Responses Processes in Assessing Academic Content of Non-native Speakers: Literature Review and ONPAR Summary

Rebecca J. Kopriva and Laura Wright
University of Wisconsin

Building on Mislevy (1994) and Kane's (2013) work, Kopriva, Thurlow, Perie, Lazarus and Clark (2016) outline the dominant conceptual framework this chapter will use as we discuss considering response process data of non-native speakers to validate academic score meaning. Kopriva et al. assert that valid measurement involves more than: (a) identifying intended assessment content and cognitive demands, (b) proposing consistent claims to infer from the operationalization of constructs, (c) building clear tests that consistently measure these demands, and (d) using evidence and theory to support the proposed interpretations for specific uses. They argue that valid measurement must also include identifying and addressing the the test takers, who are as much a part of the testing operation as the content being testing, the testing machinery that presents and delivers the content and responses, and the evaluations of validity and reliability. The authors refer to this as the "Person Dimension", which is akin but more pervasive and inclusive than simply addressing current definitions of fairness. The Dimension involves several aspects of impact. It includes considering student background and abilities not associated with particular test content; past and current experiences that aggravate access to typical test conditions; and how these influences impact how students attend to items in non-standard or unintended ways, how and how well they process, and how they perform on tests. While the aspects have an impact in different ways, the result is they call into question the proper interpretation of the academic test scores for certain students. The dimension is particularly pertinent when the impact involves the interaction of individual students' characteristics with construct-irrelevant testing attributes within test situations where responses are generated and evaluated.

However, the problematic interactions can often subtle if not well understood. The focus must first be on addressing the construct irrelevant situations known to act as sufficient barriers to access and measurement of the intended content and cognitive skills, knowing that this is a necessary but not sufficient step in producing defensible score interpretations. For most assessments tied to ongoing learning and the classroom, the impact of the testing situations the test takers experience in assessment situations should be understood and act as catalysts for assessment adaptation at a more nuanced level. Adaptation is accomplished such that the target constructs remain constant across students and score meaning is improved for various individual students. Kopriva et al. contend that this should be occurring even when the situations don't reach the level of an inaccessible barrier, and that not doing so will decrease the validity of the interpretations. These authors argue that sufficient documentation of evidence and theory associated with the impact of and interaction with the Person Dimension should be a necessary and essential component of validating score meaning for all test takers. For non-native speakers two particularly pertinent person x test interaction aspects have a direct impact on the quality of score meaning—students successfully comprehending what items are asking, and having access to response situations that allow them to communicate their solution reasonably well. A third, access to internal problem solving strategies and related skills, is also key but less well understood.

This chapter first reviews salient response process literature associated with these three aspects, and then summarizes a promising approach to testing non-native speakers called ONPAR that attempts to improve the validity of the scores by introducing possible solutions to a number of the key problems raised in the literature. For the chapter review the small-scale response process research that will be reviewed involves the following data collection methods: individual interviews, direct observations using field data retrieval techniques,

qualitative inspections of items or other testing features, and small exploratory research studies where data analyses focus on test taker intent and behaviors. Every effort has been made to choose research that appears to be of high quality and uses defensible protocols and systematic analytic procedures. The reviews are not comprehensive but representative of the types of findings available for validation purposes. Using an integrated systems approach to designing items, the purpose of the ONPAR section is to introduce an assessment model that holistically increases points of access, and uses multi-semiotic and load-sharing linguistic techniques to improve the validity of score meaning for a variety of non-native speakers. By tackling a number of the problems raised in the response data literature in the first section, the ONPAR methodology research to-date summarizes how, when and why such techniques appear to broaden access to test item questions and provide better evidence of intended score inferences than was evident using primarily traditional approaches.

I. Review of Relevant Response Process Literature

1. Comprehending What the Test Items are Asking

Most of the large- and small-scale research to date agrees that without comprehension of what the items are asking, test scores of non-native speakers will tend to remain distorted by their meaning negotiations over languages and diverse demographic, academic and cultural experiences. For non-native speakers comprehension refers to understanding the target measurement focus of the items, including how meaning is conveyed. Winter et al. (2006) distinguished between full and partial comprehension. ELs who fully comprehended what items were asking performed substantially better than those with partial comprehension, where those with partial comprehension understood the gist of the question, but missed the nuanced overlays or more subtle aspects.

Linguistic Features A distinction is made here between linguistic complexity and cognitive content complexity. Studies have shown that when more challenging academic content involving greater cognitive content complexity is measured, greater linguistic complexity often becomes a key construct-irrelevant factor impacting content score meaning. This is because the vocabulary, language, and discourse structures used to convey sophisticated concepts, reasoning and skills are generally also sophisticated. In theory some may argue that this puts more challenging content off limits to those with lower language proficiency. However, there is ample empirical literature showing that those who have not yet developed full proficiency in a second (school) language can and do learn complex content (i.e. Gee, 2002; Schleppegrell, 2004). The challenge for teachers and measurement professionals is to utilize the 'work arounds' students use to understand and convey meaning in an assessment while not altering the cognitive demands of the content.

