Implications of Technology for Assessing ELLs

In an article in the New York Times in March, 2013, Joel Klein, former chancellor of New York City schools argued that "…if we didn't see a dramatic technological change (in K-12 public schools) we were not going to be able to move this country forward (Chozick, 2013)". Embracing technology in large-scale assessments includes using technology to expand how items and performance tasks can be designed and built, adaptations (including accommodations and item variations) can be assigned to students, tests can be administered, and algorithm-based scoring and reporting can be integrated into the testing infrastructure. This is a very exciting time in testing of all students, but particularly those such as ELLs who have had problems accessing the content of items in traditional paper-and-pencil testing contexts, or being able to respond so their answers will reflect their knowledge, skills and abilities. For the testing community the challenge will be to use this opportunity to address, up front, the constraints ELLs faced in the past by making use of the *expanded capacities* available through the use of technology. This also includes being sensitive to and addressing unintended consequences for ELLs as assessment developers work with this new technology to produce tests for the general student population. Four areas of test development and implementation seem particularly relevant: (1) addressing the range of item complexity, (2) expansion of response environments, (3) accommodation implications for the future, and (4) consideration of student characteristics.

Addressing the Range of Item Complexity. As defined above, linguistically accessible items are those that present the focal content and response environments clearly with no ambiguity in the language. In an article for the Applied Measurement in

Education Journal, Abedi (in press, 2013) argues that one advantage of computer adaptive testing, is that more linguistically dense items could, in principle, be swapped out for accessible items with less linguistic density that measure the same content at the same cognitive demand level for some ELLs. Cook and McDonald (2012) have introduced a tool to evaluate the linguistic complexity of items, and at least one of the multi-state assessment consortia have expressed interest in using it in some way to differentially tag their items (Cook, 2012, personal communication). Certainly, technology's capacity to easily direct more accessible items versions to students who need them seems like a plus. It is important to keep in mind, however, that discourse studies and cognitive linguistics consistently indicate that the text in more challenging content items typically involves more complex language usage and structures to handle the elevated linguistic demands (for instance see Schleppegrell, 2004; Talmy, 2003). So, the questions here are: 1) If ELLs will be tested only or mainly on items commensurate with their level of language proficiency skills, is the same level of content complexity really held constant, especially for students who do not have the formal language required for that type of item? If this is not the case, 2) in their CAT algorithms, how is SBAC planning to guard against truncating the cognitive complexity of test questions for lower ELLs when it may be language rather than content that is causing students to answer more challenging questions incorrectly? How is PARCC planning to address the issue for these same students? 3) Are *any* of the current language accommodations sufficient to effectively minimize the effects of more sophisticated language structures and non-relevant but challenging terms in more complex content items while still conveying the full level of content meaning to ELLs that are not yet proficient in this level of academic language? Kopriva & Mislevy (2005), Carr (2009), and Cawthon, Lippo, Carr, & Kopriva (2013) have suggested that simplified language and typical ancillary supports may often not be sufficient or adequate for lower English proficient ELLs, especially when the items are more challenging. To date, other research has not focused on the interactions of cognitive complexity in items and the effectiveness of current test accommodations.

It does not seem to be in ELLs' best interests to be tested on only more basic items where meaning can most often be with conveyed primarily with simplified language (with aforementioned accommodations as supports), while the general test takers are being measured on a fuller range of content knowledge, skills and abilities.

Kopriva and others (2009a, 2011a, 2013a) demonstrated in a set of experimental studies that retaining the content complexity seems to be possible for less proficient ELLs and some students with disabilities, if semiotic representations beyond text explicitly carry essential portions of the non-target relevant meaning needed to capture the nuances required in the challenging content. The methodology they have used, called ONPAR, builds multi-semiotic task-specific problem contexts and surrounds, uses movement via animation and simulation, ongoing interactions between students and the computer screens, multiple redundancies, auditory as well as on-screen elements, and targeted text as needed for precision. But this requires careful identification of non-target relevant elements in items and careful planning in how the meaning is conveyed. Further, it often includes extending how students are allowed to respond.

On a related front, Kopriva and Winter (2012), along with a small study group of seasoned assessment professionals, recently completed a paper arguing that content complexity seems to be inadequately captured by identifying just the traditional depth of knowledge rubric in items. In this case, developing items with the same level of cognitive demand may misrepresent the actual content complexity of items and mislead the compatibility of items sharing the same cognitive demand. Extending the work of Shavelson, Ruiz-Primo and colleagues (2007, 2011?), among other researchers, the study group proposed a two dimensional *matrix of item complexity*. On one dimension were levels of *cognitive demand*, and on the other were levels of *content demand*. In this instance, cognitive demand followed a Webb-type structure and content demand was captured in terms of the kinds of questions that get at different types of knowledge, such as declarative, procedural, systematic and schematic. Luecht (cites) has also focused on what it means for the content in the target item construct to be more or less complex as he re-considers automatic item generation algorithms using some of the same ideas.

