Language and the Performance of English-Language Learners in Math Word Problems

MARIA MARTINIELLO
Educational Testing Service

In this article, Maria Martiniello reports the findings of a study of the linguistic complexity of math word problems that were found to exhibit differential item functioning for English-language learners (ELLs) and non-ELLs taking the Massachusetts Comprehensive Assessment System (MCAS) fourth-grade math test. It builds on prior research showing that greater linguistic complexity increases the difficulty of English-language math items for ELLs compared to non-ELLs of equivalent math proficiency. Through textual analyses, Martiniello describes the linguistic features of some of the 2003 MCAS math word problems that posed disproportionate difficulty for ELLs. Martiniello also uses excerpts from children’s think-aloud transcripts to illustrate the reading comprehension challenges these features pose to Spanish-speaking ELLs. Through both DIF statistics and the voices of children, the article scrutinizes the appropriateness of inferences about ELLs’ math knowledge based on linguistically complex test items.

One of the most pressing issues in the current focus on high-stakes educational assessment is ensuring valid measurements of content-area skills for students who are not proficient in English. The No Child Left Behind Act mandates the reporting of English-language learners’ (ELLs’) math and English language arts (ELA) scores in the states’ accountability systems. Schools failing to show adequate yearly progress (AYP) for their ELL groups may face corrective actions (e.g., replacement of staff or curriculum) or restructuring (e.g., replacement of principals and state takeover) (U.S. Department of Education, 2002). This system presents a number of challenges for schools, since ELLs are among the lowest-scoring groups in both national and state assessments. The pervasive gap in scores between ELLs and non-ELLs in large-scale mathematics and reading assessments has been well documented (Abedi, 2006; Abedi...

The use of testing in education presupposes that a student’s test score is an accurate reflection of her mastery of a particular content area. However, if the student is an ELL and the math test includes questions the student might have trouble understanding, it is unknown whether the low score is due to the student’s lack of mastery of the math content, limited English proficiency, or both. Numerous reports commissioned by the National Research Council discuss the questionable validity of inferences from content-based assessments for students who are not proficient in English (August & Hakuta, 1997; National Research Council, 2000, 2002). As stated in their 2000 report, “A test [of proficiency in a content area] cannot provide valid information about a student’s knowledge or skills if a language barrier prevents the students from demonstrating what they know and can do” (National Research Council, p. 20). Research suggests that items with unnecessary linguistic complexity are a potential source of bias when assessing ELLs’ subject mastery (Abedi, 2004; Abedi, Leon, & Mirocha, 2003; Martiniello, 2006, 2007). Experts argue for the need to distinguish the language skills of ELLs from their subject-area knowledge (Abedi, 2004; Abedi & Lord, 2001), but they recognize the difficulty in doing so, because all assessments administered in English are also measures of English proficiency (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985; August & Hakuta, 1997; National Research Council, 2000).

The study on which this article is based explored the nature of linguistic complexity in math word problems that present comprehension difficulties for ELLs. I implemented differential item functioning (DIF) procedures to identify those items posing greater difficulty for ELLs than for non-ELLs of comparable math proficiency on the fourth-grade Massachusetts Comprehensive Assessment System (MCAS) mathematics test administered statewide in 2003 (68,839 students). I solicited expert reviews and textual analyses of test items to examine the complexity of their nonmathematical language. To investigate reading comprehension challenges ELLs encounter during the test, I administered test items to a sample of fourth-grade Spanish-speaking ELLs using think-aloud protocols. In these think-alouds, ELLs reported and verbalized their understanding of and responses to the items as they read them. (For a discussion of think-aloud protocols, see Ericsson & Simon, 1994.)

In this article I present an in-depth review of six items found to favor non-ELLs over ELLs in the DIF analysis. Textual analysis and think-aloud data confirm that such items exhibit complex syntactic and lexical features that challenge ELLs’ text comprehension. Differences in content among test items also
account for some of the disproportionate difficulty for ELLs. I also discuss the literature on text comprehension, paying particular attention to specific syntactic and lexical features of math word problems and their impact on the differential performance of ELLs and non-ELLs. The notion of differential item functioning is defined, and empirical studies on sources of DIF for ELLs in math word problems are reviewed. Additionally, DIF indices and empirical item characteristic curves are presented to illustrate differences in conditional difficulty for ELLs and non-ELLs for six DIF items favoring non-ELLs over ELLs. Selected transcripts with children’s responses are shown along with the analyses of the text’s syntactic and lexical features. Finally, I discuss implications for item writing and score interpretation.

Research Background and Context

Reading comprehension has been characterized as being a meaning-making process that involves reciprocal exchanges between readers and text for a particular purpose or task (National Reading Panel, 2000) and as being part of a broader sociocultural context (RAND Reading Study Group, 2002). Successful text comprehension requires that readers recognize and decode words quickly and effortlessly (reading fluency); interpret the appropriate meaning of each word given the context (vocabulary knowledge); understand the syntactic arrangement of words and phrases in the sentence (syntactic knowledge or sensitivity); and, finally, extract and construct meaning of the string of words or sentences based on the semantic, contextual, and structural relationships among them (discourse comprehension) (Adams, 1990; National Reading Panel, 2000; RAND Reading Study Group, 2002; Snow, Burns, & Griffin, 1998). Comprehension may also be influenced by sociocultural factors such as familiarity with cultural subject matter, group differences in access to texts, and instruction (RAND Reading Study Group, 2002).

Research has shown that for text comprehension to take place, approximately 90 to 95 percent of the words in any given passage or text must be known to the reader (Carver, 1994; Nagy & Scott, 2000). Difficulty understanding words is related to their frequency of use. High-frequency words are likely to be known by students, since repeated exposure to words in multiple contexts is a good predictor of word acquisition (McKeown, Beck, Omanson, & Pople, 1985). Furthermore, high-frequency words are more likely to be recognized and decoded fluently, while low-frequency words are less likely to be recognized, thus slowing down the reading process, increasing the memory load, and interfering with text comprehension. As Adams (1990) explains, “The greater the time and effort that a reader must invest in each individual word, the slimmer the likelihood that preceding words of the phrase will be remembered when it is time to put them all together” (p. 141). The degree of polysemy — the number of “possible meanings [to] be considered in order to settle on a precise meaning for the local context” (Bravo, Hiebert, & Pearson,
Most measures of text comprehensibility, such as readability indices, rely on sentence length (average number of words per sentence) as an overall indicator of linguistic complexity (Abedi, Lord, & Plummer, 1997; Adams, 1990; Chall & Dale, 1995; Mestre, 1988). Sentence length functions as “an estimate of the number of meaningful ideas that the reader must interrelate in interpreting the sentences” (Kintsch & Keenan, 1973, as cited in Adams, 1990, p. 154). The greater the sentence length, the greater the number of individual words to decode and process. In addition to length, the syntactic characteristics of sentences and clauses also affect comprehension. The transparency or clarity of the syntactic relationships among words, clauses, and phrases is critical for text comprehension (Adams, 1990). For instance, constructions such as subordinate clauses are more difficult to understand than coordinate clauses. Also, verbs in passive voice are more difficult than those in active voice.

Research on ELLs’ text comprehension of math word problems is consistent with the empirical findings about text comprehension by proficient English speakers. Linguistic features of natural language that create comprehension difficulties for ELLs relate to complex vocabulary (lexical complexity) and sentence structure (syntactic complexity) (Abedi & Lord, 2001; Abedi, Lord, & Hofstetter, 1998; Abedi et al., 1997; Butler, Bailey, Stevens, Huang, & Lord, 2004; Spanos, Rhodes, Dale, & Crandall, 1988). Lexical features of complexity that have been studied in previous research include number of low-frequency words, abstractions, polysemy of words, and idiomatic and culture-specific nonmathematical vocabulary terms. Syntactic features include mean sentence length in words, item length in words, noun phrase length, number of prepositional phrases and participial modifiers, syntactically complex sentences, use of passive voice in the verb phrase, and complex sentences, which are sentences with relative, subordinate, complement, adverbial, or conditional clauses.

