Do linguistic features of science test items prevent English Language Learners from demonstrating their knowledge?

Tracy Noble[1]

Rachel Kachchaf[2]

Ann Rosebery[1]

Beth Warren[1]

Mary Catherine O'Connor[3]

Yang Wang[4]

[1] *TERC*
[2] *Smarter Balanced Assessment Consortium*
[3] *Boston University*
[4] *Education Analytics*

Paper presented at the Annual Meeting of the National Association of Research on Science Teaching, April, 2014, Pittsburgh, PA

Abstract

Little research has examined individual linguistic features that influence English language learners (ELLs) test performance. Furthermore, research has yet to explore the relationship between the science strand of test items and the types of linguistic features the items include. Utilizing Differential Item Functioning, this study examines ELL performance on 162 Grade 5 large-scale science multiple-choice test items and its relationship to two discourse features, as well as the distribution of these features across three science strands. We also interviewed 52 ELLs to examine their interaction with these two features. Results indicate that these two features were most frequent on items under the Life Science strand. Additionally, both of these features were significantly correlated with DIF disfavoring ELLs indicating that the presence of these features potentially hinder ELLs' abilities to understand test items containing these features. Interviews confirmed that these two features in combination interfered with ELLs' abilities to make sense of the items, which often resulted in students answering incorrectly, even when they demonstrated knowledge of the content. Because the features are most frequent on Life Science items, ELLs' content knowledge on these topics may be severely underestimated for this strand.

Do linguistic features of science test items prevent English Language Learners from demonstrating their knowledge?

Accurately assessing English language learners' (ELLs') knowledge in science is a complex and pressing issue. While many existing assessments have been shown to be inaccurate measures of ELLs' content knowledge (Abedi, et al., 2005; Authors, 2012; Sato et al., 2010), new and increasingly linguistically complex assessments are being developed. To better assess ELLs' science knowledge, many argue that current testing practices need to (a) incorporate theories that acknowledge that ELLs draw from two language systems (Valdés & Figueroa, 1994), (b) consider sociolinguistic aspects of interactions between students and test items (Solano-Flores, 2006; 2008), and (c) include ELLs in the test development process (Abedi & Hefri, 2004).

Testing accommodations are the main tool used to improve the accessibility of science tests for ELLs, and can include the use of bilingual dictionaries, extra time, translated tests, and linguistic simplification (Durán, 2008; Rivera, et al., 2006). While results on the effectiveness of testing accommodations have been mixed (Abedi & Hefri, 2004; Kieffer, et al., 2009; Pennock-Roman, & Rivera, 2011; Sireci, Li, & Scarpati, 2003), some show improved ELL performance with reduced linguistic complexity of test items (Abedi, Courtney, & Leon, 2003; Sato et al., 2010).

However, linguistic complexity of test items is not consistently defined from one study to the next, and specific linguistic features that contribute to linguistic complexity are frequently neither defined nor operationalized (Authors, in preparation). As a result, it is difficult to replicate findings from individual studies and to accumulate knowledge about how specific linguistic features of test items interfere with the performance of ELL students. Furthermore,

much of the research on the effects of linguistic complexity of test items on ELLs' performance has focused on mathematics test items, although the challenges of the language of science test items for ELLs are likely to be even greater (Penfield & Lee, 2010). In addition, studies that have investigated the linguistic complexity of science test items have not explored the relationship between the science strand of test items and the types of linguistic features the items include. We conclude that more research is needed to understand the sources of problematic linguistic complexity for ELL students on science test items, and the relationship between linguistic complexity and science strand. Consequently, this study examines ELL performance on science test items from the Grade 5 Massachusetts Comprehensive Assessment System (MCAS) and its relationship to specific linguistic features of these test items, as well as the distribution of these features across three science strands: Earth and Space Science (ESS), Life Science (LS), and Physical Sciences (PS).

