Issues and Considerations for the Second Generation of Researchers Interested in Assessing ELLs on Academic Content: A Review of Four Papers

Rebecca J. Kopriva
2-8-13

Rigney, Wiley and Kopriva (2008) trace the evolution of academic assessment over the last century, summarizing the purposes and issues, including fairness, inherent in ongoing test theory and academic test development. In the 1920's multiple-choice methodology was heralded as a valuable solution to concerns about fairness because of its standardization in the questions and objectivity in scoring. Differences among students were not considered salient at this point. Translation considerations and student representation in field tests consistent with population percentages were among the primary nods to student access over the next 75 years or so. In the 1990's exploration of 'authentic assessments' argued that variation in response could be judged to yield the same score for the same item (for instance see Resnick, 1987)). Mislevy's address published in *Psychometrica* in 1994 discussed what has become Evidence Centered Design (ECD). It stressed that the argument for comparability of common academic test inferences could be made on conceptual grounds rather than the procedural grounds that had led to the requirement for common products and testing conditions. His conceptual argument focuses on developers providing adequate evidence about the target knowledge and skills, necessary observations, properties of tasks to elicit the observations, and assessment situations. One important implication is that data may be collected under alternate conditions, as long as there is proper documentation that the evidence yields inferences about the same targeted constructs.

The first generation of research on accessibility of academic tests primarily flowed from the nationwide need to hold schools accountable for teaching the same rigor of content to *all* their students, including language minority students identified as English language learners. Researchers interested in accessibility in academic tests for diverse student groups built upon ECD, investigating which challenges lead to distortions in valid inferences for some students, and how to build and/or adapt assessments so that the defensibility of valid common inferences across students and condition variations could be sustained. Rudimentary large-scale accommodations for ELLs were empirically and pragmatically distilled over time. Over the last 7 years researchers argued that ELL accommodation needs seem to be informed by student background factors including, but not limited to, English language proficiency (ELP) levels. For a number of reasons the assessment field has been slow to adopt this advice when it comes to systematically matching students with certain needs to proper accommodations or adaptations, even though the literature abounds with evidence that current methods are not working very well.

The papers for this review (Solano-Flores, Abedi, Ercikan et al., Noble et al., and Oliveri, et al.) highlight some of the key issues outstanding as academic assessment accessibility research for linguistic minorities enters its second generation. As the authors suggest, their recommendations are relevant for both large-scale and classroom testing. In all, the papers make many excellent points, but three in particular stand out for comment.

First, most of the papers argue that more needs to be done to disentangle crucial student and environment characteristics connected with the language minority population because current methods are still distorting inferences about these students' knowledge, skills, or abilities. The contexts in their papers were different, but the message of the importance of understanding how certain linguistic, language development, environment, and student background characteristics interact with aspects of items seems constant.

Solano-Flores' theoretical paper underscored the assumptions that appear to be shared by most authors. In arguing that the relationships among linguistic elements and socio-cultural characteristics for most language minority students are dynamic, he links these ongoing changes to fluctuations in how most students perceive the content and cognitive demands in items. He also suggested that the interactions between items and student features may often vary from item to item. Further, formal language features, in and of themselves, do not definitively mark one level of proficiency or set of background attributes versus another, and cannot be effectively considered without the broader contexts of student strengths, factors influencing differential meanings, and other variables. Ercikan et al. pointed to the impact of language at home and in the immediate environment the students frequent, and also tackled the topic of the Francophone language minority population and diverse countries of origin of their parents. Noble et al. focused on the ongoing vulnerability of students who have recently exited from English language services considering a variety of interactions of content demand, linguistic aspects and student background. Oliveri et al.'s simulation study highlighted that amount of language proficiency differently impacts identification of differential item functioning in items.

