). Of these, five of the 4th grade items and two of the 8th grade items were polytomously scored.

Table 8. Grade 4 Item Regressions

	Item Name	DOK*	IV	Traditional Form				ONPAR Form			
				n	B	Wald	Sig.	n	B	Wald	Sig
Dichotomous	Springs	1	EL	144	0.66	6.77	.009	186	-0.03	0.13	.72
			Rating		0.09	0.04	.84		0.37	3.14	.08
	Crocodile Life Cycle	2	EL	143	0.03	0.10	.75	185	-0.06	0.38	.54
			Rating		-0.02	0.009	.92		-0.30	1.86	.17
	Substance	2	EL	142	-0.09	0.54	.46	185	-0.02	0.02	.90
			Rating		-0.39	2.15	.14		-0.75	7.95	.005
	Food Web Crises	2	EL	144	-0.04	0.15	.70	186	-0.21	3.61	.06
			Rating		0.22	1.03	.31		-0.11	0.21	.65

	Settling	2	EL	143	0.11	1.25	.26	188	-0.11	1.41	.24
	Rocks		Rating		-0.20	0.83	.36		0.09	0.21	.65
	Earth	2	EL	141	-0.06	0.27	.60	185	-0.10	0.75	.39
	Materials		Rating		-0.27	1.33	.25		-0.53	4.39	.04

	Item Name	DOK	IV	Traditional Form				ONPAR Form			
				n	B	Wald	Sig.	n	B	Wald	Sig
Polytomous	Buoyancy	2	EL	144	-0.25	1.51	.22	181	-0.06	0.10	.75
			Rating		-0.19	0.44	.51		-0.34	1.69	.19
	Heated	2	EL	141	-1.30	5.02	.03	183	-0.59	6.07	.01
	Sand		Rating		0.18	0.12	.73		-0.23	0.55	.46
	Rolling	2	EL	136	-0.56	5.77	.02	181	-0.36	2.91	.09
	Balls		Rating		-0.33	1.30	.26		0.29	1.02	.31
	Pond	3	EL	137	-0.49	3.32	.07	181	0.04	0.04	.85
			Rating		-0.51	1.66	.20		-0.39	1.23	.27
	Magnets	3	EL	140	-0.16	0.23	.63	188	0.05	0.07	.79
			Rating		-0.53	1.30	.25		-0.51	3.46	.06

* Depth of knowledge (an indicator of cognitive complexity).

Table 9. Grade 8 Item Regression

	Item Name	DOK	IV	Traditional Form				ONPAR Form			
				n	B	Wald	Sig.	N	B	Wald	Sig
Dichotomous	Animal	1	EL	145	-0.03	0.11	.74	145	0.17	3.24	.07
	Cell		Rating		-0.17	1.22	.27		-0.30	3.14	.08
	Saturn	1	EL	146	-0.07	0.80	.37	151	0.12	1.78	.18
			Rating		-0.10	0.45	.50		-0.41	5.49	.02
	Changing	2	EL	143	0.02	0.04	.84	145	-0.21	3.13	.08
	Phases		Rating		-0.36	4.88	.03		-0.22	1.17	.28
	Molecules	2	EL	145	0.21	6.28	.01	149	0.11	1.39	.24
			Rating		-0.40	6.44	.01		-0.06	0.13	.72
	Force	2	EL	146	0.11	1.79	.18	145	0.14	0.90	.34
	Box		Rating		-0.23	2.29	.13		-1.23	16.89	.000
	Muscle	2	EL	147	0.28	10.006	.002	151	-0.06	0.18	.67
	Respiration		Rating		-0.43	7.56	.006		-0.73	9.23	.002
	Heat	2	EL	144	0.32	12.56	.000	150	0.11	1.28	.26

	Transfer		Rating		-0.59	12.71	.000		-0.45	6.33	.01
	Density	2	EL	141	-0.05	0.37	.54	143	0.10	1.02	.31
			Rating		-0.01	0.003	.96		-0.58	8.66	.003
	Photo-synthesis	2	EL	144	0.04	0.22	.64	147	0.03	0.04	.84
			Rating		-0.63	11.80	.001		-0.93	12.72	.000
Polytomous	Item Name	DOK	IV	Traditional Form				ONPAR Form			
				n	B	Wald	Sig.	N	B	Wald	Sig
	Toolkit Electricity	2	EL	140	-0.75	25.60	<.0001	146	-0.20	2.43	.12
			Rating		-0.21	0.66	.42		-0.19	0.69	.41
	Power Plant	2	EL	140	-0.61	17.42	<.0001	147	-0.38	5.93	.01
			Rating		-0.39	2.78	.10		-0.10	0.19	.66