Conceptual Framework

Understanding a test item is crucial for accurately solving it (Leighton & Gokiert, 2008). Discourse features are essential for constructing this understanding, as they affect students' comprehension of the item as a whole. We conceptualize the process of item comprehension as the integration of text and background knowledge to create mental representations that capture the details of what is read (e.g., who, when, why) in the form of a Situation Model (Zwann & Radvansky, 1998). In assessments, the student must also construct a Problem Model, which consists of what the student needs to know from the text and what needs to be done with that information to correctly answer the item (Nathan, Kintsch, & Young, 1992).

We hypothesize that some discourse features may prevent ELLs from creating the intended Situation and Problem Models for test items. In previous work, we found one such feature,

Forced Comparison, to interfere with the construction of the intended Situation and Problem Models for students from diverse backgrounds (Authors, 2008; 2012). The Forced Comparison feature occurred in items asking the student to compare all the answer choices and select the option with an extreme value, such as the *best* or *most likely* choice. Subsequent analysis found that this feature often co-occurred with another discourse feature, Reference Back, which required students to return to a previous sentence to find information necessary to answer the question (e.g., *How would this change affect the plant population?)*. To better understand the role of these discourse features in ELLs' performance on science test items across strands, we ask:

- What is the relationship between science strand and ELL performance?
- What are the frequencies and distributions of these two discourse features across science strands in these test items?
- What is the relationship between these discourse features and ELLs' performance?
- What do student interviews reveal about how ELLs interact with these features?

Methods

This report is part of a 4-year study, currently in progress, investigating the effects of specific linguistic features of test items on ELLs' performance on large-scale standardized multiple-choice items. These features include aspects at the word level, sentence level, and item level (see Kachchaf et al, Submitted). For the purposes of this paper, we focus on two item level features, discussed below.

Quantitative Data Sources

Students. Student performance for the correlation analysis was calculated for Grade 5 students from three classifications of English proficiency: non-ELLs, Limited English Proficient

(LEP), and Formerly Limited English Proficient (FLEP), as determined by the state. For each year, this statewide study included (a) 52,694 – 56,991 non-ELL students, (b) 2,645 - 3,804 LEP students, and (c) 1,761 - 2,466 FLEP students.

Items. We correlated student performance on 162 publically released science multiple-choice items from a state mandated science exam for the years 2004-2010. These items covered three science strands: Earth and Space Science, Physical Science, and Life Science.

Linguistic Features of Test Items. A comprehensive literature synthesis filtered existing studies that investigated the role of linguistic complexity in ELL test performance (see Noble, Kachchaf, & Rosebery, In Preparation). Across the 11 studies identified in the literature synthesis, over 60 linguistic features were identified as potentially influencing ELL performance. From these 60 features, the project team selected 13 features at the word, sentence, and item level to analyze (for details on features not discussed in this study see Kachchaf et al., Submitted). However, the majority of item level features found in previous research were difficult to replicate due to a lack of information provided on how they were operationalized. Therefore, we included an item level feature from our own previous research (the Forced Comparison feature) as well as identified a new item level feature that arose during preliminary analysis of items. Both of these features are discussed below.

Forced Comparison. This feature was defined as an item that typically (a) used *Which of the following,* (b) named a category of what was sought: "Which of the following *drawings…"*, (c) asked students for an end of scale value (e.g., *best shows* or *most likely result*), and (d) had a verb or noun associated with the end of scale value (e.g., *best shows* or *most likely result*). A question statement containing the Forced Comparison is: *Which of the following drawings best shows the*

life cycle of a berry bush? Items with the presence of the Forced Comparison received a score of a 1.

Reference Back. During the pilot study, preliminary analysis of test items with the Forced Comparison feature uncovered another feature that was related to the relationship between sentences in the item. We defined the Reference Back feature as a question sentence that required the student to return to the text of a previous sentence in the item to identify information necessary to answer the question. In some cases, this feature was instantiated in an explicit anaphoric reference. For example, if an item's question statement asked, *Where did this rock most likely form?*, students would need to refer back to a previous sentence to find out what *this rock* was. In other cases, the Reference Back feature occurred when the question sentence had no explicit reference to prior information but nonetheless required students to refer back to previous parts of the question to construct an understanding of what the question asked. This feature score was dichotomously scored.