The papers considered here neither collectively identify a set of clearly articulated variables that impact access, nor how those discussed here might be prioritized or applied. Rather, the strength of these articles seems to be that, moving forward, those interested in accessibility and valid inferences need to think in a more nuanced fashion about how to better measure the academic knowledge and skills of language minority students. The authors discussed a wider scope of academic testing beyond just large-scale assessment, and certainly classroom assessment expands how various characteristics might be better considered. Solano-Flores mentioned some work he had undertaken in schools where teachers are trained to incorporate more differential information in how they pose questions to their diverse students and how they provide situations so the students could tell them what they know. On the whole, however, the authors appear to be setting the stage for future research where these crucial questions can be specified and investigated. Of interest, this call for more nuance in content testing for diverse students is not limited to language minority accessibility researchers. Cawthon et al. (in press) have recently highlighted this issue in their evaluation of accessible item adaptations for students who share language and/or literacy challenges but where the limitations originate from very different etiologies (ELLs, students with learning disabilities, and deaf and hard of hearing students).

Second, to prepare for large-scale and classroom assessments where the dynamic interplay of multiple features can be made an integral part of how questions are asked and answered for students, assessment techniques, as a methodology, will most likely have to be expanded to handle the increased variations. How might this occur? Abedi commented that many large-scale accommodations, well known from paper-and-pencil days, might be made more standardized and easier to provide in online contexts, thereby increasing the likelihood that students would receive them, especially those who differentially need them. However he also mentioned that some ELLs might have problems responding to constructed response items. Right now, one implication of the constraint that almost all open-ended responses need to be typed is that it restricts English learners from drawing graphics or using mix-media to make their points. Further, computerized scoring algorithms at this point do not seem not be sensitive to detecting meaning as the student intended when their written responses include code-switching or immature language, phrase or sentence constructions. Under computer adaptive testing, Abedi poses that simplified language versions of items that are less linguistically dense might be made available to students with less English proficiency. This could be a plus, but it will be important that the cognitive demand of the simplified versions remain consistent with that of the general version of the item. This is easier to do when item targets are more basic, but much more difficult, if not impossible, to do with more challenging content unless other item adaptations beyond just simplified language are incorporated.

The underlying capabilities of computer- (or tablet-) based testing could build more complex algorithms of item variations that consider more student variables and item techniques than those in use today. It appears that an empirically grounded framework web needs to be systematically and coherently outlined for language minority students and needs to contain specific elements. These include a) which student variables are most salient for signaling different item versions and under what conditions, b) which item techniques need to be considered for what particular outcome, c) how can arrays of item variations be methodically built that properly use a principle-based set of techniques to allow each version to conform to the same or similar intended content meaning at the same cognitive level of complexity, and d) what are the nuts and bolts of how items arrays might connect to profiles of students that are sophisticated enough to handle the nuances such as those discussed in this Issue, so that e) a system can be built that will effectively match items with particular features to students with particular needs and backgrounds. Pieces of this web are being developed. Undergirding most of the papers in this Issue is a stream of work by the authors where they are ferreting out particular student variable connections, and/or particular item variation techniques. Solano-Flores's guidance to teachers regarding dialect adaptations and other features is one great example, as is, I expect, the sociocultural and linguistic protocols behind Noble et al's item evaluations. Ercikan et al.'s findings that there were NO common DIF items across different groups of language minority students, suggests a basis from which to work. Further, Ercikan et al. pointed out that linguistic meaning-making can be quite substantially mediated by environmental influences of the dominant language and culture, and referencing Oliveri and her colleagues (Roth et al., in press), Ercikan et al. posited that some dissimilar DIF patterns may reflect different forms of reasoning. Such student variables would seem to be crucial for understanding how meaning remains constant across language minority groups, much less the rest of the general population.