Overall, seven of the 11 items in grade 4 and 10 of the 11 items in grade 8 indicated a significant difference in EL group and/or students' science ability as rated by the teachers. Specifically, results indicate that ONPAR items are much less likely than the traditional versions to have EL status as a significant predictor. In total, six of the traditional items (two in grade 4 and four in grade 8), show EL to be a significant predictor (with positive Betas in the dichotomous items; negative in the polytomous items), compared to zero ONPAR items. For two of the polytomous item pairs (one in each grade), the EL group was found to be a significant predictor on both forms, where, of note, the Betas were again consistently negative.

There may also be some indication that the independent teacher ratings may better predict performance on the ONPAR items as five ONPAR items (two in grade 4 and three in grade 8) showing student accomplishment as a significant predictor compared to two 8th grade traditional items. In three cases (all in 8th grade) both dichotomous item versions were significant. Interestingly, all the significant Betas for science ability showed an inverse relationship whether they were traditional or ONPAR items.

5. Discussion

This proof-of-concept experimental study investigated a novel approach to conveying meaning in challenging science assessment items. This technology-based methodology called ONPAR, uses multisemiotic stimuli to convey meaning in substantive ways, rather than employing it as "window dressing" to text. The intent of this work is to support learning of challenging content and practices articulated in recent science education standards such as NGSS, and encourage that all students are included in the teaching and learning of more complex content, reasoning, and strategic skills that extend beyond basic recall and procedural application. This is important because this study, as well as the cognitive labs (Kopriva & Wright, 2017) strongly suggest that ELs, without the requisite amount of

formal academic language, still learn more challenging content if we give them a chance. This approach represents a way of tracking their progress.

Finally, the items and tasks developed using this approach are designed to provide a set of innovative techniques and strategies for developing standardized summative and diagnostic tests capable of being effectively used by students who otherwise cannot access the nuanced language typically associated with more challenging content. While the focus here is ELs, other studies suggest that disadvantaged students and others who learn and explain what they know in non-standard ways, may also benefit (Kopriva et al., 2013; Kopriva, Wright, Malkin, & Myers, 2019).

To these ends, four hypotheses were investigated in this study, three at the test level of analyses and one at the item level. The focal groups were students with low English language proficiency, ELs with mid and high levels of English language proficiency, and native English speakers. Data from a total of 648 students who took the ONPAR or traditional science items measuring grades 4 and 8 science content and skills have been analyzed in this study.

5.1 Hypothesis #1: Focus on the Performance of Lower Proficiency ELs

Would lower proficient ELs be able to demonstrate what they know in science significantly better on items using ONPAR techniques? The results tended to support that the ONPAR methodology is effective for grade 8 students with lower levels of language proficiency to access what is being required and to demonstrate what they know. Further, 8^{th} grade comparisons of low proficiency ELs with non-ELs finds that while there is significant differences between these groups on both forms, the adjusted mean differences on the traditional form ($F= 31.18$ with a p of $<.0001$) far outweigh the differences on the ONPAR form ($F=4.01$, $p<.05$). This distinction by form between the two groups suggests that the ONPAR form is providing access for lower English proficient ELs. While the grade 4 ONPAR adjusted mean scores were not significantly different than their traditional scores for lower proficiency EL students (or from non-ELs on the respective forms for that matter), it seems that low statistical power may be at least partially at fault, since the cognitive labs with younger students certainly suggested differences in access favoring ONPAR items. Clearly, further research needs to be completed to confirm this.

It is noteworthy to mention that, with much higher n's per form, the difference between form types was never significantly different for native English speaking non-ELs in either grade. This is good news because it suggests that either form is accessible to this group meaning that this approach could be used in performance tasks for all students, as well as in classroom-based products for all students. Additionally, non-ELs will not tend to be disadvantaged if and as students may be given an ONPAR item adaptation rather than its traditional counterpart in large-scale summative testing.