Qualitative Data Sources

Student interviews. In addition to calculating performance of students statewide, we gathered detailed qualitative data of how students interacted with items containing these two linguistic features. We interviewed 52 ELLs from 3 districts about 32 different test items with and without the Forced Comparison and Reference Back features. For each interview, students answered six multiple-choice items in either (a) the original form containing these features, or (b) a modified form created by the project team that removed these features. After students selected an answer for each item, bilingual interviewers asked the students (1) how they solved the items, (2) whether they understood specific linguistic features of the items, and (3) whether they knew the science content being assessed.

Data Analysis

Quantitative Coding and Analysis. Two coders with experience in teaching and educational research independently were trained to code all 162 items for the presence or absence of features, including the Forced Comparison and Reference Back. For the purposes of coding, the Forced Comparison was identified any item containing an end of scale value (e.g., *best* or *most likely)*. To code the reference back feature, coders were given items with only the question statement and the answer options. If the coders deemed it possible to answer the item with only seeing the question statement, the item was coded as not having the Reference Back feature. If the coders decided it was not possible to answer the item when only reading the question statement, the item was coded as containing the Reference Back feature. Coders were given items in three rounds, randomly determined, to code independently. After coding these items independently, the coders met to discuss any discrepancies and to arrive at consensus.

We calculated Differential Item Functioning (DIF) using the Standardization method (Dorans & Kulick, 1986) to determine which test items showed differences in the probabilities of answering correctly for LEPs, FLEPs, and non-ELLs who were at the same ability levels. DIF values calculated using a second method, HLM-LR, were significantly correlated with DIF values calculated using Standardization method (.873, $p < .001$). Spearman's Rank correlation measured the association between items' DIF values and the presence of the two item level features: Forced Comparison and Reference Back.

Qualitative Coding and Analysis. Two coders independently coded student interviews and discussed discrepancies to arrive at consensus. Drawing from student responses, the coders identified if the student: (1) selected the correct response, (2) understood specific linguistic features of the item, and (3) demonstrated knowledge of the construct the item assessed.

Insert more on data analysis for FC & RB here.

Results

To answer the first research question, *What is the relationship between science strand and ELL performance?*, we investigated the frequency of non-negligible levels of DIF favoring non-ELLs over ELLs across the three science strands included in the STE MCAS. Although the total corpus of 162 items was evenly distributed across each science strand, of the 62 test items with non-negligible DIF, 30 were Life Science (LS) items, while 16 were Earth and Space Science (ESS) items, and 16 were Physical Science (PS) items, indicating that there is a pattern of greater ELL difficulties with LS test items on the STE MCAS. To investigate the reasons for this pattern, we pursued our second research question: *What is the frequency and distribution of these discourse features in Grade 5 science multiple-choice test items*? We calculated the average feature score for each of these features, shown in Table 1 below.

Table 1. *Distribution of Linguistic Features*

Feature	Strand			Total (n=162)
	LS (n=58)	ESS (n=51)	PS (n=53)	
Forced Comparison	0.66	0.55	0.36	0.52
Reference Back	0.34	0.24	0.19	0.26

Table 1 shows the average feature score for all items in the last column, regardless of the science strand that they fell under. The last column shows that half of the items contained the Forced Comparison Feature, one fourth of the items contained the Reference Back feature.

The distribution of these features differed across science strands. Table 1 shows a general trend of LS items having the highest feature score for each discourse feature. PS items had the lowest feature score for each of these features, while Earth and Space had a mean greater than PS items but less than LS items. Some of these differences were quite large. In fact, for LS items,

the mean feature score for the Forced Comparison feature only and the mean feature score for the Forced Comparison and Reference back combined was almost twice as great as scores for the PS items. It appears that item writers tended to utilize certain discourse features when assessing topics in LS as compared to topics under the other science strands.