For example, Martiniello (2008) used six state mathematics test items in English and conducted think-aloud interviews with 24 4th grade students whose first language was Spanish and who had been in U.S. schools two or more years. She found that students made errors in item sentences that used multiple clauses, long noun phrases, and limited syntactic transparency, and problems with lower frequency words and those with multiple meanings. Even when students were able to decode and literally comprehend the meaning of the text, expressions were problematic that referenced unfamiliar contexts and cultural references (often subtle and such a part of the U.S. cultural lexicon the problems were not identified in reviews). These findings are generally consistent in other subjects and with students from other language backgrounds (Logan-Terry & Wright, 2010; Logan-Terry, 2011; Noble, Rosebery & Suarez, 2015; Noble, Suarez, Rosebery, O'Connor, Warren, & Hudicourt-Barnes, 2012; Wright, 2008; Wolf & Leon, 2009. Winter, Kopriva, Chen, and Emick's

(2006) cognitive lab study with 156 3rd and 5th graders found that phrasal verbs and verb tenses are also problematic for non-native speakers. Such expressions and grammatical forms are rarely found in English glossary accommodations and need to be properly negotiated for non-native speakers, as do atypical and colloquial language (see also Carr, 2008).

Prosser and Solano-Flores (2011) conducted cognitive labs with 78 Spanish-speaking and monolingual English speakers. They found that how ELs versus non-ELs acquired their understanding of text-based test items differed in how they interacted with the science and non-science content terms, and how they approached and understood more linguistically challenging sentences. In navigating native and L2 languages, Solano-Flores and Li's analysis of differential student level variance (2006) found that, because of specific linguistic, cultural, and localized influences of particular item elements, the Haitian-Creole speakers sometimes performed better with English or standard Haitian-Creole text. Roth, Oliveri, Sandilands, Lyons-Thomas and Ercikan (2013) conducted think-alouds with a group of expert translators considering English and French versions of items, and noted differences in the length of French and English versions, syntactic and semantic differences and differences in the logical structure of item content or form as well as cultural issues.

A few projects have looked at the interaction between linguistic complexity and content complexity. Martiniello (2008) noted that ELs had more problems overall with the language in more difficult content items than in less difficult items. Carr's (2008) qualitative review of English language traditional and 'access-based' versions of mathematics, science and social studies items found that the greatest improvement in EL scores across versions was for items measuring more basic knowledge and skills. Much smaller differences were noted for ELs on the adapted versions vs. the traditional when the items measured more cognitively challenging constructs. Carr noted that the language adaptations used in the more complex

content items still resulted in greater linguistic complexity than language changes on the items measuring more basic abilities, even though relevant static visual and format adaptations were added. Non-text features appeared to be insufficient to provide enough support to successfully offset the increased linguistic complexity. In her discourse analysis of a middle school science classroom, Wright (2008) found ELs tended to use 'muddled' language while in the throes of learning challenging science content even as they demonstrated facility with more sophisticated academic language in general. She argues that using less sophisticated language and other supports on tests of challenging content may likely provide more valid data on students' science abilities. In their large cognitive lab project Chen and Yi (2005) reported larger effects of traditional vs. language adaptated versions in more rudimentary items for elementary ELs than for more challenging items.

Cultural Features The interrelationships between home and majority cultures, convergent and divergent expectations, and ongoing experiences in and between these cultures within classrooms and socially are but a few examples of why and how the literature suggests cultural features impact score meaning for non-native speakers (see Basterra, Trumbull, & Solano-Flores, 2011; Gee, 2002). Unfortunately, most of the process studies did not specifically focus on cultural features, but rather look at these features within the context of the linguistic features as discussed above (e.g Martiniello, 2008; Noble et al, 2015; Noble et al., 2012; Prosser & Solano-Flores & Li, 2006; Solano-Flores, 2011). In an exploratory systematic language review of DIF items by two French-speaking education experts, Ercikan, Roth, Simon, and Sandilands (in press) note that the nature and frequency of access to the mainstream culture outside of school may contribute to differential score meaning. In her observational analysis of assessment items and answers from majority and minority language speakers in a middle school science classroom, Logan-Terry (2011) found that native English

speakers were able to more successfully notice subtle contextualization cues in test item prompts than emergent bilingual students. She noted that culturally-nuanced details of assessments, such as sequencing of questions and cultural understandings of visuals all contributed to miscues by emergent English speakers. Mann and Emick (2006) described findings from interviews with teachers about parents of new, non-English speakers. For those with limited exposure to mainstream U.S. schooling, many parents didn't understand the purpose or nature of formal and regular classroom testing in U.S. classrooms or of different question types. Teachers said that parent misunderstandings negatively affected the way a number of their students interacted in at least some of the testing situations, and that the students' confusion negatively impacted their test scores as compared to what the students had exhibited otherwise in the classrooms.