The relationship between some of these linguistic features and the differential difficulty of math word problems for ELLs and non-ELLs has been investigated in elementary, middle, and high school students (Abedi, Bailey, Butler, Castellon-Wellington, Leon, & Mirocha, 2005; Abedi et al., 1997; Abedi et al., 1998; Abedi & Lord, 2001; Lord, Abedi, & Poosuthasee, 2000; Bailey, 2005; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006). Although many of these studies did, as predicted, find a relationship between linguistic complexity and ELLs’ performance in math word problems, the effect of specific linguistic features varies from test to test and from one grade to another. Of the features studied, only item length showed relatively consistent negative effects on item difficulty for ELLs and non-ELLs in a variety of math tests and grade levels in national and state samples.

Longer items exhibited greater differences between ELLs and non-ELLs than did shorter items in the NAEP math test (grade eight, in Abedi et al., 1997) and the Stanford Achievement Test 9th Edition (SAT9) (grade eight,
Language and the Performance of ELLs in Math Word Problems

MARIO MARTINIOLLO

Delaware sample, in Lord et al., 2000). Also, longer items presented more difficulty than shorter items for both ELLs and non-ELLs in the Kansas general math assessments (KGMA) (grades four, seven, and ten, in Shaftel et al., 2006).

Results for other linguistic features were relatively inconsistent. For example, items with more prepositions, pronouns, and words with multiple meanings were more difficult than items with fewer or none of these features for both ELLs and non-ELLs in the KGMA grade four, but not in grades seven and ten (Shaftel et al., 2006). Also, ambiguous vocabulary showed significant effects on item difficulty in the KGMA grade four, while the presence of low-frequency vocabulary showed no significant effect in the SAT9 grades three and eight. Neither the number of passive-voice sentences nor the number of subordinate clauses has shown a significant effect in any of the tests and grades studied (NAEP grade eight, in Abedi et al., 1997; SAT9 grades three and eight, in Lord et al., 2000; KGMA grades four, seven, and ten in Shaftel et al., 2006).

Differential item functioning refers to the discrepancies in difficulty an item presents for members of two groups who have equivalent levels of proficiency on the construct the test is intended to measure (Dorans & Holland, 1993; Thissen, Steinberg, & Wainer, 1993; Wainer, 1993). Research on the sources of DIF for ELLs in math word problems finds that the linguistic complexity of items and their curriculum learning strand predicted items’ differential difficulty for ELLs and non-ELLs with comparable math proficiency (Martiniello, 2006). Employing item response theory DIF detection methods, Martiniello (2006, 2007) estimated differences in item difficulty for ELLs and non-ELLs in the 2003 MCAS math test and showed that DIF could be predicted from items’ syntactic and lexical complexity. In contrast to linguistically simple items, more complex items in the test tended to show greater DIF favoring non-ELLs over ELLs. In other words, ELLs tended to have a lower probability of correctly answering linguistically complex items than non-ELLs with equivalent math ability. At the highest end of the linguistic complexity range, items contained complicated grammatical structures that were essential for comprehending the item, along with mostly low-frequency, nonmathematical vocabulary terms whose meanings were central for comprehending the item and could not be derived from the context. These were items for which students with limited English proficiency had greater trouble deriving the accurate interpretation. Likewise, items measuring the learning strand data analysis, statistics, and probabilities tended to present greater difficulty for ELLs than for non-ELLs with equivalent math scores, compared to items measuring the rest of the learning strands in the test (Martiniello, 2006).

These prior studies investigated composite measures of linguistic complexity and their association with relative difficulty for ELLs and non-ELLs of comparable math proficiency (Martiniello, 2006, 2007). They did not examine specific syntactic or lexical features of the items. My study builds on these
earlier findings using a detailed examination of the linguistic features of the items showing DIF disfavoring ELLs and triangulates this information with think-aloud responses to the 2003 MCAS math items from Spanish-speaking ELLs.

Methods
I analyzed the linguistic complexity of MCAS math items and used think-aloud protocols to gather evidence of comprehension difficulties for Spanish-speaking ELLs. In order to identify those items that posed greater difficulty to ELLs than to non-ELLs with comparable math proficiency, I employed two DIF detection methods. I examined items flagged for evidence of DIF disfavoring ELLs and described them in terms of the degree of DIF, the learning strand they represent, their linguistic complexity, and children’s responses to them in the think-aloud interviews.

Instrument
The test studied is the English version of the fourth-grade MCAS mathematics exam administered statewide in the spring of 2003. This is a standards-based achievement test aligned with the Massachusetts Mathematics Curriculum Framework. It includes five major learning strands: number sense and operations; patterns, relations, and algebra; geometry; measurement; and data analysis, statistics, and probabilities (Massachusetts Department of Education [MDOE], 2003a). The items analyzed included a total of thirty-nine publicly released items, twenty-nine items in a multiple choice format and ten of constructed response (five of short-answer format and five open-ended).

Measures
Measures of linguistic complexity for each item were derived from a detailed microanalysis of the text’s syntactic complexity, as well as from expert ratings of the items’ overall syntactic and lexical complexity.

Textual analysis. A coding system was developed to identify elements of complexity in the structural relationships among words, phrases, and sentences in the items, such as the number of clauses, noun phrases, verbs, and verb phrases. The coding system captured the length of these grammatical elements by marking the beginning and ending of each clause or phrase. Additional codes further specified the syntactic function or type of all elements and the syntactic order of clauses.

Two linguists were trained to use this microanalytic coding manual with items from a different version of the MCAS fourth-grade math test. The first linguist coded all thirty-nine items, while the second linguist coded 20 percent of the items independently and reviewed the first rater’s original coding for the remaining 80 percent. The agreement between these two raters in the microanalytic coding, adjusted for chance agreement, was high (Cohen’s
kappa coefficient = 0.89). They discussed discrepancies and, when necessary, recoded items.

**Vocabulary frequency and familiarity.** To assess whether the item vocabulary was likely to be known by the majority of fourth-grade students, we cross-checked words against *A List of 3,000 Words Known by Students in Grade 4* (Chall & Dale, 1995) and *Living Word Vocabulary* (LWV) (Dale & O’Rourke, 1979). LWV is a national vocabulary inventory that provides familiarity scores on 44,000 written words (tested in grades four, six, eight, ten, twelve, thirteen, and sixteen) (Dale & O’Rourke, 1979). *A List of 3,000 Words Known by Students in Grade 4*, a subset of the LWV comprising 3,000 words commonly known by 80 percent of fourth graders, is used in the calculation of readability formulas (Chall & Dale, 1995).

— Think-Aloud Interviews

Think-aloud protocols using items from the 2003 English version of the MCAS math test were administered to a nonprobability sample of twenty-four fourth-grade ELLs attending six inner-city Massachusetts public schools in the spring of 2005. Two of these schools offered dual immersion programs (Spanish and English). The students interviewed were first- or second-generation Latin American immigrants from Colombia, El Salvador, Mexico, Guatemala, Peru, and the Dominican Republic, as well as some students from Puerto Rico. They had between two and four years of schooling in the U.S., came from homes where Spanish was the primary language, and were identified by their teachers as ELLs. The sample was gender balanced. Parents signed a written consent (in Spanish) allowing their children to participate in this study. Based on their teachers’ ratings, the children interviewed represented a wide range of mathematics and English-language proficiencies. Children who could not read or communicate in English were not interviewed since they would not be able to read the test at all.

I conducted think-aloud interviews individually in either one or two sessions that lasted between thirty and ninety minutes each. I audiotaped and later transcribed these sessions. Interviews were conducted primarily in Spanish, although I did encourage the children to respond in the language they felt most comfortable speaking. The children who were more fluent in English often used both languages during the interviews.