5.2 Hypothesis #2: Focus on the Performance of High, Mid-Level and Non-EL Students

How might high and non-EL students demonstrate what they know using ONPAR as well as traditional testing formats? Would mid level ELs will follow the trend of smaller differences as they acquired more English? As expected, high English proficient ELs behave like non-ELs with no significant mean

score differences between forms or with non-ELs in either grade (although small samples in grade 8 suggest this finding needs to be confirmed). One caveat to this finding comes from ONPAR research in high school biology classrooms that suggests how high proficiency ELs might behave over grades (Kopriva et al., 2013). Here, the researchers noted that, while ELs significantly preferred ONPAR tasks as compared to traditional tasks measuring the same content at a similar level of cognitive complexity, closer inspection found that most of these ELs had high English proficiency (very few lower proficiency ELs were in the sample). This finding suggests a possible age effect, and further research needs to be completed to understand differences in access between younger and older ELs.

Results for mid-proficiency ELs at both grades were a surprise in that the adjusted mean differences across forms were relatively low and insignificant (16 points difference in grade 4, and 18 in grade 8). This held true even when sample sizes were small (in grade 4) and more robust (in grade 8). In fact, in grade 8, the traditional adjusted mean score was higher than the ONPAR, which is consistent with high proficiency and non-ELs, whereas in grade 4, slightly higher mean scores favored the ONPAR form, a preference that was consistent with low-proficiency EL mean score comparisons. Two previous studies using static forms had found that mid-level English proficiency ELs are more overall a volatile group in terms of how they access items (Emick & Kopriva, 2007; Carr, 2008). Applied linguistic experts have discussed that, in many cases, there is a mid-point within the development of English proficiency where ELs shift from focusing more holistically on meaning making in items to becoming hyper-focused on small elements in the text or assessment tasks. This shift often looks like it decreases comprehension, especially of novel stimuli and/or more complex content. However, as these students attain higher levels of proficiency this hyper-focus relaxes once again and they are able to grasp more nuanced meaning than low and low-mid proficiency ELs and can again discriminate between preferred and non-preferred communication modalities (for instance, see Graves, August, & Mancilla-Martinez, 2013). Clearly, more research needs to be conducted here as well.

5.3 Hypothesis #3: Focus on the Performance Within and Across Groups and Forms

Would there be significant differences between low ELs and non-ELs on the traditional form, but less difference on the ONPAR form? This study tracked the spread of scores between each EL group and their non-EL peers on both forms. Across grades, findings with one exception consistently indicated that the spread of the mean score differences on the traditional forms were substantially more pronounced for each EL group, as compared to the spread of mean differences on the ONPAR forms. This occurred everywhere except for high- proficiency ELs versus non-ELs at grade 4 where the differences were rather flat (differences of two score points on the traditional form versus nine score points on ONPAR) compared to the magnitude of the spread in all other cases. Looking at the spread of traditional mean scores for three dyads of students, the spread between EL group and non-ELs averaged 26, a spread of 17 for mid-proficiency ELs and non-ELs was reported in grade 8, and a whopping 71 point spread of mean differences was found for grade 8 low vs. non-ELs.

5.4 Hypothesis #4: Focus on the Methodological Impact

At the item level, when and how might ONPAR tasks minimize the impact of group status on the performance in the items and capture the KSAs of low English ELs better than the traditional methods? There seemed to be indications in the item regression examinations that effect of group status may be reduced using the ONPAR method but low sample sizes complicate the findings, especially in grade 4. Likewise, teacher rating estimates of science ability were a little more strongly related to lower EL performances on five ONPAR versus traditional items with significant negative Betas, as compared to their performance on traditional items in two pairs.

This study also tracked the spread of scores between each EL group and their non-EL peers on both forms. Across grades, findings with one exception consistently indicated that the spread of the mean score differences on the traditional forms were substantially more pronounced for each EL group, as compared to the spread of mean differences on the ONPAR forms. This occurred everywhere except for high- proficiency ELs versus non-ELs at grade 4 where the differences were rather flat (differences of two score points on the traditional form versus 9 score points on ONPAR) compared to the magnitude of the spread in all other cases. Looking at the spread of traditional mean scores for three dyads of students, the spread between EL group and non-ELs averaged 26, a spread of 17 for mid-proficiency ELs and non-ELs was reported in grade 8, and a whopping 71 point spread of mean differences was found for grade 8 low vs. non-ELs.