Multisemiotic Features Multisemiotic communication is the use of multiple communication modes, signs or representations to convey meaning, rather than relying primarily on the modes of language, especially the language of the majority culture (Peirce, 1931-1958). Most classrooms are inherently multimodal today and the multiple modalities have been considered part of best practice in teaching ELs, especially when more challenging conceptual and reasoning skills are being learned. A small number of projects have focused on how other modalities might be used together with written text to successfully convey meaning to non-native speakers in assessment. While Wright (2008) and Logan-Terry (2011) found rich multisemiotic communication during teaching, they noted these non-text representations, if used at all in assessment, were almost always used in an auxiliary position to language to convey the meaning of concepts and skills. Kopriva et al. (2007) researched the links between item features and EL needs, investigating how specific multi-mode adaptations geared to the students' needs (such as static visuals tied to verb phases and low

frequency nouns) affected their performance. They found that student scores significantly increased when they used representations associated with student need, but scores didn't significantly change when the alternative modes weren't needed. Solano-Flores (2010) and Solano-Flores and Wang (2011) found important cultural differences associated with illustration features for Chinese students. Carr (2008) inspected the features of the visuals in access items, for ELs who performed better, worse or similarly on the traditional and adapted items, and relative to their peers with learning disabilities (LD), students with hearing impairments (DHH) and control students. While no distinct score patterns emerged for the controls, for ELs, higher scores on specific access items seemed to be associated with the gestalt of the visual features. For the LD, higher scores appeared more related to the clarity of the individual features, and higher DHH scores seemed more dependent on gestures and facial expressions of the people.

2. Problem-Solving and Response Processes

There seems to be very little process literature that focuses in detail on how non-native speakers approach and conduct problem-solving during tests, or literature that considers how different response opportunities interact with student needs and preferences. However, there is speculation that these factors impact the validity of inferences.

Problem Solving Walqui and Heritage (2009), Wright (2015), and others explain that non-native speakers still learning the academic language of the school while also learning challenging content use multimodal symbol systems and meaning representations to acquire and expand their mental learning maps about the content. These internal multimodal structures will often be non-standard, meaning that students tend to use their home language and the language they are taught in, but use them in incomplete language structures that integrate language with a variety of other schematic representations. Some suggest that these

non-standard learning internal maps may signal differences in problem solving strategies in testing (Gee, 2004; Moschkovich, 2012; Santos, Darling-Hammond & Cheuk, 2012).

In order to evaluate how selection and application of problem solving strategies interact with item comprehension and response for ELs and others, Winter et al. (2006) hypothesized that three response process aspects and the score would reflect a recursive chain relationship: apprehension → strategy → application → response. Investigating the probabilities of using an appropriate solution strategy given a level of comprehension, results from 156 labs indicated that using an appropriate strategy increases sharply as comprehension goes from partial to full for open-and close-ended items, particularly in 5th versus 3rd grade. On the other hand, the relationship between the appropriateness of the strategy used and the accuracy of its application was found to be stronger in 3rd rather than 5th. The recursive regression results indicated that, for both grades, increasing the degree to which students comprehend a task affects the probability that they will select an appropriate solution strategy. Fullness of comprehension also mediated the correctness of the application of that strategy. These distinctions between partial and full comprehension and their relationships to the subsequent processes were particularly strong for ELs versus non-ELs in both grades, and were evident in how ELs responded in their labs to both basic and adapted items. Based on her observations in classrooms, Wright (2008) suggested that traditional text-laden methods of testing may actually inhibit problem solving, or at least inhibit students' use of their problem solving skills in items presented in this way. She observed that the middle-school science teachers often provided multisemiotic learning opportunities while they were teaching, and that these methods not only facilitated learning but also seemed to act as catalysts to deepen or encourage more sophisticated reasoning and problem solving skills. She noted that when ELs were asked to explain their thinking they frequently used multiple

representations to get their points across to their teachers and peers. On the other hand, when these same students took traditional text predominant tests in English, their responses and scores did not reflect the reasoning and skills they had previously exhibited in class.