I presented items from the 2003 MCAS math test to the students, who read aloud the item text in English. I paid attention to decoding errors, such as words students stumbled over, could not pronounce correctly, skipped, or paused to consider. Children explained what the item was asking them to do. Probe questions assessed whether children could a) understand the text in English; b) rephrase the text in Spanish or English to demonstrate their understanding of its meaning; c) identify which aspects (if any) of the English text they could not understand; and d) figure out what the item was requiring them to do, even if they could not understand the text in its entirety.
recorded the types of mistakes children made when interpreting linguistically complex text as well as their ability to answer the question correctly based on their understanding of the text.

**DIF Detection: Selection of Items for In-Depth Analysis**

I identified items showing differential difficulty for ELLs and non-ELLs with equivalent math proficiency (i.e., equal total math scores) with the aid of two commonly used DIF detection methods: standardization (Dorans & Kulick, 1986) and Mantel-Haenszel (Holland & Thayer, 1988), using a refined total test score matching criterion. DIF indices were categorized according to classification guidelines employed by the Educational Testing Service (ETS; Dorans & Holland, 1993). I flagged ten items as showing some evidence of DIF disfavoring ELLs: nine items coded as category B (slight to moderate DIF) and one as category C (moderate to large DIF). Two of these ten items exhibited large DIF (STND P-DIF > 0.1) using criteria suggested by Zenisky, Hambleton, and Robin (2003).

— **DIF Detection Sample**

The sample used for DIF detection was a subsample of all the fourth-grade students in Massachusetts who took the test in the spring of 2003. In total, the DIF sample included 68,839 students, 3,179 of whom were ELLs. This sample excluded students classified as former ELLs and ELL students who took the test in Spanish.

— **Sample Descriptive Statistics**

In the 2003 MCAS math test, the average raw score for ELLs (25.1, SD = 10.6) was almost one standard deviation below the average raw score for non-ELLs (34.3, SD = 9.9) (see Table 1). In the 2003 MCAS English language arts (ELA) test, the average raw score of ELLs was 38.9 (SD = 12.4) compared with 52.3 (SD = 10) among non-ELLs. The distribution of scores for ELLs was slightly more dispersed than that for non-ELLs in math and was relatively greater in ELA, revealing greater heterogeneity in ELLs’ English-language skills. The average mean difference between the groups was slightly larger for the ELA test than for the math test (1.28 and 0.92 SDs, respectively), consistent with findings by Abedi et al. (2003) that the performance gap between groups is greater in areas that have a greater language load. The math and ELA test scores are moderately correlated ($r = 0.72$, $p < 0.001$), a finding that is consistent with analyses of other tests.

In this study, language proficiency status was highly associated with socioeconomic status. Eighty-two percent of ELLs received free lunch, compared to 27 percent of non-ELLs. As suggested by their mean scores, there was a wide gap between the groups’ proficiency levels in math. Only 14.4 percent of ELLs scored proficient or above on the test, while 42.7 percent of non-ELLs did. ELLs were overrepresented in the lowest performance level. They made up 13.6
Results

I examined six of the ten MCAS items identified as showing DIF favoring non-ELLS over ELLs in order to gauge the possible contribution of nonmathematical linguistic complexity to the unusual difficulty experienced by ELLs. Following are the items showing the most severe DIF disfavoring ELLs. Each item is described in terms of DIF and its associated learning strand in the Massachusetts Mathematics Curriculum Framework, followed by a textual analysis of its syntactic and lexical features. In addition, transcripts from think-aloud interviews with fourth-grade Spanish-speaking ELLs responding to the 2003 MCAS math items are presented to illustrate comprehension errors that these ELLs make when interpreting complex text in these DIF items.

High DIF Items

Items 2 and 8 have the largest DIF indices favoring non-ELLS over ELLs with equivalent math proficiency. For comparative purposes they are discussed along with items 21 and 30, respectively. The first pair of items (items 2 and 21) illustrates marked differences in the degree of text linguistic complexity, size of DIF disfavoring ELLs, and curriculum content measured. The second pair (items 8 and 30) illustrates differences in linguistic complexity features and size of DIF with identical curriculum content.

— High DIF Item 2

Differential item functioning. One way to represent DIF is to plot for each group the conditional proportion answering the item correctly ($p$ value) as a function of total test score. For an item with negligible DIF, the two curves will be similar. For an item showing DIF, the curves will be different. As shown in Figure 2, at nearly all score levels a considerably larger proportion of non-ELLS than ELLs answered item 2 correctly. The odds of answering the item correctly were nearly twice as high for non-ELLS than for ELLs with the same test scores.

### TABLE 1 Average Raw Math and ELA Scores for ELLs and Non-ELLS

<table>
<thead>
<tr>
<th>Test</th>
<th>ELLs</th>
<th>Non-ELLS</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
<td>$N$</td>
</tr>
<tr>
<td>ELA</td>
<td>38.88</td>
<td>12.41</td>
<td>3173</td>
</tr>
<tr>
<td>Math</td>
<td>25.16</td>
<td>10.62</td>
<td>3179</td>
</tr>
</tbody>
</table>

percent of the warning category, though they comprised only 4.6 percent of the total sample of students in the state.\(^4\)
FIGURE 1  Item 2

To win a game, Tamika must spin an even number on a spinner identical to the one shown below.

Are Tamika’s chances of spinning an even number certain, likely, unlikely, or impossible?
A. certain
B. likely
C. unlikely
D. impossible


FIGURE 2  Empirical Item Characteristic Curve — Item 2. Category C DIF

M-H Odds ratio = 1.92; STND P-DIF = 0.15
Learning strand. Item 2 measures the learning strand data analysis, statistics, and probabilities. Responding to this item correctly indicates that students “understand and apply basic concepts of probability,” specifically that they know how to “classify outcomes as certain, likely, unlikely, or impossible by designing and conducting experiments using concrete objects” — in this case a spinner (MDOE, 2000, p. 44). Assuming familiarity with spinners and the mathematical even number concept, the item requires students to discern that five of the eight numbers on the spinner are even, and therefore it is likely that Tamika will spin an even number.

Linguistic complexity. The item stem consists of two long multiclausal sentences (see Figure 1). The first sentence starts with an introductory adverbial clause that includes a nonpersonal form of a verb (to win). The main clause starts with an uncommon proper noun (Tamika) that functions as subject. The verbal phrase displays a complex verb that includes the modal verb must, a noun phrase as a direct object (an even number), and a highly complex prepositional phrase with an embedded adjectival phrase that includes the past participle shown. The second part of the item stem is a question. This question contains a complex noun phrase with a possessive construction and a prepositional phrase that includes a nonpersonal form of a verb (gerund spinning). The following words are not part of the 3,000-word list known by 80 percent of fourth graders: even, spinner, identical, likely, and unlikely (Chall & Dale, 1995).

Think-aloud interviews. Think-aloud interviews conducted in the spring of 2005 revealed that this item was difficult for fourth-grade Spanish-speaking ELLs to understand. Presented here are transcriptions from two children whose poor text comprehension led to incorrect solutions to the problem. They explain their understanding of the English text in the item:

Child 1:  

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tamika hizo un juego de éstos.</td>
<td>Tamika did one of these games.</td>
</tr>
<tr>
<td>Y tiene que caer en esto ...</td>
<td>And it has to fall in this ...</td>
</tr>
<tr>
<td>[Child points to the spinner.]</td>
<td></td>
</tr>
<tr>
<td>¿Cuáles posibilidades hay de que caiga en uno?</td>
<td>What is the likelihood that it would fall in one [the number-one slot]?</td>
</tr>
<tr>
<td>Maybe puede caer, maybe no puede caer.</td>
<td>Maybe it can fall, maybe not.</td>
</tr>
<tr>
<td>“Likely,” es posible.</td>
<td>“Likely” means is possible.</td>
</tr>
<tr>
<td>“Impossible” es que no va a caer.</td>
<td>“Impossible” is that it will not fall.</td>
</tr>
<tr>
<td>“Certain” es que va a caer.</td>
<td>“Certain” is that it will fall.</td>
</tr>
<tr>
<td>“Likely” es posible, tal vez va a caer.</td>
<td>“Likely” it is possible, maybe it will fall.</td>
</tr>
<tr>
<td>“Unlikely” tal vez no va a caer.</td>
<td>“Unlikely,” maybe it will not fall.</td>
</tr>
<tr>
<td>[Child pauses to reason his response to the item.]</td>
<td></td>
</tr>
<tr>
<td>Es “unlikely,” tal vez no va a caer.</td>
<td>It is “unlikely,” maybe it will not fall.</td>
</tr>
</tbody>
</table>
Of all the content words in the item, this child understood the words game, number, and one, and identified Tamika as a proper noun. He did not know what spin, spinner, or spinning meant, but he was able to recognize the picture of the spinner. He figured out that Tamika played with the spinner and that he must evaluate her chances of getting a particular number.