Sireci and Wells (2010) have found success in boldly connecting written item text in one language with oral in another. Our work (Kopriva and others) uses some targeted text, oral in a number of languages, and assessment techniques with multiple redundancies. These techniques are coupled with screens that spend time building multi-semiotic task specific problem contexts and surrounds, and methods of using movement via animation and simulation as well as ongoing interactions between students and the computer screens. All of these aspects in the ONPAR methodology promote involvement for purposes of increasing meaning in ways beyond just text. Novel response spaces and techniques also have been piloted that allow students without much literacy to show or tell us about what they know. August et al.'s article (in preparation) explores the theoretical underpinnings of how challenging content can be assessed in the classrooms of English language learners. This paper would seem to help undergird the formative portion of the framework, and most likely, would be informative for large-scale testing as well.

Third, more nuanced variations and novel methodology will need to be properly evaluated as well, both in terms of validity of inferences within the language minority omnibus group, and comparability between interpretation of scores of language minority students and the general student population. Validation methodology, such as item evaluations by experts and cognitive labs with students discussed by Noble et al., generalizability methodology referred to by Solano-Flores, randomized experimental trials noted by Abedi, as well as other methods such as those discussed in the multi-volume text *The APA Handbook of Testing and Assessment in Psychology* (in press, 2013) refer to procedures for planning and implementing well-conceived studies which can lead to defensible inferences and add to the validation evidence of particular relationships. As more sophisticated sets of variables and item methods are considered, the researchers here seem to be very aware that systematic research agendas designed to empirically investigate hypotheses and rule out alternative explanations need to be undertaken. Perhaps the framework elements in the prior comment could be extended to methodically outline the empirical work that would attend each of the linkages in the web. Without logical arguments and empirical evidence of this caliber, it most likely will be difficult to convince the measurement professionals that this type of more fine-grained academic assessment is warranted in order for inferences to accurately reflect knowledge and skills of the language minorities.

Ercikan et al. and Oliveri et al. use a few differential item functioning (DIF) methods to investigate measurement comparability. I'm sure the researchers would agree that measurement comparability is more complex than determining if items behave differently across groups, although this type of technique has its place. Ercikan et al. allude to a fuller definition of measurement comparability in their discussion, and it seems that making an argument for multiple types of evidence would be essential in determining if and when the test scores of the heterogeneous language minority population are producing comparable and valid inferences internally and compared to other student groups. Kopriva and Albers (in press) summarize some of the recent literature regarding this topic. In general, a broader understanding of measurement comparability seems to be in terms of providing evidence of equivalence in content/cognitive demands and in terms of score equivalence. For instance, documentation of score equivalence could include evidence that scores are behaving in the same (enough) way for students with the same abilities, evidence of similar score distributions, and evidence of similar internal structures of tests for different groups of students. The first chapter in Winter (2010) outlines the underpinnings, measurement context, function, and types of arguments associated with this broader sense of comparability. The rest of the book summarizes some of the promising methods a set of researchers used to evaluate comparability for various purposes and contexts, most of which were related to investigating comparability for diverse student groups in relation to the general population of test takers.

Oliveri et al. noted that higher levels of simulated DIF led to lower correct detection rates in her simulation study. She surmised that this trend might be related to the degree of contamination of the matching variable that is the total test score in almost all DIF methodology. I repeatedly have made the argument that DIF techniques underrepresent differential item detection for many ELLs (especially those with lower and mid-levels of English proficiency) because the issues of language and literacy tend to be pervasive to some or greater degree across most items in a large-scale academic tests (see Kopriva, 2008, pgs 298-304). As Oliveri et al. and Erickan et al. point out, item specific issues such as language more likely to lead to differential meaning, unusually complex language load, or certain sociocultural referents may still signal DIF detection in any case, and may help explain Ercikan et al.'s results regarding no common items with DIF within the language minority population. As we move forward, it seems that those of us interested in identifying differential item functioning for students with language and literacy challenges will likely have to develop a method that is more defensible for our purposes.

The questions and considerations of the four articles reviewed here strongly suggest that the next generation of accessibility research promises to be insightful and fruitful for improving academic testing of language minority students. My concluding thought for these authors is to consider how we might generalize this thinking to improving testing for the general population of test takers as well. It is hard to believe that it is as homogenous as we might simulate them to be, and while we may be helping to lead the field towards more differentiated assessment for ‘our’ students, there may come a time when the measurement community or their consumers ask for a more nuanced version of content testing for the ‘not identified’ students as well.