Response Based on her observations Wright (2008) and Wright et al. (2013) noticed what she felt was a causal link when items more closely mirror how students are making meaning in classrooms. For ELs with low through mid level English language proficiency, this seemed to include allowing non-standard formats and multiple sign systems. When there was a dissonance between how students made meaning during learning and on traditional text-based assessment, she observed this resulted in lower scores (as students translated their non-standard learning and process maps into English text).

In their cognitive labs, Chen and Yi (2005) investigated how elementary ELs interacted with traditional multiple choice and constructed response (CR) math items and their adapted accessible counterparts. They highlighted problems in responding to traditional multiple-choice options because of the difficulties of properly interpreting what amount to English 'shorthand' option phrases. The adapted versions were sometimes able to minimize the shorthand language problems but not always, particularly when the criteria for adapting items did not include a close look at the language of English over and above addressing the literacy level. The researchers also reported that lower, mid level and some higher English proficient ELs more often used non-standard response methods to fully explain what they knew in responding to CR items than did native speakers. The methods included use of mixed home and majority text, home language conventions or phonetics applied to their written English, and greater use manipulatives, diagrams and drawings to express themselves.

Using a draft of a CR scoring guide designed to interpret EL responses (Kopriva & Sexton, 1999), Kopriva and Lara (1997) investigated the effects of the document and a one-

hour training integrated into the traditional NAEP scorer training for hand-scoring CR responses. After training, the participants, about 15 monolingual English speaking scorers not involved in the initial scoring, blindly rescored responses from mostly higher proficient ELs and a random sample of non-ELs to a set of middle school NAEP science items. Comparing the scores from this study with the original scores of the same responses, the researchers found that the scores for ELs from this study were generally higher than their original scores; they also noted score differences for some native speakers. In interviews conducted at the end of scoring, participants said training prompted them to read and analyze the quality of the responses more carefully and to not prematurely judge responses based on their non-standard presentation methods (e.g. interspersing language and drawing or graphical devices, responses having 'organization issues', or responses they referred to as 'colloquial').

3. Discussion

To-date, what does this literature suggest about how to improve intended inferences for this population? As Kopriva et al. (2016) argue, the processes for evaluation of score meaning and producing assessments useful for non-native speakers rest in framing argument paths and collecting sufficient empirical and argument-based evidence linking tests with particular features and conditions that facilitate valid meaning to profiles of students who benefit from these features or conditions. The profiles not only specify needs and levels but also strengths that can be drawn upon to enable valid meaning. The literature above suggests that this process starts with understanding what traditional methods are insufficient, why, and for whom, and when and how the methods lead to distortion of score meaning. It also provides some clues about how to mitigate these problems. These include what kinds of parameters and evidence are necessary to, first, improve the intended inferences of scores for

non-native speakers and, second, to improve a constant score meaning over native and non-native students with varieties of profiles.

Much of the literature outlined above has focused on problematic native language linguistic features and it's impact on intended meaning to and from the students. To some degree non-native speaking students from different languages and cultures share native linguistic challenges and the need to negotiate between their home languages and cultures as they are taught in the majority language. In general, second language acquisition tells us that receptive and productive skills associated with learning more challenging coursework lags behind their understanding and expression of more basic knowledge and skills. These challenges seem to have heightened their compensatory strengths to make meaning in and out of school that in turn allows them to learn more challenging and cognitively complex content than their current levels of native language proficiency might suggest. The students are also heterogeneous, extending beyond their various home languages and cultures to their own temperaments, personal strengths, challenges, interests, and experiences. As students navigate their learning, different preferred compensatory methods and strategies are encouraged or inhibited in a dynamic fashion over time and content areas.

For test developers and classroom teachers, these aspects, as well as others, begin to suggest a framework for improving score meaning and better understanding the knowledge, skills and abilities of non-native speakers. First, for purposes of assessing content, linguistic and content cognitive complexity need to be de-coupled as much as possible. Non-native speakers are learning challenging content—we need to better document how to build assessment opportunities that mirror their successes. Second, it is important of develop parsimonious but representative student profiles that capture the key student characteristics associated with assessing the content-related concepts, reasoning and other skills of those

students. The profile criteria should include both strengths and challenges, and differ in detail for test developers and teachers. Third, students are dynamic learners of both content and the native language, and preferred assessment methods will change over time, over items, probably over content areas, and at different ages. This, coupled with different student profiles, suggests that assessment should have a more fluid, multi-method and flexible quality of presenting and accepting communication to and from students within established constraints of maintaining defensibility. Assessment defensibility over students requires that what is being measured be held constant, and that retention of the same targets needs to be supported by evidence over profiles and for smaller as well as larger components of tests, for instance items. Varying features and conditions should be designed explicitly to impact non-target aspects of the items and not cue responses or lower the complexity of the construct targets across some variations. Finally, improving score meaning for non-native speaking students means that there is an ongoing need for projects that will begin to apply what we know to date about how to improve assessment opportunities for non-native speakers. This includes process studies that contribute, refine, research, and communicate aspects of viable and valid assessment frameworks, profiles, and assessment methodology that works and leads to improved score meaning. In large part this means that ongoing work should focus on how these and other aspects dynamically interrelate, for summative and formative assessment opportunities, and in ways that are feasible, accessible and available.