To answer the second research question, *How do these features relate to ELL performance?*, we correlated these features with the items' DIF values, as shown in Table 2. The Forced Comparison feature was significantly and positively correlated with DIF disfavoring LEP and FLEP students. The Reference Back feature was significantly and positively correlated with DIF disfavoring LEP students only.

Table 2.

Correlations between Linguistic Features of Test Items and Item DIF Values

Features	LEP	FLEP
Forced Comparison	.194*	.192*
Reference Back	.192*	.101

To answer the third research question, *How do ELLs interact with these features?* we analyzed student interviews. Because they were significantly correlated with DIF disfavoring ELLs, we specifically focused only on items with both the Forced Comparison and Reference Back features. Here, we summarize one case. Yolanda, a native Spanish-speaking student classified as LEP, incorrectly answered the Earthworm item (shown below) that contained the Forced Comparison and Reference Back features. She chose C. *by staying where it was placed*.

An earthworm was placed on top of a thick layer of moist topsoil in a pan. The pan was placed in a room with the lights on. *How did the earthworm **most likely** respond to these conditions?*

A. by burrowing under the soil
B. by crawling around in the pan

C. by staying where it was placed
D. by trying to crawl out of the pan

Figure 1. Earthworm item: Forced Comparison in italics, Reference Back underlined.

Although Yolanda knew all of the words except for two terms (*most likely* and *respond)* she stated that she "didn't really get" this item. Focusing on the phrase, *was placed on top of a thick layer of moist topsoil*, Yolanda said C. was correct, "because it says it was on top of a moist topsoil in a pan. If it was on top, it would just be there staying steady." Yolanda thought the phrase *most likely respond to these conditions* asked her to choose the best answer for what the earthworm was doing. It was not clear to her that *respond to these conditions* required her to go back to the first two sentences of the item to determine the earthworm's reaction to (a) being on top of moist topsoil, and (b) being placed in a room with the lights on. Rather than referring back, she thought *these conditions* referred forward to the answer choices. Her interpretation of *most likely,* the extreme value asked for by the Forced Comparison feature, only intensified her difficulty, as she thought it meant *the best answer choice.* A key phrase, *the lights on* was buried in the middle of the item, where Yolanda stated she did not notice it. Nevertheless, she know a lot about earthworms. Later, the item was asked in a simplified form without these features. Yolanda immediately and correctly chose A. *by burrowing under the soil,* stating "earthworms do not like light, so it would go under the soil." It appears that, while she knew most of the words, the discourse features Forced Comparison and Reference Back prevented her from demonstrating her knowledge.

Significance

This study provides insight into discourse features of science test items that can be systematically identified and analyzed. Our results showed that the features were frequent, but that the distribution of these features differed across the three science strands. Each feature was

most frequent on LS items. This is problematic because the Forced Comparison and Reference Back features were significantly correlated with DIF disfavoring LEP and FLEP students indicating that they potentially hinder ELLs' abilities to understand test items containing these features. Interviews confirmed that these two features in combination interfered with ELLs' abilities to make sense of the items, which often resulted in students answering incorrectly, even when they demonstrated knowledge of the content. Therefore, tests may not be obtaining accurate measures of ELL knowledge on items containing these features. Because the features are most frequent on LS items, ELLs' content knowledge on these topics may be severely underestimated by these tests. These results call for further investigation into the ways that ELLs interact with these features as well as the reasons for the exceptional incidence of these features arising in the LS strand. We hypothesize that item writers may have utilized these features to assess complex LS standards in a multiple-choice format when an open-response format may have better assessed the content.

Author Note

The authors would like to thank the school and district administrators, teachers, parents, and students who made the interview study possible, and to whom this work is dedicated. The research reported in this paper was supported by the Institute of Education Sciences, U.S. Department of Education through Grant # R305A110122. The opinions expressed herein are those of the authors and do not reflect the opinions of the funding agency.