References

Abedi, J. (in press, 2013). The use of computer technology in designing appropriate test accommodations for English language learners. In *Applied Measurement in Education*.

August D., Valdez,G., Heritage, M., Herman, J., Kopriva, & Bailey, A. (in press, 2013). Instructional Strategies and Tools. Accepted for Special Issue of Educational Assessment.

Cawthon, S., Leppo, R., Carr, T., & Kopriva, R. (in press, 2013). Towards accessible assessments: The promises and limitations of test item adaptations for students with disabilities and English language learners. *Educational Assessment.*

Ercikan, K., Roth, W-M, Simon, M., Sandilands, D, and Lyons-Thomas, J (in press, 2013). Tests fair for all linguistic minority students? Validity and fairness of measurements for divese linguistic minority students. In *Applied Measurement in Education*.

Kopriva, R.J., Gabel, D., and Cameron, C. (2009). *Overview of the ONPAR Elementary and Middle School Science Experimental Study with ELs and non-ELs: A Promising New Approach*. At www.onpar.us.

Kopriva, R.J., Winter, P.C., Triscari, R. Carr, T.G. (2013). *Assessing the Knowledge, Skills, and Abilities of ELs, Selected SwD's 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.* At www.onpar.us.

Kopriva, R.J. (2008). *Improving testing for English language learners: A comprehensive approach to designing, building, implementing, and interpreting better academic assessments*, pgs 298-304. New York, NY: Routledge Publishers.

Kopriva R.J. & Albers, C. (in press, 2013). Considerations of achievement testing for students with individual needs. In *APA Handbook of Testing and Assessment in Psychology*.

Mislevy, R.J. (1994). Evidence and inference in educational measurement. *Psychometrika*, 51, pp 177-195.

Noble, T, Rosebery, A. & Suarez, C. (in press, 2013). How do English language learners respond to science test items? Evidence related to validity. In *Applied Measurement in Education*.

Oliveri, M.E., Ercikan, K. & Zumbo, B.D. (in press, 2013). Effects of population heterogeneity on accuracy of DIF detection. In *Applied Measurement in Education*.

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

Resnick, L.B. (1987). *Education and learning to think. Committee on Mathematics, Science and Technology Education, Commission on Behavioral and Social Sciences and Education.* Washington DC: National Academy Press.

Rigney, S., Wiley D.E., and Kopriva, R.J. (2008). The past as preparation: Measurement, public policy and implications for access. In R.J. Kopriva, *Improving Testing for English Language Learners: A Comprehensive Approach to Designing, Building, Implementing, and Interpreting Better Academic Assessments.* New York, NY: Routledge Publishers.

Sireci and Wells (2010). Evaluating the comparability of English and Spanish video accommodations for English language learners. In P.C. Winter (Ed.), *Evaluating the Comparability of Scores from Educational Achievement Test Variations*. Washington, DC: Council of Chief State School Officers.

Solano-Flores, G. (in press). Probabilistic approaches to examining linguistic features of test items and their effect on the performance of English language learners. In *Applied Measurement in Education*.

Winter, P.C. (2010, Editor). *Evaluating the Comparability of Results from Educational Achievement Test Variations.* Washington, DC: Council of Chief State School Officers.

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.* At www.onpar.us.

Wright, L.J., & Carr, T. (2011) *ONPAR Elementary and Middle-school Mathematics Cognitive Lab Results: How Students Interact with Traditional and Dynamic Assessment Tasks with Novel Features.* At www.onpar.us.

Wright, L.J., Staehr-Fenner, D., Moxley, K., Kopriva, R.J. & Carr, T. (2013). *Highlights from Cognitive Labs Conducted with ONPAR End-of-course High School Biology and Chemistry Interactive Assessment Tasks: Knowing What Diverse Learners Understand.* At www.onpar.us.