II. An illustration: ONPAR

Below is a brief explanation of one promising measurement approach developed to explicitly improve academic score meaning of non-native speakers as well as others. Built from the ground up as a way to embrace and address the differentiated needs of students, it seeks to respond to a number of the problems summarized above, the recent literature

associated with needs of various students, and needs to broaden assessment item types that better measure today's challenging coursework in a systematic way. This methodology, called ONPAR, has demonstrated success in using technology-based multisemiotic representations that include but are not limited to native speaker text in order to improve the validity of score meaning to and from students. This approach is discussed here for two reasons. First, ONPAR seeks to mitigate many of the problems raised in process studies for non-native speakers such as those summarized above. In so doing the findings in this section respond to this literature and outline how well one set solutions seem to work to address some of these concerns. Second, critical evaluations of the new techniques using response process studies are outlined here as well, in an effort to expand the focus of methods involved in the ongoing validation discourse going forward.

Developed over the course of three federally funded research grants and private funds, the assessment items and tasks simultaneously use multiple modalities on screens to broaden the inclusion of students who differentially utilize different sign systems in different situations to access meaning. Drawing from linguistics and semiotic theory (e.g. Jewitt, 2008; Kress, 2003; Kress 2010; Kress & van Leeuwen, 2001), ONPAR capitalizes on the affordances of different modalities to create a multisemiotic 'grammar' of assessment design that may allow developers to better communicate to test takers and hear from them in novel ways. In general, the questions ONPAR focus on are ones that measure a variety of skills and depth of knowledge conventionally assessed through tasks requiring substantial language. To convey what questions are asking students to do, the items and tasks utilize representations such as simulations, animations, image rollovers, sound, interactive sequences, and some L1 and L2 text and oral support as needed for precision. Depending on the nature of the questions, students are asked to respond by building, modeling, assembling, categorizing, or

producing relational or inferential explanations using screen stimuli. To date approximately 20 different types of response spaces are used. Supportive elements include the standardization of several aspects of screens within and across tasks undergirding the novel presentation and response screens without overwhelming or confusing students. For instance, consistent color hues and layouts, placement of screen elements, novel onscreen assists, and accessible dashboards are used as well as animated and static visuals, oral English and L1, and various directional nontext rollovers that support text without cueing any particular response. Additional techniques that facilitate communication rely on careful placement of interactive buttons, target questions, and response spaces, pacing, and task introduction (approach and length). Underlying algorithms capture and score responses, conceptual threads and screen interactive processes and strategies in real time and individualized student reports are available immediately. Readers are encouraged to visit the website at www.iiassessmnt@wceruw.org for more information.

The research and feasibility studies investigated the validity of methodology for measuring challenging mathematics and science in elementary, middle school and high school, for English learners (ELs), students with high and lower abilities in the content areas, native English speakers, and students with learning and other communicative and attention disabilities. Across studies 161 cognitive labs (of focal and control students) researched when and how variations in language and other representations can be integrated within and across sign systems to achieve effective and efficient communication to and from students (Wright and others, 2009, 2011, 2013). Three experimental (Kopriva and others, 2011; 2013; under review) and a correlational study (Carr & Borkon, 2012) found that, controlling for content ability in most cases, the focal groups (ELs, low income, and others with literacy, processing, and attention issues) generally scored significantly higher and in preferred rank orderings

using ONPAR as compared to traditional testing methods measuring the same content. Control groups scored more similarly using both methods, which the researchers suggest, underscore that this approach does not artificially inflate scores, and that it reflects a viable and valid method for measuring challenging concepts for these students as well. Two new grants are researching formative ONPAR methodology within classroom-embedded settings.