In the phrase identical to the one shown below, the word one is used as a pronoun to refer to the spinner displayed on the page. The child interpreted one as referring to the numeral 1 on the spinner. Recognizing the word one among many unknown words, he deduced that the spinner’s arrow must fall in the number 1 slot. He failed to recognize the syntactical function of the word one, used as a pronoun in this sentence, and instead misinterpreted it as the numeral one. Based on this linguistic misinterpretation, he offered a reasonable answer: “It is ‘unlikely,’ maybe it will not fall.”

It is evident from his interpretation of the item distractors that this child was familiar with the English words certain, likely, unlikely, and impossible, which were used to classify the mathematical likelihood of the event spinning an even number. However, he was not able to identify correctly what event the item referred to, possibly due to the large proportion of unknown words in the first sentence. To illustrate the child’s lexical understanding of this sentence, I substitute unknown words with blanks:

To —— a game, Tamika must —— an —— number on a —— —— to the one —— ——.

Despite having a greater vocabulary, Child 2 made a similar mistake to Child 1.

Child 2: Para ganar el juego, Talia ... Ta ... 
   ka ... 
   no puedo decir ese nombre ... 
   Tamika necesita hacer así alrededor.
   To win the game, Talia ... Ta ... 
   ka ... 
   I cannot say that name ... 
   Tamika needs to do this around.

[Child makes a spinning gesture.]
Language and the Performance of ELLs in Math Word Problems

MARIA MARTINIELLO

para tener el número del “spinner” “identical” con el número uno “shown below,” igual al 1.

¿Cuáles son las chances, como la posibilidad que ella tiene ese número? “certain, likely unlikely, impossible”

Ella necesita ganar con el número 1.


Es “unlikely” que ella va a tener ese número …

[Child points to the numeral 1 in the spinner.]

porque sólo hay uno de 1. because there is only one 1.

Child 2 was the only student interviewed who knew that the word identical meant equal. However, she was unable to understand the syntactic boundaries of the phrases and the syntactic function of the word one. She interpreted the phrase identical to the one shown below as modifying the noun number rather than the noun spinner. She resolved the syntactic ambiguity of identical to the one [spinner] shown below as identical to the [numeral] one [1] shown below. On completing the item, the student was asked what even number meant, and she responded:

un número con pareja, 2, 4, 6 … a number with couple/pair, 2, 4, 6 …

The child was instructed to reread the item one clause at a time, so as to scaffold the parsing of the syntactic boundaries:

To win a game,
Tamika must spin an even number on a spinner identical to the one shown below.

After reading, she provided a new answer.

Oh! Es “likely,” porque hay 2, 3, 4, 5; cinco “even numbers.”

[Child counts numbers on the spinner.]

In this particular item, the layout of the text does not facilitate the delimitation of the syntactic boundaries for children who may not be fluent readers. For instance, the sentence To win a game, Tamika must spin an even number on a
spinner identical to the one shown below is displayed in three lines. The first line says To win a game, Tamika must spin an even number on a spinner identical to the one shown below, and the third line says shown below. In the example above, Child 2 did not perceive the word even as modifying the noun number, even though she knows the meaning of the phrase even number. Likewise, the word one, which ends the second line, is perceived as separated from shown below.

Syntactically, the clause one shown below could be rewritten as one that is shown below or the spinner that is shown below. It is possible that the layout favored a perception of the word one as a numeral standing on its own and separated from the next line, instead of a pronoun associated with the past participial phrase shown below, which is in the next line.

In both examples above, the children understood the concept of probability of an outcome using spinners, knew the mathematical meaning of the English words likely and unlikely, and could correctly classify the likelihood of a particular event occurring. Nonetheless, they could not answer the item correctly because they were unable to understand the text providing them with the information about the event to be classified.

Another important challenge for the ELLs interviewed was their lack of familiarity with the item’s vocabulary, particularly some of the distractors. One hundred percent of the children knew the meaning of impossible, a Spanish-English cognate that is a high-frequency word in Spanish. In contrast, few children understood the words identical (4%) and certain (33%). These are also Spanish-English cognates, but, unlike impossible, they are infrequently encountered in conversation because more-colloquial synonyms, such as igual (equal) and seguro (sure), are available. About half of the children either ignored or confused the meanings of the words likely and unlikely, as shown in the transcripts that follow. Child 3 misinterpreted the word identical:

Child 3: Para ganar Tamika debe … To win, Tamika must …

[Child makes a spinning gesture.]

para tener un “even” número de … have an “even” number of …
twins que enseña abajo. twins that show below.

Si Tamika trata de … If Tamika tries to …

[Child makes a spinning gesture.]

un número “even,” si es imposible. an even number, if that is impossible.

Probing questions revealed that this student interpreted identical as meaning twin (in Spanish, gemelos) because of the expression identical twins.

Interviewer: ¿Qué crees que te están preguntando aquí? What do you think they are asking you here?
Language and the Performance of ELLs in Math Word Problems
MARIA MARTINIHELLO

[Child points to the distractors certain, likely, unlikely, impossible.]

Child 3: Si hay de verdad la chance, If there is truly a chance
un chin, [referring to “certain”],
mucho, a bit [“likely”],
o imposible, no hay chance. a lot [“unlikely”],
or impossible, there is no chance.

While Child 3 exchanged the meanings of the English words likely and unlikely, Child 4 switched the meanings of certain and likely:

Child 4: Para jugar, Tamia … tiene que … To play, Tamia … has to …
el “spinner,” the “spinner,”
tiene que coger she has to get
un número que sea “even” an “even” number
que es multiplicado por 2, that is multiplied by 2,
como 2 por 2 es 4, like 2 by 2 is 4,
4 es un “even number,” 4 is an “even number,”
entonces, ¿tiene que ser el chance, then, will the chance be
“certain”? certain?
Que es como puede pasar pero no pasa Which is like it can happen but it
seguro, won’t happen for sure,
“likely,” que puede pasar seguramente, “likely,” that can happen for sure,
“unlikely,” puede pasar, pero “unlikely,” it can happen, but
no estoy tan segura, I am not that sure,
imposible, que nunca va a pasar. impossible, that it will never happen.

— Comparison Item 21

Item 21, with the lowest ranking on linguistic complexity in the test, provides some comparison in both the degree of linguistic complexity and magnitude of DIF.

Differential item functioning. Unlike item 2, item 21 shows no differential item difficulty when comparing ELLs and non-ELLs with equivalent math scores. In Figure 4, the lines describing the conditional proportion correct for ELLs and non-ELLs are similar: ELLs and non-ELLs with the same test scores have similar odds of answering the item correctly.

Learning strand. Item 21 assesses the learning strand number sense and operations, specifically students’ understanding of the base-ten number system by reading decimals (MDOE, 2000).

Linguistic complexity. Items 2 and 21 illustrate the extremes in the test’s linguistic complexity range. The length in words of item 21 is one-fourth the length of item 2. Item 21 has only eight words in a single sentence compared with thirty-two words (sixteen per sentence) in item 2. (Item length in words is
a well-accepted index of linguistic complexity and has been shown to generate comprehension difficulty.) Item 21 has only one clause per sentence, which, despite its passive-voice verb, has a relatively simple structure that follows a prototypical question format. Finally, all the words in item 21 are included in the list of 3,000 words known by 80 percent of fourth graders in the U.S. (Chall & Dale, 1995).  