As Abedi pointed out in his AME article, ease of manipulating among compatible items using algorithms that undergird testing systems and selecting the proper item option for students with particular issues seems to be a key component of the future of content testing for ELLs. Embracing the idea of an accurate way to categorize the rigor of complexity in items is still a work in progress. For ELLs, the challenge, it seems, will be to expect a linguistically accessible version that does not 'dumb down' the test by focusing only on linguistic density to the exclusion of other item aspects that may be necessary to keep the content complexity of more rigorous items consistent across test takers.

Extension of Response Environments. In the past, some hand-scored paper/pencil constructed responses allowed students to draw, diagram, code-switch and/or use L1 in addition to responding in written English. Shaw's work, summarized in Shaw, Abedi and Kopriva (in press, 2013), suggests that this freedom seems to be at least partially responsible for narrowing the score gap for ELLs and others in performance tasks. As electronic scoring of constructed responses have grown in popularity, one consequence has been that, as of now, almost all tests require responses using only text and only in the language of the test. This constraint may allow for easier scoring overall but the unintended consequence hits lower English proficient ELLs quite hard. The problem arises when students *cannot* explain themselves with only text. Some researchers (for instance, Wright and Logan-Terry, 2013) have found that many lower ELLs who could not convey meaning even if they used code-switching and L1 text in addition to rudimentary English. It would seem that accessibility scoring reviewers for both consortia would want to address the substantial barriers associated with this problem.

A related unintended consequence that severely hampers the automatic scoring of written responses come from those whose formal written English is not yet developed to be consistent with the general population (or formal written L1 if the test is in L1). It would seem that the markers algorithms use to differentially score the items, perhaps even short answer constructed response items, will be somewhat to very different for most ELLs. While certainly flags might be used to identify written responses that need to be set aside for hand-scoring, accessibility reviewers in both consortia will need to be able to evaluate how sensitive the flags are to specific non-standard language usages of ELLs. Lara and Kopriva (1997) found that training monolingual hand-scorers in how a range of ELLs might explain themselves appeared to improve how ELL responses were hand scored. It would seem that training tips such as those used in this study could be valuable in building into scoring or flagging algorithms.

A large advantage of online testing is that it can handle a larger set of item types. In their technology-based classroom science tasks, Quellmalz and others (see, for example, Quellmalz, Golbert, & Clarke-Midura, J., 2012; Quellmalz, Clarke-Midura, Kopriva, Golbert, & Davenport, 2011) have successfully investigated the viability of classroom technology-based performance tasks for the general student population that include animations, student driven simulations and even virtual reality, in addition to text, as they present coherent, multi-layered and challenging problems for students to address. Further, linking research materials and other stimuli to selected items has been explored. As noted above, Kopriva and others have shown that many of these techniques can be useful in minimizing problematic language in the presentation of questions as well.

However, for many ELLs, the extension of response environments isn’t just for challenging items and tasks. Just as when constructed response items used to allow for drawing diagrams and pictures, we argue that extended response environments are important in a wide range of items for a number of ELLs because it broadens how they can respond about what they know.

Classroom teachers have long been aware of incorporating extended communication options for ELLs into lessons and evaluations (see Trumbull and Farr, 1997, for examples). To date, some assessment researchers have studied a few limited ‘technology-enhanced’ response methods (for instance see Winter et al., 2012). Kopriva and others (2009a and b, 2011a and b, 2013a and b) have investigated a more extended set in the three ONPAR projects and 156 cognitive labs. Essentially, by expanding beyond a reliance on language that is either text-heavy or language ‘shorthand’, ELLs can explain themselves and their knowledge, skills and abilities by, for instance, engaging with stimuli and dragging and dropping, building, assembling, and manipulating connections in ever-changing environments. This has been shown to be the case with numerous multiple-choice items as well as an increasing percentage of constructed response items. Scoring algorithms undergird these extended response types (visit http://onpar.us to view examples of the response environments in sample items and testlets).

These projects have also piloted the blank statement frame item type where students complete the frames by choosing among a wide set of supported, non-construct-relevant text, symbols, pictures, conjunctive and causal connective devices, and so on in order for students to explain reasoning, rationales, justifications or meta-cognitive processes. All frames are internally scored using algorithms that track responses and, for process scores, sometimes what choices were discarded or how frames were completed. When appropriate, some frame elements are color-coded by parts of speech so scoring rules can utilize the English language structure as students are learning it’s conventions. Statement frame versions also include those where students choose among various statement frame configurations to complete their explanations, and those where blank frames structure basic or complex relationships students are expected to complete.

Accommodation Implications for the Future. Technology’s impact on the form and administration of accommodation tools has largely been covered in Abedi’s 2013 AME article. The nature of the accommodations in an online environment are consistent with what is noted above. However electronic administration and how the accommodations are interpreted (for instance one bilingual glossary or set of bilingual pop-ups will be used for a test administration) are standardized across students, classrooms, states, etc. As assessment professionals consider how to integrate and implement the accommodations within online academic testing systems, however, a couple additional possibilities, beyond utilizing ONPAR techniques, are worth noting.