1. ONPAR Comprehension

The ONPAR labs typically focused on accessing comprehension of test items and providing accessible response spaces. In examining comprehension, the researchers found that even low English proficient students could access the nuanced meaning of most ONPAR items. The labs confirmed that the consistent ordering of task screens and careful placement of information across screens aided students in expecting certain types of information and comprehending the flow of information. Further, non-native and native students in general approached visual representations and animations by themselves and as part of a gestalt with other semiotic elements such as movement, language, or interactive engagement in a similar way as they would read text (left to right, top to bottom). Non-text stimuli were largely successful in serving in a primary position to substantively convey meaning, especially in introduction and problem-building screens. Native speakers and higher English proficient ELs more often asked for English text in addition to the multi modal stimuli to confirm what they were seeing, especially in high school. The lab investigators reported that the semiotic quality of involving virtual movement in items (on the screen versus movement of the test takers themselves) seemed effective in conveying substantive meaning denoting action sequence explanations or over time changes as relevant and germane to the context or target question. Additionally, the lab researchers reported that frequent interactive opportunities were useful in keeping students focused throughout the tasks, keeping them involved in and

curious about solving cognitively challenging tasks with substantial content demand that sometimes unfolded over a number of screens. Items that allowed students the opportunity to interact early on, through moving screen stimuli or asking simple questions, were particularly useful. When the meaning of visual representations was not clear, the labs found that written text, serving as anchorage on the screen, was often a sufficient way to constrain the context or communicate a precise meaning. Text labels of the visual stimuli, usually an individual word or a noun or verb phrase, were sometimes adequate; otherwise, full simple sentences in English worked best (e.g. "This is a number machine").

The first set of labs examined characteristics of written English text in the target questions per se for elementary and middle school science items (Wright & Kopriva, 2009). By systematically varying the amount of written English text used in the questions while holding all other parts of the ONPAR items the same, researchers found that low English proficient students and others on the whole performed best with full, complete, succinct sentences using precise content language as relevant, and with context-relevant, target irrelevant, words or phrases supported by other semiotic representations. Succinct questions were most often possible when the meanings of non-target language were 'learned' on prior problem-building screens. Support rollovers of non-target verb and sometimes adjectival and adverbial phrases, versus individual words were found most useful as opposed to word-by-word supports. To retain the precision of the item questions the native language option was added on these screens, and generally found not to be necessary on other screens, even for lower proficient ELs. One benefit of less language is less to translate, greatly reducing translation error and allowing for multiple language translations of the questions (provided in ONPAR). Overall, lower and mid-level English proficient elementary and middle ELs sometimes used oral L1 to make meaning of ONPAR item questions, sometimes they relied more heavily on non-text

modes, and sometimes they used oral and/or written English. High English proficient ELs also seemed to benefit from these features in high school more so than in lower grades.

2. ONPAR Problem-solving and Response

Problem-solving The 161 labs investigated some strategies students used to address the ONPAR items, but the two current projects will focus more heavily on assessing these processes. In one strand of data gathered in the earlier labs researchers noticed that students linked understanding of ONPAR items to substantive, pertinent knowledge, procedures, and reasoning they learned in and outside the classroom. Further, the level of quality and relevance of such knowledge or procedures was generally consistent with the quality of their responses. Students more often reported that the higher quality understanding of the prior knowledge also seemed to stimulate and activate strategies they would use to build upon for solving the problems, if they knew how to do so. Their repertoire included not only describing different data organization and interpretation methods learned in their classes, but the ability to link the methods with the content specified in the item, and place and use this information within the larger problem-solving context as it was required to reach a solution.

For students who didn't know the target content, their links to the prior knowledge and skills, while relevant, seemed vague and ill-formed, and their strategies seemed to follow suit. Poorly constructed response spaces sometimes further confused these students but didn't appear to usually be the cause of their incorrect answers. In addition, the investigators reported that more and less knowledgeable students used roundabout logic, incorporating outside experiences and related concepts, skills and knowledge; others were more direct. Logic streams also differed but not by content ability—some began conceptually while others began more procedurally. On the other hand, ELs with little knowledge and using primarily guessing or trial and error to respond were most often unsuccessful.

Response As noted above, some authors argue that key to improving the validation of non-native speaker scores can be traced back to the multisemiotic ways they express meaning in the content classrooms. ONPAR's varied response environments range from capturing more basic answers to responses conveying more in-depth conceptual, reasoning or multi-step integrative or interpretive concepts and skills. Many of the response screens use open as compared to hotspot technology and are scored algorithmically so reports are available immediately. Categories of response types include 1) demonstrating concept knowledge by manipulating screen elements; 2) assembling models, or using diagrams or other symbolic forms to represent systems and meta-systems relationships; 3) drawing; 4) manipulating and/or creating graphs; 5) categorizing; 6) filling-in basic to more complex structures, including basic to complex relational and causal chains; 7) predetermined or choice-based statement frames where visuals, symbols and supported item-irrelevant language are placed in syntactic relationships to form explanations or articulate reasoning, from simple to complex; and 8) open response environments with numerical, pictorial, symbols, and/or language response elements students can use to create proofs or otherwise capture their thinking. Most screens are designed so guessing is at a minimum.