Think-aloud interviews. All ELLs interviewed understood that item 21 required a translation between the written version and the numerical version of the decimal number. Even though some children did not know the word following, they derived its meaning from context, since the rest of the words in the sentence are familiar and the syntactic structure is relatively simple.
Child 5: Este puntito.

[Child points to decimal point in the number 5.10.]

Mi maestro nos explicó que representaba “and,”
5 and,
y luego necesito hacer la fracción.
It’s not 5.10 because that’s five and a tenth.

This little dot/period.
My teacher explained to us that it represented “and,”
5 and,
and then I need to make the fraction.
It’s not 5.10 because that’s five and a tenth.

— High DIF Item 8 and Comparison Item 30

Item 8 has the second-largest DIF index disfavoring ELLs in the test. For comparison purposes, item 8 is discussed along with item 30, which measures the same curriculum content.

Differential item functioning. Compared with item 30, item 8 is much more difficult for ELLs than for non-ELLs with equivalent math scores. The odds of responding to item 8 correctly for non-ELLs are close to double the odds for ELLs, while for item 30 the odds for non-ELLs are about 1.3 times those for ELLs. Figures 6 and 7 illustrate the discrepancy in the difficulty gaps across groups in these two items. Since both items measure students’ ability to count combinations, the discrepancy between their DIF indices cannot be attributed to differences in proficiency in this skill.

Learning strand. Like item 2, items 8 and 30 assess the learning strand data analysis, statistics, and probabilities. They require children to apply basic concepts of probability by counting the number of possible combinations of objects from two or three sets (MDOE, 2000).

Linguistic complexity. A close examination of the items’ linguistic complexity shows that both items’ syntactic structures are relatively complex. These two items display particularly lengthy sentences and so might pose a challenge for ELLs. Item 8 has three sentences with an average of twelve words per sentence. Item 30 also has three sentences with an average of sixteen words per sentence. Experts’ reviews of the items’ linguistic features ranked these items similarly in their syntactic complexity but rated item 8 as lexically more complex than item 30.

All the words in item 30 are on the list of 3,000 words known by 80 percent of fourth graders (Chall & Dale, 1995). Two words in item 8 do not appear in this list: chores and vacuum. According to the LWV list, chores is known by 67 percent of fourth graders tested nationwide (Dale & O’Rourke, 1979).

Think-aloud interviews. In item 30, all the ELLs interviewed understood the meaning of the words students, notepad, pencil, ruler, day, school, and colors, while only a few had trouble with the past-tense verb gave and the preposition below. In contrast, most of the students were unacquainted with the word chores, the
FIGURE 5  Items 8 and 30

8 Every Saturday in the fall, Martin has to do 1 inside chore and 1 outside chore. The chores are listed below.

<table>
<thead>
<tr>
<th>Inside Chores</th>
<th>Outside Chores</th>
</tr>
</thead>
<tbody>
<tr>
<td>vacuum</td>
<td>rake</td>
</tr>
<tr>
<td>wash dishes</td>
<td>weed</td>
</tr>
<tr>
<td>dust</td>
<td></td>
</tr>
</tbody>
</table>

How many different combinations of 1 inside chore and 1 outside chore can Martin make?
A. 3
B. 5
C. 6
D. 9

30 Mr. Garcia gave each of his students a notepad, a pencil, and a ruler on the first day of school. The chart below shows the different colors of the notepads, pencils, and rulers.

<table>
<thead>
<tr>
<th>Colors of Notepads, Pencils, and Rulers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notepads</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>Yellow</td>
</tr>
<tr>
<td>White</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

How many different combinations of 1 notepad, 1 pencil, and 1 ruler can Mr. Garcia make?


phrases inside chore and outside chore, and the different words listed in the table in item 8. Only two girls knew what vacuum, wash dishes, and dust meant. None of the children knew the meaning of the words rake and weed.

About 88 percent of the ELLs interviewed were familiar with the mathematical meaning of the word combination, a Spanish-English cognate, and understood that they were supposed to create and count combinations of things. However, in item 8 they did not know what to combine, while they did know what to combine in item 30. For these ELLs, the proportion of unknown words was larger in item 8 than in item 30, making their comprehension of the former much more difficult than the latter.

The ELLs interviewed showed a clear understanding of school-related words, whereas words related to home posed bigger challenges. It seems the experts were able to identify this additional hurdle when rating the lexical complexity.
FIGURE 6  Empirical Item Characteristic Curve — Item 8. Category B DIF

M-H Odds ratio = 1.8; STND P-DIF = 0.11

FIGURE 7  Empirical Item Characteristic Curve — Item 30. Category B DIF

M-H Odds ratio = 1.3; STND P-DIF = 0.05
of item 8. In this case, word-frequency lists do not offer any insights as to the lexical challenge for ELLs contained in item 8, since all content words (except vacuum) are high-frequency words known by at least two-thirds of fourth graders tested in the U.S. (Chall & Dale, 1995; Dale & O’Rourke, 1979).

**Low DIF Items**

This section discusses in depth a subset of items that show slight degrees of DIF disfavoring ELLs. Items 7 and 5 are offered as examples of a larger group of items flagged as category B DIF.

— Low DIF Item 7

**Differential item functioning.** On average, the non-ELLs’ odds of responding to item 7 correctly are 1.3 times the odds of ELLs doing so (see Figure 9).

**Learning strand.** Item 7 assesses the learning strand number sense and operations, specifically children’s understanding of the relationship between common fractions (MDOE, 2000).

**Linguistic complexity.** This item is characterized by the use of long noun phrases (see Figure 8). The subject of the second sentence is a noun phrase that includes a chain of two prepositional phrases. Most of the content words appear in the 3,000-word list commonly known by fourth graders (Chall & Dale, 1995).

**Think-aloud interviews.** The complex noun phrase in the second sentence was particularly difficult to understand for the children in the think-aloud interviews, as shown in the following transcripts.

<table>
<thead>
<tr>
<th>Child 6:</th>
<th>Duncan paso chibolas azules y verdes en la bolsa.</th>
<th>Duncan put green and blue marbles in the bag.</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Exactly” … no sé qué es “exactly” …</td>
<td>“Exactly” … I don’t know what exactly is …</td>
<td></td>
</tr>
<tr>
<td>Tres cuartos de las chibolas … en la bolsa azul.</td>
<td>Three-quarters of the marbles … in the blue bag.</td>
<td></td>
</tr>
<tr>
<td>¿Cuál … puede ser el total de números de chibolas en la bolsa?</td>
<td>Which … could be the total of numbers of marbles in the bag?</td>
<td></td>
</tr>
<tr>
<td>No sé qué es “following.”</td>
<td>I don’t know what “following” is.</td>
<td></td>
</tr>
<tr>
<td>Creo que me están preguntando</td>
<td>I think they are asking me</td>
<td></td>
</tr>
<tr>
<td>¿Cuántas chibolas hay ahí?</td>
<td>How many marbles are there?</td>
<td></td>
</tr>
<tr>
<td>Son tres cuartos chibolas.</td>
<td>There are 3/4 marbles.</td>
<td></td>
</tr>
<tr>
<td>No entiendo esto.</td>
<td>I don’t understand this.</td>
<td></td>
</tr>
</tbody>
</table>

[Pointing to Exactly 3/4 of the marbles in the bag are blue.]

In addition to not knowing the meaning of exactly and following, Child 6 thought that the adjective blue referred to the noun bag instead of the noun
FIGURE 8  Item 7

Duncan put blue marbles and green marbles in a bag. Exactly \( \frac{3}{4} \) of the marbles in the bag are blue. Which of the following could be the total number of marbles in the bag?