First, in their study with ELLs whose native language is Spanish or Arabic, Sireci and Wells (2010) found success in boldly connecting item text written in English with test administration using oral L1. This 'mismatch language' approach is not often considered but may be worth future study because of its apparent success. This may be true especially for low and mid level ELLs for students taught in English who may or may not be literate, but who still need more dense support from their L1 than glossaries seem to be able to give. Conceptualized as a version of a dual language test, this approach may be worth considering. Second, in a number of instances, the language ELLs find confusing in items are actually phrases, tense or word variations not easily found in glossaries or dictionaries, or words that take more difficult language to explain than ELLs may be able to read or understand (see Basterra, Trumbull, and Solano-Flores, 2011). Extending some of these language links would seem to be useful. Further, some work has been done re. pop-ups or roll-overs that can utilize the types of multi-semiotic techniques mentioned above to accommodate a broader range of language in the online glossaries. Recent reviews of plans within both assessment consortia suggest some of these language elements will added to their online English and bilingual glossaries and they are considering the use of some visual pop-ups (Chia, 2013; Powers & Strain-Seymour, 2013). *Consideration of Student Characteristics.* A number of authors have argued that more needs to be done to disentangle crucial ELL student, language, and environment characteristics and their interactions because current testing methods are still distorting knowledge, skills, or abilities of many students. For years, Solano-Flores (e.g. 2006, 2008, 2009) has emphasized that the relationships among linguistic elements and socio-cultural characteristics for most English language learners are dynamic. He links these ongoing changes to fluctuations in how most students perceive the content and cognitive demands in items, including variations from item to item. Further, he suggests that (a) formal language features, in and of themselves, do not consistently reflect a particular level of English proficiency, (b) adaptations to linguistic elements do not definitively increase accessibility in a certain way, and (c) language characteristics cannot be effectively considered without the broader contexts of student strengths and other related variables. Solano-Flores (in press, 2013), Ercikan, Roth, Simon, Sandilands, and Lyons-Thomas (in press, 2013), and others focus on the impact of L1 and the relationships between it and the language of schooling as well as the influence of the immediate environment students frequent. Noble, Rosebery, and Suarez, (in press, 2013) address the ongoing vulnerability of students who have recently exited from English language services, and Ercikan et al. and Roth, Oliveri, Sandilands, Lyons-Thomas, and Ercikan (in press, 2013) argue that dissimilar DIF patterns may reflect different forms of reasoning.

These researchers and their work highlight a number of compelling characteristics which need to be addressed as possible in classroom through large-scale testing. To-date, however, no comprehensive sets of clearly articulated variables have been definitively coalesced for various assessment purposes, and a clear sense of how the variables might be prioritized or applied is not fleshed out. This call for more a rigorous but nuanced approach to content testing is not limited to addressing the needs of ELLs. Cawthon et al. (in press, 2013) have recently highlighted this issue in their evaluation of accessible item adaptations for students with learning disabilities, deaf and hard of hearing students, and ELLs, all of whom who share language and/or literacy challenges but where the limitations originate from very different etiologies.

Two places where it seems that incorporating more nuanced considerations of ELLs into testing would lead to substantially improving the validity of the score interpretations are in (1) increasing the number and types of interchangeable item/accommodation variations to meet the needs of different students, and (2) building and using matching systems that identify and use levels of salient student characteristics and then link them to appropriate item/accommodation variations. Technology has the capacity to handle a wide variety of assessment design algorithms that can be integrated into assessment systems in order to capture a greater range of adaptations at the item level. If conceptualized correctly, it appears that these adaptations can facilitate strengths and minimize construct-irrelevant interference at a more nuanced level. The challenge will be to design adaptation schemes that meet the ‘comparable enough’ standards being discussed by researchers (for instance see the studies compiled in Winter, editor, 2010).

There has been some work focusing on how differentiating student characteristics might be applied to assigning accommodations. Rivera and Collum (2008) laid out guidance arguing how English language proficiency levels may impact student needs. Kopriva, Koran, and Hedgspeth, (2007) and Kopriva and Koran (2008) identify 8 student variables, and 4 domains for both L1 and English, and outline how the levels of each of these 16 characteristics are differentially applied to capture dozens of distinct ELL student profiles. To-date, this work has been used to link each specific profile to their proper accommodations. Koran and Kopriva (2006) and Kopriva, Emick, Hipolito-Delgado, and Cameron, (2007) are two studies that were completed to validate the accommodation linking system, called STELLA, along with a number of focus groups, interviews, and expert algorithm reviews. Kopriva and Koran (2008) reviewed the literature examining attempts at systematizing accommodation guidance for ELLs and students with disabilities to-date, and Kopriva and Lara (2009) discussed the challenges in systematizing accommodation guidance for NAEP testing over the last 20 years. In her paper, Carr (2009) used the differential student characteristics and their levels to compare and contrast the types of ELLs who would receive different accommodations under the STELLA system.

Next steps would seem to be for experts to convene to refine and vet relevant sets of characteristics and variables that might be used for different assessment purposes. This group might then propose and prioritize relevant research agendas. As findings from this work are available, the capacities of technology can be applied to advance the more nuanced validity program for ELLs.