The Impact of Discourse Features of Science Test Items on ELL Performance

Rachel Kachchaf[1]

Tracy Noble[2]

Ann Rosebery[2]

Yang Wang[3]

Beth Warren[2]

Mary Catherine O'Connor[4]

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

Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA, April, 2014.

Abstract

Most research on linguistic features of test items negatively impacting English language learners' (ELLs') performance has focused on lexical and syntactic features, rather than discourse features that operate at the level of the whole item. This mixed-methods study identified two discourse features in 162 multiple-choice items on a standardized science assessment. After analyzing the frequency of features, we correlated the presence of these features with values of Differential Item Functioning (DIF) favoring non-ELLs over ELLs. Next, we analyzed 52 interviews with ELLs to examine the interactions between students and items with and without these features. Our results indicate that these two discourse features negatively impacted ELLs' item performance and item comprehension, indicating they should be avoided when developing test items.

The Impact of Discourse Features of Science Test Items on ELL Performance

Objectives

Accurately assessing English language learners' (ELLs') knowledge in domains such as science and mathematics 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; Noble et al., 2012; Sato et al., 2010), new and increasingly linguistically complex assessments are being developed. Many argue that inaccuracies stem from a lack of (a) the incorporation of bilingualism theories that acknowledge ELLs as drawing from two language systems (Valdés & Figueroa, 1994), (b) the consideration for sociolinguistic aspects (Solano-Flores, 2006; 2008; 2009), and (c) the inclusion of ELLs in test development procedures (Abedi & Hefri, 2004).

To improve science and mathematics assessments, research has investigated aspects of language that disadvantage ELLs, but are not related to the content being assessed. Based partly in this research, Universal Design (UD) provides a useful guide to develop items with unnecessary linguistic complexity (Thompson et al., 2002; 2008). Next Generation Assessment Systems—charged with creating assessments to be used by multiple states—have drawn from UD principals to guide item development (PARCC, 2013; SBAC, 2013). This is promising, as it implies that ELLs' needs are being considered during item development. However, while UD provides useful rules to avoid linguistic complexity, more concrete information on operationalizing its principles is needed.

Testing accommodations are another tool used to minimize the role of language in ELL performance, and these include the use of bilingual dictionaries, extra time, translating tests, and linguistic simplification (Butler & Stevens, 1997; 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 a direct link between linguistic complexity and ELL performance (Abedi, Courtney, & Leon, 2003; Sato et al., 2010).

Although research has begun to define linguistic complexity, a recent review of the literature (Noble, Kachchaf, & Rosebery, In preparation) showed great diversity in the types of features investigated across studies. In addition, more attention was given to word-level and sentence-level features than to item- or discourse-level features, even though the latter may profoundly affect students' comprehension of test items as a whole. More research is needed to pinpoint specific discourse features that influence ELL performance. Consequently, this mixed-methods study examines two discourse features as they relate to ELL performance.

Perspectives

Understanding the test item is crucial for accurately solving it (Leighton & Gokiert, 2008; Polya, 1973; Pretz, Naples, & Sternberg, 2003). Discourse features are essential to constructing this understanding, as they affect the item as a whole. Studies have shown that the discourse features that may interfere with ELLs' abilities to make sense of the item include item length (Lord et al., 2000), lack of cohesion or clarity among the parts in the item (Abedi, Courtney, & Leon, 2003; Sato et al., 2010), and unfamiliar item contexts (Martiniello, 2008; Sato et al., 2010; Winter, et al., 2006). However, the latter two of these features can be defined in multiple ways. Additionally, few discourse features have been extensively examined, suggesting that further exploration of discourse features is needed.

To better understand how students make sense of the text, Zwaan and Radvansky (1998) utilize the Situation Model to describe the integration of text and background knowledge to create mental representations that capture the details of the text (e.g., who, when, why). In

assessments, the student must construct a Situation Model for the context described by a test item, and 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 (Coquin-Viennot & Moreau, 2007; Nathan, Kintsch, & Young, 1992).

We hypothesize that certain 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 (Hudicourt-Barnes et al., 2008; Noble et al., 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 occurred in items with question sentences requiring students to return to a previous sentence to find information necessary to answer the item. Often this consists of an explicit anaphoric reference, as in, *How would this change affect the plant population?*.

To better understand the role of these two discourse features in ELLs' performance, we ask:

- How frequently do these two discourse features, Forced Comparison and Reference Back, appear in science multiple-choice items in Grade 5?
- What is the relationship between these discourse features and ELLs' performance?
- How do ELLs interact with these discourse 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, *How frequently do these two discourse features appear on science multiple-choice items*?, we calculated descriptive statistics, shown below.