References

Abedi, J., Courtney, M., & Leon, S. (2003). *Effectiveness and validity of accommodations for English language learners in large-scale assessments*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Abedi, J., & Hefri, F. (2004). Accommodations for students with limited English proficiency in the national assessment of educational progress. *Applied Measurement in Education, 17*(4), 371-392.

Abedi, J., Bailey, A., Butler, F., Castellon-Wellington, M., Leon, S., & Mirocha, J. (2000/2005). *The Validity of Administering Large-Scale Content Assessments to English Language Learners: An Investigation from Three Perspectives* (CSE Report No. 663). Retrieved from National Center for Research on Evaluation, Standards, and Student Testing website:http://www.cse.ucla.edu/products/reports.asp

Durán, R. (2008). Assessing English-language learners' achievement. *review of Research in Education 32,* 292-327.

Hudicourt-Barnes, J., Noble, T., O'Connor, M. C., Rosebery, A., Suarez, A., Warren, B., & Wright, C. (2008). *Making sense of children's performance on achievement tests: The case of the 5th grade science MCAS.* Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Kachchaf, R., Noble, T., Rosebery, A., O'Conner, M. C., Warren, B., & Wang, Y. (Submitted). A closer look at linguistic complexity: Pinpointing individual linguistic features of science multiple-choice items associated with English language learner performance. Manuscript submitted for publication.

Kieffer, M. J., Lesaux, N. K., Rivera, M., Francis, D. J., (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research, 79*(3), 1168-1201.

Johnstone, C. J., Thompson, S. J., Bottsford-Miller, N. A., & Thurlow, M. L. (2008). Universal design and multimethod approaches to item review. *Educational Measurement: Issues and Practice*, *27*(1), 25-36.

Leighton, J. & Gokiert, R. (2008). Identifying potential test item misalignment using student verbal reports. Educational Assessment, 13, 215-242.

Noble, T., Kachchaf, R., & Rosebery, A. (In Preparation). Assessment and English language leaners: Synthesizing research on linguistic features and construct irrelevant variance. *Manuscript in preparation.*

Noble, T., Suarez, C., Rosebery, C., 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 test and what they know about science. *Journal of Research in Science Teaching, 49*(6), 778-803.

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

Penfield, R. D., & Lee, O. (2010). Test-based accountability: Potential benefits and pitfalls of science assessment with student diversity. *Journal of Research in Science Teaching, 47*(1), 6-24.

Pennock-Roman M. & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice, 30*(3), 10-28.

Polya, G. (1973). *How to solve it: A new aspect of mathematical method.* Princeton University Press.

Pretz, J. E., Naples, A. J., & Sternberg, R. J. (2003). Recognizing, defining, and representing problems. The psychology of problem solving, 3–30.

Rivera, C., Collum, E., Wilner, L. N., & Sia, J. K. (2006). Study 1: An analysis of state assessment policies regarding the accommodation of English language learners. In C. Rivera & E. Collum (Eds.), *State assessment policy and practice for English language learners: A national perspective* (p. 1-136). Mahwah, NJ: Lawrence Erlbaum.

Sato, E., Rabinowitz, S., Gallagher, C., & Huang, C. S. (2010). Accommodations for English language learner students: The effect of linguistic modification fo math test item sets.

Sireci, S. G., Li, S., & Scarpati, S. E. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*(4), 457-490.

Solano-Flores, G. (2006). Language, dialect, register: Sociolinguistics and the estimation of measurement error in the testing of English language learners. *Teachers College Record, 108*(11), 2354-2379.

Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. *Educational Researcher, 37*(4), 189-199.

Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Ablex.

Winter, P. C., & Kopriva, R, Chen, C., & Emick, J. (2006). Exploing indivudal and item facors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning and Individual Differences,* 16, 267-276.

Zwaan, R. A. & Ravansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*(2), 162-185.