These results suggest that the ONPAR methodology may be useful in helping to closing the performance gap due to communication methods and level the performance “playing field” between ELs and non-ELs, especially for lower English proficient students. It also supports the notion that individual ONPAR items may be a viable item adaptation for some ELs while other students take the traditional counterparts.

6. Implications

All in all, this study suggests that the ONPAR approach holds promise, although it has also highlighted possible pitfalls that need to be more fully understood. Fundamentally, the methodology illustrates how a number of multisemiotic techniques, properly assembled and implemented, can be used to substantively convey meaning. This approach broadens the status of these techniques in measurement beyond what has typically been their peripheral role vis a vis written text in presenting problems and scenarios to both EL and non-EL test takers, and in configuring how they are able to respond. While by no means eliminating text, ONPAR depicts how a number of the multisemiotic representations can be used to convey meaning in tandem with language or on their own.

Potential uses of such a methodology include broadening meaningful assessment of challenging content for students who have barriers to demonstrating their abilities using traditional formats, and enhancing how online performance tasks are designed to elicit real time demonstration of conceptual understanding and skills as well as identify students’ strategies, reasoning and other processes important for developing deeper mastery of content. In particular, multisemiotic presentations and

novel response methods, like those exemplified here, seem to be well-suited to improve how we measure educational content in at least two distinct ways.

First, by carefully integrating multiple sources of stimuli together within and over screens to convey intended constructs and supplement response opportunities, this approach can selectively yet pervasively increase accessible communication avenues so students for whom language, literacy, or, most likely, language processing challenges are often a barrier can more effectively access problems and better demonstrate their understanding. Here, the findings imply that the ONPAR methodology would seem to be useful and effective for measuring these more complex knowledge and skills, for students whose language is less well developed as well as for students with adequate literacy skills. We and others (for instance, see Gee, 2008; and Roth, 2005, among others) contend though, that supporting assessment access to more complex content knowledge and skills has the potential to encourage learning of more nuanced academic language by engaging students in the content material and leveraging that interest as impetus to learn the academic language associated with this content. As Roth (2005) shows, when learning science, students often engage in "muddled" discussions about the phenomena; as their conceptual understanding grows through hands-on experience, so too, does their ability to "talk science." A 2018 National Academies of Science, Engineering, and Medicine report concurs: Academic language development must be anchored in conceptual understanding.

Second, the researchers assert that this study defends the viability of more directly presenting "live" (albeit technology-based) performance scenarios in testing versus traditional modalities that rely almost exclusively on indirect descriptions or hands-on performances. As the testing field begins to tackle how to utilize the capabilities of technology, ONPAR researchers have spent considerable time deliberating and illustrating how to concurrently utilize multiple stimuli without overwhelming or confusing students, how to integrate various presentation and response avenues, including language, into seamless communication packages, how specific stimuli or combinations of stimuli work to convey various kinds of meaning for a number of purposes, and how to capture and concurrently score different types of performances, reasoning arguments, and a range of relational and meta-cognitive explanations using representations that are generated rather than selected by students.

While the technology capabilities referenced in ONPAR are rather straightforward and familiar to most of us who utilize the internet today, what makes ONPAR 'work' is not the individual techniques but rather HOW they are combined to suit a large number of purposes. These purposes range from clarity and defensibility of the targeted material and issues of access for literate students, those without much English, and those in-between, to building tailored situations and site spaces for presenting construct-relevant multimodal scenarios and capturing a wide set of responses. The ONPAR methodology developed thus far is templating how some these types of purposes might be met. Continuing research is of course necessary to validate how well the approach is producing evidence demanded by the test inferences, for whom, and under what conditions. The proof-of-concept study of this methodology presented here seems to be a viable step forward in that direction.

Acknowledgements

This article was partially supported by US Department of Education, Grant No. 368A20061.

References

American Association for the Advancement of Science (AAAS) Science Assessment. (2007). Project 2061 science assessment website [Test item database]. http://assessment.aaas.org/

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Batalova, J., & Zong, J. (2016). Language diversity and English proficiency in the United States. *The Online Journal of the Migration Policy Institute*.

Brown, B., Donovan, B., & Wild, A. (2019). *Language and cognitive interference: How using complex scientific language limits cognitive performance. Science Education*, *103*(4), 750-769. https://doi.org/10.1002/sce.21509

Callahan, R. M. (2013). The academic achievement of immigrant adolescents: Exploration of school factors from sociological and educational perspectives. In M. Gowda, & A. Khanderia (Eds.), *Educational achievement: Teaching strategies, psychological factors and economic impact* (pp. 53-74). Nova Science.