A. 3  
B. 6  
C. 7  
D. 8


FIGURE 9  Empirical Item Characteristic Curve — Item 7. Category B DIF

M-H Odds ratio = 1.3

marbles. This caused him to be confused about the question because he understood the initial statement to mean there are \( \frac{3}{4} \) marbles in the blue bag. The child in the next transcript knew the words that were unknown to Child 6 (exactly and following) but was still unable to comprehend the sentence.
Duncan puso azules … marbles y verdes en una bolsa.

Exacto 3/4 de las canicas en la bolsa azul …

¿Cuál de los siguientes puede ser el número de canicas en la bolsa?

Duncan put blue marbles and green in a bag.

“Exact” 3/4 of the marbles in the blue bag …

Which of the following can be the number of marbles in the bag?

Interviewer: ¿Cómo lo respondes?

Child 7: No lo entiendo.

Interviewer: ¿Qué es lo que no entiendes?

Child 7: Es difícil de entender, después de “exactly.”

Most of the children interpreted the sentence Exactly 3/4 of the marbles in the bag are blue, to mean 3/4 of the marbles in the blue bag. Confused by the proximity of the verbal phrase are blue to the noun bag, the children attributed the quality of blue to the bag rather than to the more distant noun marbles, making the sentence unintelligible. Children’s inability to understand the sequence of the syntactic constituents of this important clause resulted in an incorrect representation of the problem.

In addition, half of the children interviewed could not interpret the meaning of the adverb exactly. The word is a Spanish-English cognate. However, it is likely that these Spanish-speaking ELLs would be more familiar with the adjective exacto (exact) than with the adverb exactamente. Even if they knew the Spanish adverb, they may not have seen it frequently enough in print to be able to recognize that exactly and exactamente share the same root. Interpreting what exactly means in this context is key to answering the question correctly because the construction Exactly 3/4 indicates that the total number of marbles must be a multiple of the number 4.

— Low DIF Item 5

Differential item functioning. On average, the odds that non-ELLs respond to item 5 correctly are 1.2 times the odds for ELLs along the total math score distribution (see Figure 11).

Learning strand. Item 5 measures the learning strand number sense and operations, specifically the use of appropriate operations to solve problems, which here involve money (MDOE, 2000).

Linguistic complexity. The item stem is relatively long (forty-one words, including numbers; see Figure 10). The first two sentences average eight words per sentence and have simple noun-verb-direct-object structures. The last sentence, however, is fairly complex and long (twenty-five words). It has one introductory conditional clause with a long and complex noun phrase that includes the expression coupon for $1.00 off. The polysemous off is a high-
frequency word that appears in the list of 3,000 words known by 80 percent of fourth graders (Chall & Dale, 1995). Its specific meaning in context (taken away) was readily understood by 86 percent of the fourth graders sampled for the LWV inventory (Dale & O’Rourke, 1979). In contrast, the word owe is low frequency; it does not appear in the 3,000-word list. Neither does coupon (Chall & Dale, 1995), though coupon is a familiar term known by 79 percent of fourth graders in the LWV list.

Think-aloud interviews. The think-aloud interviews revealed that the item’s use of cultural references (coupon for $1.00 off), conditional clauses (If he uses
. . .), and critical unfamiliar words (owe) were barriers to item comprehension, as illustrated below:

Child 8: Miguel quiere 3 bolsas de “potato chips.”

Each bag of potato chips, um … costs 2 dólares con sesenta y nueve centavos.

And if he is going to use …

[Pause as the child stumbles for words.]

Coupon of 1 dollar.

¿Qué “price,” precio de la bolsa… qué precio de la bolsa?

¿Cuánto Miguel va, va, a, a … “owe, owe” para las 3 bolsas de potato chips?

Probing questions revealed that the boy was familiar with the verbs buy and costs and the nominal potato chips. He was not familiar with the words coupon and owe or the expression coupon for $1.00 off the price. He was not able to figure out the meaning of the polysemous word off in context, and got confused about whether he was being asked about the price of the bag. His limited vocabulary, background knowledge, and familiarity with the usage of the English language hindered his comprehension of the problem.

Interviewer: ¿Sabes qué es “owe?”

¿Qué crees que te están preguntando?

¿Cuánto Miguel va, va, a, a … de vuelta?

¿Cuánto le van a dar de vuelta?

Do you know what “owe” means?

What do you think they are asking you here?

Yes, I think they are asking me,

How much will Miguel … back?

How much [change] will they give him back?

Discussion

Previous research has found that greater linguistic complexity increases the difficulty of math word problems for ELLs as compared with non-ELLs of equivalent math proficiency (Martiniello, 2006). Through textual analyses and children’s responses to think-aloud protocols, this study sought to illustrate some of the linguistic characteristics of math word problems in the 2003 MCAS math test that posed disproportionate difficulty for ELLs.
Linguistic Features of DIF Items Disfavoring ELLs

Items identified as showing evidence of DIF disfavoring ELLs were analyzed for challenging syntactic and lexical features. Findings indicate that these items share some of the following characteristics hindering reading comprehension:

— Syntax

• **Multiple clauses.** Item sentences have multClausal complex structures with embedded adverbiaL and relative clauses, which are more difficult to understand than other types of clauses (e.g., coordinate).
• **Long noun phrases.** Long phrases with embedded noun and prepositional phrases lead to comprehension difficulties.
• **Limited syntactic transparency in the text,** such as a lack of clear relationships between the syntactic units. This point is related to the previous two. Embedded clauses along with long noun phrases between syntactic boundaries obscure the syntactic parsing of the sentence. During think-aloud responses to items 2 and 7, this lack of transparency prevented some ELLs from interpreting the syntactic boundaries in the text correctly, resulting in a distorted interpretation of the string of words in the item.

— Vocabulary

• **Some words in these DIF items are unfamiliar to most fourth graders,** according to the LWV list (Dale & O’Rourke, 1979). Other words more commonly known by English speakers, such as *chores* and *certain*, proved to be quite challenging for the ELLs interviewed. The role of vocabulary knowledge as a predictor of reading comprehension for both ELLs and non-ELLs is well established. Since, by definition, ELLs do not have the breadth of vocabulary that fully proficient students do, more words will be unknown to them. Sentences in some of these DIF items have too many words unknown to ELLs, thus making it very difficult for them to infer the meanings from the context.
• **These results suggest that ELLs might have difficulties not only with more-sophisticated academic words,** but also with words usually learned at home (e.g., *chores*). Most likely, these children will know those words in the language they speak at home with their parents.
• **Polysemous words,** those with multiple meanings, pose additional challenges to reading comprehension. In some DIF items, children could not figure out the syntactic function of a polysemous word in the sentence and ended up misinterpreting the word’s appropriate meaning in context, as when they confused *one* as a pronoun with *one* as a numeral in item 2.
• **Since their English vocabulary lacks depth,** ELLs may know only the most familiar connotation of a polysemous word. In some DIF items, children tended to assign the most familiar meaning to such words regardless of
their context, misinterpreting the text. For instance, in the phrase the pictograph below shows the amount of money each fourth-grade class raised for an animal shelter (item 18; MDOE, 2003b), some students interpreted the verb to raise to mean to increase or to rear animals rather than to raise funds.

- English words and expressions that signify particular referents of mainstream American culture may be unfamiliar to ELLs, as in the case of coupon in item 5. Another example was spelling bee championship in item 34 (MDOE, 2003b). In these examples, students’ lack of background knowledge interferes with their text comprehension.

- Not having one-to-one correspondence between the syntactic boundaries of clauses and the layout of the text in the printed test may challenge ELLs who exhibit poor reading fluency, particularly when an item contains complex sentences with multiple clauses and long noun phrases. For instance, in item 2 the line breaks between the words even and number and between one and shown below may have hindered ELLs’ identification of the grammatical structures. This is consistent with research on the relationship between visual-syntactic text formatting and reading comprehension (Walker, Schloss, Fletcher, Vogel, & Walker, 2005).