Table 1

Frequency of Features

Feature	Min	Max	Mean	SD	n[1]
Forced Comparison	0	1.00	0.52	0.50	85
Reference Back	0	1.00	0.26	0.44	42
Forced Comparison and Reference Back	0	2.00	0.78	0.72	99

[1]Indicates the number of items that had at least one feature. Total number of items in this study was 162.

The Forced Comparison feature was quite frequent, occurring in over half the items. The Reference Back feature was less frequent, occurring in one fourth of the items. There were 99 items that had either a Forced Comparison or a Reference Back and 28 items that had both features.

To answer the second research question, *What is the relationship between these discourse features and ELL performance?*, we correlated the presence of features with items' DIF values. Table 2 shows the results.

Table 2

Correlations between Discourse Features and Item DIF Values

Features	LEP	FLEP
Forced Comparison	.194*	.192*
Reference Back	.192*	.101
Forced Comparison and Reference Back	.258**	.200*

* $p < .05$; ** $p < .01$.

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. For both LEP and FLEP students, the combination of these features was also significantly and positively correlated with DIF. Moreover, the correlation for the sum of these features was higher than the correlation of either feature in isolation, indicating an even stronger relationship between these two features and DIF disfavoring LEP and FLEP students.

To answer the third research question, *How do ELL students interact with these discourse features?*, we analyzed student interview data. Table 3 shows that the percentage of students correctly answering items with both features present was lower than students answering the same items with these features removed. Due to limitations of space, the full paper will provide details on our item modification method.

Table 3

ELL Performance on Items with and without Forced Comparison and Reference Back

Item Version	Answered Correct
Forced Comparison and Reference Back Present	48%
Forced Comparison and Reference Back Removed	67%

Students taking items containing the Forced Comparison and Reference Back features correctly answered the items 48% of the time. This increased to 67% when these features were removed. Indeed, analysis of interview transcripts revealed that these features prevented some ELLs from understanding the item. Here, we summarize one case in which Reference Back appeared to be a great issue. Yolanda, a native Spanish-speaking student classified as LEP incorrectly answered the Earthworm item (shown below) that contained both 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.

Yolanda started her interview by stating she "didn't really get" this item. However, she reported knowing most of the words in the item, except two terms from the question sentence: *most likely* and *respond.* 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 that *respond to these conditions* required her to go back to the first two sentences in the item to determine the earthworm's reaction to (a) being on top of moist topsoil,

and (b) the change of being placed in a room with the lights on. Rather than referring back, she thought *these conditions* referred forward to the answer choices. Notably, Yolanda's confusion was increased by thinking *respond* meant *select the correct answer choice*. 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 focus when first reading the item. Nevertheless, she demonstrated knowing 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, although she knew most of the words in the item, the discourse features Forced Comparison and Reference Back prevented her from demonstrating her knowledge.

Significance

This study provides insight into two discourse features that hinder ELLs' abilities to understand multiple-choice science items. In addition to being significantly correlated with DIF, interviews confirmed that these features contributed to ELLs' difficulties comprehending the items. The Forced Comparison and Reference Back are discourse features of test items that negatively impact ELLs' performance but can be systematically identified and modified. These features have not previously been explored by the research on linguistic modification, and thus our findings provide new and valuable information for item writers as it specifies types of linguistic complexity to avoid.

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 University of California, National Center for Research on Evaluation, Standards, and Student Testing website: http://www.cse.ucla.edu/products/reports.asp

Butler, F. A., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations* (Vol. 448). Center for Research on Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Coquin-Viennot, D. & Moreau, S. (2007). Arithmetic problems at school: when there is an apparent contradiction between the situation model and the problem model. *British Journal of Educational Psychology, 77*, 69-80.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 23(4), 355-368.

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.

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.

Partnership for Assessment of Readiness for College and Careers (PARCC). (2013). PARCC Accessibility Accommodations and Fairness. *Partnership for Assessment of Readiness for College and Careers.* Retrieved June 15, 2013. From: http://www.parcconline.org/parcc-accessibility-accommodations-and-fairness.

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.

Smarter Balanced Assessment Consortium (SBAC). (2012). Item writing and review. *Smarter Balanced Assessment Consortium*. Retrieved June 15, 2013. From: http://www.smarterbalanced.org/smarter-balanced-assessments/item-writing-and-review/

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.

Solano-Flores, G. (2009). The testing of English language learners as a stochastic process: Population misspecification, measurement error, and overgeneralization. In K. Ercikan & W. M. Roth (Eds.) *Generalizing from Educational Research*. New York: Routledge.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). Universal design applied to large scale assessments (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved June 7, 2012, from http://education.un.edu/NCEO/OnlinePubs/Syntehsis44.html

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.

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.