Carr, T. G. (2008). *Qualitative review of items that worked and didn't work*. Paper presented at the National Council of Measurement in Education annual meeting, New York, NY.

Choi, J., & Yi, Y. (2016). Teachers' integration of multimodality into classroom practices for English language learners. *TESOL Journal*, *7*, 304-327. http://dx.doi.org/10.1002/tesj.204

Emick, J., & Kopriva, R. J. (2007). *The validity of large-scale assessment scores for ELLs under optimal testing conditions: Does validity vary by language proficiency?* Presentation at the American Education Research Association annual meeting, Chicago, IL, April.

Escamilla, K. (2015). Schooling begins before adolescence: The case of Manuel and limited opportunities to learn. In D. Moelle, E. Sato, T. Boals, & C. Hedgepeth (Eds.), *Multilingual learners and academic literacies; sociocultural contexts of literacy development in adolescents* (pp. 210-227).

Gebhard, M. (2019). *Teaching and researching ELLs' disciplinary literacies: Systemic functional linguistics in action in the context of U.S. school reform*. Routledge. https://doi.org/10.4324/9781315108391

Gee, J. P. (2008). What is academic literacy? In A. Rosebery, & B. Warren (Eds.), *Teaching science to English language learners* (pp. 57-70). NSTA Press. https://doi.org/10.2505/9781933531250

Goldenberg, C. (2013). Unlocking the Research on English Learners: What We Know—and Don't Yet Know—About Effective Instruction. *American Educator*, *37*(2), 4.

Graves, M. F., August, D., & Mancilla-Martinez, J. (2013). *Teaching vocabulary to English language*

learners. Teachers College Press.

Greenberg Motamedi, J. (2015). *Time to reclassification: How long does it take English learner students in Washington Road Map districts to develop English proficiency*? (REL 2015-092). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northwest. https://roadmapproject.org/wp-content/uploads/2018/09/REL_2015092.pdf.

Hansen, E., & Zapata-Rivera, D. (2010). *Designing Assessment-for-Learning (AfL) systems for diverse students: A game-based AfL for learning middle school mathematics*. Paper presented at the National Council on Measurement in Education, Denver, CO.

Kalantzis, M., & Cope, W. (2009). A Grammar of multimodality. *The International Journal of Learning: Annual Review*, *16*(2), 361-426. https://doi.org/10.18848/1447-9494/cgp/v16i02/46137

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73. http://dx.doi.org/10.1111/jedm.12000

Kim, J., & Herman, J. L. (2009). A three-state study of English learner progress. *Educational Assessment*, *14*(3-4), 212-231. http://dx.doi.org/10.1080/10627190903422831

Kopriva, R. J. (2014). Second generation challenges for making content assessments accessible for ELLs. *Applied Measurement in Education*, *27*(4), 301-306. https://doi.org/10.1080/08957347.2014.944311

Kopriva, R. J., & Wright, L. (2017). Validating score meaning for non-native speakers. In K. Ercikan, & J. W. Pellegrino (Eds.), *Validation of score meaning in the next generation of assessments*. National Council of Measurement in Education Press, Washington D.C.

Kopriva, R. J., Gabel, D., & Cameron, C. (2009). *Overview of results from the ONPAR elementary and middle school experimental study with ELs and non-ELs: A promising new approach for measuring complex content knowledge of English learners with lower proficiency levels*. http://www.iiassessment.wceruw.org.

Kopriva, R. J., Thurlow, M. L., Perie, M, Lazarus, S. S., & Clark, A. (2016). Test takers and the validity of score interpretation. *Educational Psychologist*, *5*(1), 108-128.http://dx.doi.org/10.1080/00461520.2016.1158111

Kopriva, R. J., Winter, P. C., Triscari, R., Carr, T. G., Cameron, C., & Gabel, D. (2013). *Assessing the knowledge, skills, and abilities of ELs, selected SWDs and controls on challenging high school science content: Results from randomized trials of ONPAR and technology-enhanced traditional end-of-course biology and chemistry tests*. Institute for Innovative Assessment, University of Wisconsin-Madison. http://iiassessment.wceruw.org/research/#onpar

Kopriva, R. J., Wright, L. Kreikemeier, P., Myers, B. (2019). *Technology-Interactive Classroom-Embedded Modules for Measuring Challenging Math and Science Skills of ELs Final Report*. Institute for Innovative Assessment, University of Wisconsin-Madison. http://iiassessment.wceruw.org/research/#onpar

Kress, G. (2010). *Multimodality: A social semiotic approach to contemporary communication*. Routledge. https://doi.org/10.4324/9780203970034

Kress, G., & van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. Oxford University Press.