- As expected, the think-aloud interviews suggest that Spanish-speaking ELLs do indeed take advantage of their knowledge of Spanish-English cognates for interpreting an item’s meaning. However, it seems that this only applies to high-frequency Spanish words that are likely to be familiar to children. In these cases, children were able to recognize the morphological relationship between the English word in the test and the Spanish word they know (e.g., impossible/imposible). Other English-Spanish cognates that are less familiar in Spanish and that have more familiar substitutes were difficult for these children (e.g., certain/cierto).

This study’s think-alouds involved only Spanish-speaking ELLs. Nationwide, Spanish-speaking students constitute about 80 percent of ELLs in public schools (Kindler, 2002). Although this study did not examine differences among ELLs from different language backgrounds, we may expect that the linguistic features discussed here (including cultural referents) will hinder reading comprehension for all ELLs regardless of their primary language. In addition to looking at language-specific difficulties, further studies should investigate how generalizable these findings are to other groups of ELLs. For instance, the finding related to cognates may only be generalizable to Romance and Germanic languages but not to other languages.

Sources of Item DIF: Linguistic Complexity and Learning Strands

In general, the empirical evidence in this study tends to confirm the “linguistic complexity as one source of DIF” hypothesis. A plausible explanation for the disproportionate difficulty of DIF items disfavoring ELLs compared with
non-ELLS lies in differences in the groups’ abilities to understand the English text in these linguistically complex items. Although the present study did not analyze the linguistic features of items showing no DIF for ELLs, it builds on previous research on the relationship between measures of item linguistic complexity and DIF in the 2003 MCAS math test (Martiniello, 2006, 2007). The aforementioned studies confirmed that math word problems with low linguistic complexity do not tend to show DIF disfavoring ELLs, unlike items with greater linguistic complexity. Increased complexity seems more related to ELLs’ text comprehension than to non-ELLs’.

However, linguistic complexity is not the only plausible source of DIF disfavoring ELLs in the items analyzed. Previous studies have identified important associations among curriculum learning strand, linguistic complexity, and DIF for ELLs in the MCAS fourth-grade math test. Compared with the other learning strands in the test, *data analysis, statistics, and probabilities* items tended to be unusually more difficult for ELLs than for non-ELLS with equivalent math scores (Martiniello, 2006, 2007). In this study, half of the items identified as showing DIF disfavoring ELLs measure *data analysis, statistics, and probabilities*. The fourth-grade MCAS math test has a total of seven items covering this strand, and five of them were flagged for DIF disfavoring ELLs over non-ELLS (two with substantial and three with low DIF indices).

A possible explanation for the differences in difficulty for ELLs and non-ELLS in these items may be the differential teaching and learning of *data analysis, statistics, and probabilities* subject matter across groups. It is possible that the ELLs tested in the spring of 2003 had received less exposure to the curricular content of this learning strand than were their non-ELL counterparts, or, having had the opportunity to learn it, they did not master and acquire this knowledge as non-ELLS did for a variety of reasons about which we can only speculate. For example, their low English proficiency may have prevented them from understanding or learning the content during classroom instruction. Or they may have had teachers who were less prepared to teach this content. Research does show that teachers of ELL students have less mathematical training and lower certification rates (Abedi & Gándara, 2006; Gándara & Rumberger, 2003; Mosqueda, 2007). Nevertheless, the results of this study suggest that the unusual difficulty of this learning strand may lie in the greater syntactic and semantic complexity of items measuring *data analysis, statistics, and probabilities* as compared with items measuring the other learning strands. Martiniello (2006) found a significant and positive correlation between this learning strand and a composite measure of linguistic complexity in the fourth-grade MCAS math test (*r* = 0.455, *p* = 0.007). Further studies are needed to examine the interaction between learning strands and linguistic complexity as sources of DIF disfavoring ELLs. There may be important differences among the items measuring *data analysis, statistics, and probabilities* related to their linguistic complexity and not to their
subject matter, as illustrated by the comparison of items 8 and 30. Since both items measure exactly the same ability, variations in the groups’ knowledge of the curriculum content or learning strand are not that helpful in explaining their strikingly different DIF statistics. Item 8’s disproportionate difficulty for ELLs relative to that of item 30 is likely to be related to its unfamiliar vocabulary. Children who speak a language other than English at home are less likely to be exposed to the word chores (and types of chores: rake, weed, and dust) than English speakers, who may hear these words when interacting with their parents regarding their household tasks. In contrast, the English words pencils, notebooks, students, and colors are unarguably among the first English words ELLs will learn in the classroom setting. The differential exposure of ELLs and non-ELLs to the lexical terms in item 8 may explain why this item functions so differently across groups in contrast to item 30, even though the items measure the same curricular content.

What Is Our Inference Based on Scores?

Given the impact of linguistic complexity and curriculum content on the differential item difficulty of ELLs and non-ELLs with equal math ability, how should we interpret incorrect answers by ELLs in these DIF items? Ideally, an item should be answered incorrectly only if the student has not mastered the curriculum content measured by the item. However, the evidence from think-aloud interviews shows that our inference based on scores can be distorted by ELLs’ unfamiliarity with the English vocabulary, lack of background knowledge, and difficulty in parsing complex syntactic structures in the item. For instance, the erroneous interpretations of the word owe and the expression coupon for $1.00 off in item 5 make answering the item correctly extremely difficult regardless of the child’s mastery of the addition/multiplication/subtraction operations involved in solving this problem.

However, if we interpret this child’s zero score as an indication that he does not know how to apply operations to solving everyday transactions involving coupons, like purchasing potato chips in a U.S. supermarket, then our inference may be mostly correct. The lack of vocabulary and familiarity with U.S. culturally based coupons would likely prevent this child from either figuring out the right amount to pay or from taking advantage of the coupon discount. However, our inference may be incorrect if we assume that the child does not know how to solve problems involving money when he understands the conditions of the transaction. It is also reasonable to think that a store transaction would provide more contextual information for deriving the meaning of unknown words than could the highly decontextualized text in the item.

In item 2, a score of zero would suggest that a student does not know how to classify events as certain, likely, unlikely, and impossible. However, the think-aloud interviews revealed that some ELLs who knew the mathematical meaning of the English words likely and unlikely, and could correctly classify the likelihood of a particular event, could not provide the correct answer (likely) because
they were unable to understand the complex sentence structures providing them with the information about the event to be classified.

Many of the children interviewed did not know one or more of the English words certain, likely, unlikely, and impossible. Based on the Massachusetts mathematics curriculum framework, one could argue that these English words are in fact the mathematical terms the item intends to measure; thus, if a child does not know these words and answers the item incorrectly, the response would appropriately reflect that child’s inability to classify outcomes as certain, likely, or unlikely. However, think-aloud responses indicate that some of the ELLs interviewed had a mathematical understanding of a gradient of likelihood for a given event. They could express it in Spanish but had not yet mastered the English vocabulary required to label it correctly on the test item. Furthermore, some of these ELLs had actually learned to classify the likelihood of events using more-familiar English words than those used on the item. For instance, some children construed a likelihood continuum ranging from always to never in which certain corresponds to it will always happen, impossible to it will never happen, likely to it will almost always happen, and unlikely to it will almost never happen.

Since the words certain, likely, and unlikely have common meanings in the mathematics classroom and in everyday social language, it is reasonable to think that children who are fully proficient in English would have a greater chance than ELLs to infer the correct mathematical meanings of these words. For instance, the word certain is listed as a word known by 80 percent of English-speaking fourth graders, but in the group of ELLs interviewed only a few children knew it. This differential familiarity with the item’s vocabulary may be contributing to the large DIF in item 2.

Based on these findings, we may expect the differential performance of ELLs and non-ELLs on this item to decrease if the distractors included more familiar English words and the clauses in the item stem were less complex. The real challenge is reducing the item’s excessive linguistic complexity while preserving the integrity of the mathematical construct the item intends to measure. More research is needed in the area of linguistic modification. So far, the results of simplification studies comparing the performance of ELLs in complex and simplified versions of math word problems have been mixed (Abedi, Hofstetter, Baker, & Lord, 2001; Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006; Moreno, Allred, Finch, & Pirritano 2006).