It is easy to overcrowd response spaces and interviewers found a number of students, but particularly those with language, literacy or processing challenges, very sensitive to this dynamic. Rollovers de-clutter response screens, as does the use of symbolic forms 'learned' from and connected to meaning in earlier screens. Standardized placement of different response aspects have been found to be important, as is access to work results from prior screen(s) and the ability to go back and forth. In the lab reports Wright and others found that many but not all of the ONPAR response spaces and features were effective. For instance, investigators reported that demonstrated responses, where students move visual response

elements to provide evidence about what they know, were generally found to be successful for both native and non-native English speakers. Student difficulties with this response type mostly seemed to reflect the difficulty of the task, not the difficulty of the response space. Many of the screens use rollovers of non-target symbols, words, phrases, or other screen stimuli, and the researchers noted that different types of students, including lower proficient ELs seemed to use these in order to hear or see meaning expressed in a different semiotic form. Most fill-in response spaces were found to be clear for native and non-native speakers, even when the screen requested several different responses. This was because the screens used standardized symbols, colors or shapes. The biggest drawback to the fill-in screens was if the response contexts were confusing, or unclear visuals, language or symbols were used. As long as the screens were well formatted researchers reported that non-EL and ELs at different English proficiency levels could still readily navigate more complex screens, such as those requiring relational responses and causal chains with greater and fewer parameters.

Statement frames, where students explain, reason or interpret using text as well as visuals and symbols, are designed to provide more or less 'syntactic' structure and direction. Use of conjunctions constrained some sentences; sometimes students choose among different frames to respond. Success of the frames depended largely on clarity of the frames relative to the question, the response elements, and the non-target text supports. Some statement frames used color-coded spaces to signal object or verb positions, which tended to be very useful for less fluent English learners. When graphics or symbols were used to convey meaning across response types, their success was largely due to their universality of meaning. This universality was generally more important than the context in which it was found, and overall students responded well to novel and familiar contexts when the symbols were understood. Numbers and mathematical symbols were the easiest form of symbolic notation for students

to recognize. The broad uses of symbols in tech-based applications outside of ONPAR have standardized more and more symbols, increasing ONPAR's repertoire and their success in items. Arrows were reported to be the most easily misunderstood, as often their meaning had to be inferred by students from their embedded contexts. Researchers have called for more standardization of arrows on ONPAR screens, probably by function and purpose.

3. Closing Thoughts

Used correctly, ONPAR methods appear to be viable tools to improve score meaning for many non-native speakers, and results from random trials in all studies, as well as the response process data outlined above, have largely supported this. In general, the data demonstrate that multiple signs can and do carry substantive academic meaning, and the signs are used to communicate meaning primarily or with equal weight alongside language. They demonstrate non-text representations can carry more cognitively challenging meaning to and from students, and online screens can flexibly make use of multiple symbols targeting the same concept or skill as well, enabling students with various preferences and needs to be accommodated at the same time without being overwhelming or confusing.

ONPAR, however, is only one approach, and there is still much to do to properly understand this and other novel methodologies. Critical response process evaluations of the ONPAR techniques have suggested some of the techniques work, but future studies need to better understand when and for whom? The usefulness of student profiles and what information populates them remains an ongoing question to be addressed by small response studies designed to defensibly distinguish which multi-faceted needs and strengths are essential. Specific validation queries related to the interaction of new methods and students with different profiles need to be hypothesized during design and empirically addressed. The objective of summarizing this one novel approach here is to get this discourse started.

References

Basterra, M.R., Trumbull, E. & G. Solano-Flores, G. (2011). *Cultural validity in assessment: Addressing linguistic and cultural diversity*. New York: Routledge.

Chen, C.S. & Yi, M., 2005, Cognitive lab report. In R.J. Kopriva & R. Mislevy, *Final Research Report of the Valid Assessment of English Language Learners Project* (C-SAVE Rep. No. 259). Madison, WI: University of Wisconsin, Center for Innovative Assessment, WCER.

Ercikan, K. Roth, W., Simon, M. & Sandilands, D. (2016). Tests fair for all linguistic minority students? Validity and fairness of measurement for diverse linguistic minority students. *Applied Measurement in Education.*

Gee, J. P. (2002). Literacies, identities, and discourses. In M. Schleppegrel and M. Cecilia Colombi, eds, Developing advanced literacy in first and second languages: Meaning with power, Mahwah, NJ: Lawrence Erlbaum, pp. 159– 175.

Gee, J. P. (2004). Situated language and learning: A critique of traditional schooling. Melbourne: Common Ground.

Jewitt, C. (2008). Multimodal classroom research. *AERA Review of Research in Education, 32*, 241-267.