Kress, G., Jewitt, C., Ogborn, J., & Tsatsareis, C. (2001). *Multimodal teaching and learning: The rhetorics of the science classroom*. Continuum.

Lemke, J. (1998). Multiplying meaning: Visual and verbal semiotics in scientific text. In J. R. Martin, & R. Veel (Eds.), *Reading science: Critical and functional perspectives on discourses of science* (pp. 87-113). London: Routledge.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. http://dx.doi.org/10.1007/BF02296272

Mislevy, R. J. (2011). *Evidence-centered design for simulation-based assessment. CSE Technical Report 800*. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Myers, B. (2015). *The cognitive basis of ONPAR assessment: A white paper*. Madison: Institute for Innovative Assessment, University of Wisconsin-Madison.

National Academies of Science, Engineering, and Medicine. (2018). *English learners in STEM subjects: Transforming classrooms, schools, and lives*. National Academies Press.

New London Group. (1996). A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review*, *66*(1), 60-92. http://dx.doi.org/10.17763/haer.66.1.17370n67v22j160u

Quellmalz, E. S., & Silberglitt, M. D. (2010). *Opportunities and challenges of designing technology-based learning-centered assessments for diverse students*. Paper presented at annual meeting of the American Educational Research Association, Denver, CO.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.

Roth, W. M. (2005). *Talking science: Language and learning in science*. Lanham, MD: Rowman & Littlefield.

Rumberger, R. W., & Gándara, P. (2004). Seeking equity in the education of California's English learners. *Teachers College Record*, *106*(10), 2032-2056. http://dx.doi.org/10.1111/j.1467-9620.2004.00426.x

Schleppegrell, M. J. (2004). *The language of schooling: A functional linguistics perspective*. Erlbaum. https://doi.org/10.4324/9781410610317

Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H., Wiley, D. E., Cogan, L. S., & Wolfe, R. J. (2001). *Why schools matter*. John Wiley & Sons.

Silver, D., Saunders, M., & Zarate, E. (2008). *What factors predict high school graduation in the Los Angeles Unified School District*? California Dropout Research Project, UC Santa Barbara.

Solano-Flores, G. (2014). Probabilistic approaches to examining linguistic features of test items and

their effect on the performance of English language learners. *Applied Measurement in Education*, *27*(4), 236-247. http://dx.doi.org/10.1080/08957347.2014.944308

Solano-Flores, G., & Li, M. (2009). Language variation and score variation in the testing of English language learners, native Spanish speakers. *Educational Assessment*, *14*(3-4), 180-194. http://dx.doi.org/10.1080/10627190903422880

U.S. Department of Education/National Center for Educational Statistics. (2016). NAEP data explorer. Author.

Walqui, A., & Heritage, M. (2009). *Instruction for diverse groups of English language learners*. Understanding Language: Language, Literacy, and Learning in the Content Areas.

Williams, M., Tang, K-S, & Won, M. (2019). ELL's science meaning making in multimodal inquiry: A case-study in a Hong Kong bilingual school. *Asia-Pacific Science Education*, *5*(1), 1-35. https://doi.org/10.1186/s41029-019-0031-1

Wright, L. J. (2015). Inquire to acquire: A discourse analysis of bilingual students' development of science literacy. In D. Molle, E. Sato, T. Boals, & C. A. Hedgspeth (Eds.), *Multilingual learners and academic literacies: Sociocultural contexts of literacy development in adolescents*. Routledge.

Wright, L. J., Staehr-Fenner, D., Moxley, K., Kopriva, R. J., & Carr, T. G. (2013). *Exploring how diverse learners interact with computerized, multi-semiotic representations of meaning: Highlights from cognitive labs conducted with ONPAR end-of-course biology and chemistry assessment tasks*. http://www.iiassessment.wceruw.org.

Note

Note 1. To clearly distinguish between ONPAR techniques used in this study, text is defined here as alphabetic-based language forms.