It would be advisable to conduct some DIF studies comparing Spanish-speaking ELLs (as the focal group) who took the 2003 MCAS math test in Spanish with those who took it in English (as the reference group) to see if DIF was also present across these groups. This would help us to better understand whether the large DIF indices for some items are due to curricular differences or to difficulties comprehending English text. One would expect that the Spanish version of these linguistically complex items would be more comprehensible than the English version for Latino ELLs. However, some research
shows that this might not be the case when the subject matter assessed in the test is taught in English rather than in Spanish (Llabre & Cuevas, 1983; Wong-Fillmore, & Valadez, 1986).

Also, since ELLs vary in their English proficiency, future DIF studies should account for this. For instance, former ELLs were not included in this study's DIF sample. Conducting DIF studies with former ELLs as a reference group for ELLs or as a focal group for non-ELLs would allow us to investigate the generalizability of these findings to former ELLs who have been reclassified as non-ELLs.

More studies using think-aloud protocols should be conducted to investigate how ELLs interpret math word problems of varying linguistic complexity. A limitation of the present think-aloud study is the fact that the test used is available to the public. In fact, some teachers use previously released MCAS items to prepare their students for the fourth-grade MCAS test (which is high-stakes due to the AYP-associated sanctions). In this study, teachers regularly asked children if they had previously seen an item in an attempt to avoid presenting it again. But despite these precautions, the results might overestimate ELLs' understanding of the items.

**Implications**

This study has implications for test development, test validation, and math instruction for ELLs. Item construction for achievement tests often relies on content-area specialists who are instructed in long-standing principles of item writing: Write clearly and concisely and avoid irrelevant information (Baranowski, 2006). However, these item writers may not be aware of the specific consequences of excessive linguistic complexity on the performance of ELLs. Further training of item writers and professional test editors is needed to address this void. For instance, item writers could be provided with item performance indicators like DIF and interview transcripts from cognitive laboratory studies. Abedi (2006) recommends providing them with hands-on exercises in linguistic modification — that is, “simplifying or modifying the language of a text while keeping the content the same” (p. 384).

It is critical that the language simplification is not achieved at the expense of altering the construct or skill to be measured by the item or test. Proponents of linguistic modification do not advocate using language-free math tests for ELLs. Mathematical discourse in classrooms and textbooks combines natural and academic language, mathematical terms, symbols, and graphs. Therefore, math assessments, particularly those designed to assess “mathematics for understanding,” should do the same.

Math word problems like those on the MCAS test attempt to measure mathematical understanding by providing scenarios and novel situations in which students apply their prior knowledge and establish relationships between mathematical concepts and ideas. We want to know if our students can do that. In contrast, a test consisting of computation problems with little or no
Language and the Performance of ELLs in Math Word Problems
MARIA MARTINIELLO

Language interference would provide quite a narrow picture of the mathematical knowledge of ELLs. Still, the language of these math word problems must be scrutinized carefully and their differential functioning for ELLs studied before including them in operational tests. Studies routinely done to identify DIF for various gender and race groups should be extended to ELLs.

Cognitive laboratory research is needed to learn how ELLs interpret the text of math word problems. Relying on experts’ reviews of items may not be sufficient when developing or evaluating assessments for this population. Experts do not always anticipate the actual comprehension challenges ELLs encounter when reading the test. Although expensive, think-alouds are an important tool for examining the validity of test scores for ELLs.

This research also has implications for teachers. It confirms how important language skills are for understanding and solving mathematical problems in large-scale assessments. Thus, the teaching of mathematics to ELLs can no longer be perceived as separate from the teaching of language. Research on teachers’ perceptions has found some contradictions in the way teachers conceive of mathematics instruction (as free from language) and the kind of math assessments they use in their classrooms (with great language demands) (Bunch, Aguirre, Telléz, Gutiérrez, & Wilson, 2007). Teachers must provide sustained linguistic scaffolding for ELLs while encouraging the development of their mathematical meaning-making skills (Anhalt, Civil, Horak, Kitchen, & Kondek, 2007).

This study sought to inform the field and refine our testing practices, from the way we construct tests and write math items for ELLs to the way we gather validity evidence to support our inferences about ELLs’ math knowledge. To disentangle ELL’s math and English language proficiency in our interpretation of test scores, this research integrated three sources of information: a large-scale psychometric analysis of differential item functioning (DIF) for ELLs and non-ELLs; analysis of the language in the math items by linguists and literacy experts; and think-aloud responses to these items by ELL students. Ideally, these three sources of validity evidence should be routinely examined by testing agencies to guarantee the fairness of mainstream assessments used for ELLs. Recently researchers have started to systematically examine DIF for language proficiency groups, but this research lags decades behind that on race and gender DIF. Identifying the presence of DIF, however, is not sufficient. Studies are needed to understand the sources of DIF. Witnessing the thinking process of ELLs making sense of math word problems can illuminate this process. Although expensive, think-aloud protocols can be an invaluable tool for investigating the appropriateness of inferences about ELLs’ math knowledge based on scores. Test scores are used to make important decisions about ELL individuals and groups, from graduation requirements to track placement, to AYP calculation under No Child Left Behind. Therefore, it is crucial that thorough validity studies be conducted to ensure equity and fairness for ELLs taking these tests.
Notes

1. Between-group differences in the difficulty of a given item for two groups with different distributions of math proficiency do not represent DIF. Because of ELLs' lower overall proficiency in math, individual math items are on average more difficult for ELLs than for non-ELLs. This differential performance, which reflects the main effect of group membership, is called item impact and does not constitute DIF (Dorans & Holland, 1993; Hambleton, Swaminathan, & Rogers, 1991; Wainer, 1993). Conceptually, when we test for DIF across groups, we do not compare all ELLs versus all non-ELLs. Instead, we compare only those individuals from these two groups who have similar total test scores in math. If the proportion of correct response to a dichotomously scored math item is the same for ELLs and non-ELLs with equivalent math proficiency (i.e., if conditional item \( p \) values are the same across groups), then the item shows no DIF; if they are different, the item shows DIF.

2. The standardization procedure was implemented in two stages. In the first stage, standardized \( p \) differences (STND P-DIF) were estimated for each item using total test score as the matching criterion (39 items). The STND P-DIF is the weighted average of conditional \( p \) value differences across groups where the weights were the relative frequencies of math scores in the ELL group. In the second stage, items showing STND P-DIF with absolute values equal or greater than 0.075 were removed from the matching criterion. New DIF indices were estimated using the refined criterion (36 items). Likewise, Mantel-Haenszel statistics were first estimated for the total test score and the refined criterion.

3. Any opinions expressed in this article are those of the author and not necessarily of the Educational Testing Service.

4. MCAS results are reported according to four performance levels: advanced, proficient, needs improvement, and warning.

5. Although both items’ DIF indices fell within the ETS category B (slight to moderate DIF), an examination of the size of their DIF statistics using the standardization method is informative. Item 30 has a negligible STND P-DIF, while item 8 has a STND P-DIF equal to 0.11 favoring non-ELLs over ELLs.

References


This research was generously supported by the Spencer Foundation Doctoral Dissertation Fellowship, the Harvard University Achievement Gap Initiative, and the Harvard Graduate School of Education Dean’s Summer Fellowship. The author wishes to thank the Massachusetts Department of Education, and both the Boston and the Cambridge Public Schools for making this research possible; and Catherine Snow, Daniel Koretz, Jamal Abedi, and Paola Uccelli for their feedback on earlier versions of this manuscript. The author is indebted to many experts who contributed to various parts of this research, in particular to Paola Uccelli, who conducted the test text analysis/coding, to Young Suk Kim, who served as second coder, and to Rob Keller for assistance with the figures.

This research was conducted while the author was a member of the Harvard Graduate School of Education. Any opinions expressed in this article are those of the author and not necessarily of Educational Testing Service.