Kopriva, R. J., Emick, J. E., Hipolito-Delgado, C. P., & Cameron, C. A. (2007). Do proper accommodation assignments make a difference? Examining the impact of improving decision making on scores for English Language Learners. *Educational Measurement: Issues and Practice, 26*(3), 11–20.

Kopriva, R.J. & Lara, J. (1997) Scoring English language learners' papers more accurately. In Y.S. George and V.V. Van Horne (eds.), *Science education reform for all: Sustaining the science, mathematics and technology education reform* (pp. 77–82). Washington, DC:

American Association for the Advancement of Science.

Kopriva, R.J. and Sexton, U. (1999) *Guide to scoring LEP student responses to open-ended science items*. Washington, DC: Council of Chief State School Officers.

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.

Kopriva, R.J., Wright, L., & Triscari, R. (under review). Examining a multisemiotic approach to measuring challenging content for English learners and others: Results from the ONPAR elementary and middle school science study. *Educational Assessment*.

Kress. G. (2003). *Literacy in the new media age.* London: Routledge.

Kress, G. (2010). *Multimodality. A social semiotic approach to communication.* London: Routledge Falmer.

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

Logan-Terry, A. (2011). Achievement, assessment, and learning: A study of emergent bilinguals. Students in mainstream content classrooms. (doctoral dissertation, Georgetown University, Washington, DC).

Logan-Terry, A., & Wright, L.J. (2010). Making thinking visible: An analysis of English language learners' interactions with access-based science assessment items. *AccELLerate!*, *2*(4), 11–14.

Mann, H. & Emick, J. (2006). Language Liaisons: A novel accommodation for new ELLs. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review, 78*(2), 333-368.

Mislevy, R.J. (1994) Evidence and inference in education assessment. Psychometrika,79–85.

Noble, T., Suarez, C., Rosebery, A., O'Connor, M. C., Warren, B., & Hudicourt-Barnes, J. (2012). "I never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science. *Journal of Research in Science Teaching, 49*(6), 778-803.

Noble, T., Rosebery, A., & Suarez, C. (in press). How do English language learners respond to science test items?: Evidence related to validity. Applied Measurement in Education.

Prosser, R. & Solano-Flores, G. (2011). Examining Problem Solving Strategies on Multiple-choice Science Items Among English Language Learners Through Cognitive Interviews. Paper presented at the Annual Conference of the American Educational Research Association, New Orleans, Louisiana, April 7-11, 2011.

Roth, W.-M., Oliveri, M. E., Sandilands, D., Lyons-Thomas, J., & Ercikan, K. (2013). Investigating sources of differential item functioning using expert think-aloud protocols. *International Journal of Science Education,* 35, 546-576.

Schleppegrell, M. (2004). *Language of schooling: A functional linguistics perspective. Mahwah, NJ: Lawrence Erlbaum.*

Solano-Flores, G. (2010). *Vignette Illustrations as a Form of Testing Accommodation for English Language Learners: A Design Methodology for Use in Large-Scale Science Assessment.* Paper presented at the Annual Conference of the National Council of Measurement in Education, Denver, Colorado, April 29-May 3, 2010.

Solano-Flores, G., & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice 25*(1), 13-22.

Solano-Flores G. & Wang, C. (2011). Conceptual Framework for Analyzing and Designing Illustrations in Science Assessment: Development and Use in the Testing of

Linguistically and Culturally Diverse Populations. Paper presented at the Annual Conference of the National Council on Measurement in Education, New Orleans, LA, April 7-11, 2011.

Walqui A. & Heritage, M. (2009). Instruction for diverse groups of English language learners. Stanford, CA: Understanding Language: Language, Literacy, and Learning in the Content Areas. http://ell.stanford.edu/sites/default/files/pdf/academic-papers/

Winter, P., Kopriva, R.J., Chen, S., and Emick, J. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning and Individual Differences*, 16 (2006), pp 267-276.

Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment, 14*(3), 139-159.

Wright, L. J., & Kopriva, R. J. (2009). Using cognitive labs to refine technology-enhanced assessment tasks and ensure their accessibility: Insights from data collected to inform ONPAR elementary and middle school science task development. Madison, WI: Institute for Innovative Assessment, University of Wisconsin-Madison. Retrieved from http://iiassessment.wceruw.org/research/researchpapers/onpar.

Wright, L.J. & Staehr-Fenner, D. (2011). Results from ONPAR mathematics cognitive labs. Retrieved from http://iiassessment.wceruw.org/research/researchpapers/onpar

Wright, L. J., Staehr-Fenner, D & Moxley, K., (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. Retrieved from http://iiassessment.wceruw.org/research/researchpapers/onpar

Wright, L. J. (2008) Writing science and objectification: Selecting, organizing, and decontextualizing knowledge. *Linguistics and Education*, *19*(3), 265–293.