Understanding Differential Item Functioning for English Language Learners:

The Influence of Linguistic Complexity Features

By

Jessica M. Loughran

Submitted to the graduate degree program in Educational Psychology and Research and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

____
Chairperson Neal Kingston

____
William Skorupski

____
Vicki Peyton

____
Bruce Frey

____
Phil McKnight

Date Defended: 28 October 2014

The Dissertation Committee for Jessica M. Loughran
certifies that this is the approved version of the following dissertation:

Understanding Differential Item Functioning for English Language Learners:

The Influence of Linguistic Complexity Features

Chairperson Neal Kingston

Date approved: 20 November 2014

Abstract

Items on math assessments often include complex language that is not related to math content. English language learners (ELLs) are still acquiring proficiency in English, so the linguistic complexity of items is particularly troublesome for this group. Language demand on items can lead to lower scores for ELLs than their non-ELL peers. The present paper examines math items for evidence of differential item functioning (DIF) against ELLs using logistic regression. This study also used the general linear model to explain the DIF statistics from specific item characteristics. Results indicated that some item-level features predicted total DIF, but results for Grade 4 and Grade 8 were not consistent. Limitations and implications are discussed.

Keywords: Differential item functioning, linguistic complexity, English language learners

Acknowledgements

So many people provided support, mentorship, and assistance throughout the months of research and writing. Many thanks to Neil and Neal. Neil Loughran: Thank you for your limitless patience, words of encouragement, and for always having the appropriate Irish-ism at the ready. Neal Kingston: I owe so much of my success to you. Thank you for bringing me to the "dark side" of REMS; I appreciate your mentorship, support, and tireless optimism every day.

Thank you to Billy Skorupski for your guidance during our research project on which this dissertation is based. Many thanks to the rest of my committee for their time and feedback: Vicki Peyton, Phil McKnight, and Bruce Frey. I am also very grateful to Marianne Perie for her support and mentorship at CETE over the last few years.

I appreciate the time and expertise provided by Jennifer Brussow: thank you for your linguistic genius and for occasionally providing a much-needed respite from work. Thanks to Sukkeun Im for your assistance with coding and for generally being so pleasant; Kyle Consolver for assistance with item coding; and Elizabeth Kanost for providing your editing expertise. Stephanie Howarter, Cherie Oertel, and Wenhao Wang: thank you for the laughs and support and for being such good friends.

Last, but certainly not least, to my family for your support over the years: Steve, Jean, Mogie, Betty, and in memory of Jesse "Mutt" Thomas, Jesse Eugene Thomas, Ila Smith, and Minnie Kensinger.

Table of Contents

Chapter I: Introduction

For years, educators and measurement professionals have questioned the validity of standardized tests for students who are still learning English. As specified in the No Child Left Behind Act ("No Child Left Behind (NCLB) Act of 2001," 2002), states are required to include English language learners (ELLs) in annual state assessments for all content areas. However, ELLs often lag behind their native English-speaking classmates and do not perform as well on standardized tests. Given the specific educational issues that ELLs face, the onus is on educators and test developers to ensure these assessments are appropriate for ELL students.

The terms limited English proficient (LEP) and English language learner (ELL) are both used to describe students who are still acquiring English proficiency. The U.S. Department of Education further defines ELLs as students “who are being served in appropriate programs of language assistance” such as courses in English as a second language (US Department of Education National Center for Educational Statistics, 2013). In the 2010–2011 school year, ELLs made up about 10% (4.7 million) of U.S. public school students. These figures have increased from the 2002–2003 school year, in which ELLs made up about 9% (4.1 million) of U.S. public school students. Numbers of ELL students tend to be higher in lower grades, with recent public school data showing 14.1% of elementary students and 6.5% of secondary students were LEP during the 2007–2008 school year (U.S. Department of Education, 2007-08).

In the classroom, ELLs face the challenge of learning academic content while they are still acquiring English language skills. Achieving proficiency in academic English can take an estimated four to seven years (Hakuta, Butler, & Witt, January 2000). For many ELL students, though, learning the specific academic language needed for various content areas can take much longer. During this period of language learning, ELLs may not have the academic English they need to acquire content knowledge. Compared with their native English-speaking classmates, there is less opportunity for ELLs to learn the types of academic language that appear on content assessments (Bailey & Butler, 2003).

Indeed, there is a great deal of evidence that ELLs tend to not perform as well as their non-ELL peers on large-scale, standardized assessments. The performance gap between ELLs and native English speakers is widely documented across content areas, including reading comprehension, writing, math, social studies, and science (Abedi, 2002; Abedi et al., 2005; Abedi & Lord, 2001; Abedi, Lord, Hofstetter, & Baker, 2000; Bailey, 2005; Young et al., 2008). Evidence of this performance gap has led educators and researchers to question the appropriateness of these tests for students who are not yet proficient in English.

Assessment Validity for ELLs

According to Messick (1990), "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 1). In the process of evaluating validity evidence, researchers should consider the performance of subgroups in the target population. The inferences based on test scores for students in minority groups may differ from those of the majority group. Specifically, there is a growing body of research that indicates content assessments lack sufficient evidence of validity to be used with ELL populations.

On math assessments, test scores should indicate a student's level of math ability. This inference may be flawed for ELL students, however. Math test scores often reflect students' math ability as well as their ability to read and understand math items in their non-native languages. A test's language demand can have a significant effect on ELL students' performance. For assessments not intended to measure language ability, “the linguistic or reading demands of the test should be kept to the minimum necessary for the valid assessment of the intended construct” (AERA, APA, & NCME, 1999, p. 82).

Unfortunately for ELL test-takers, high language demand is still evident on content assessments, even assessments that are not intended to measure language proficiency. Language demand has become a ubiquitous source of construct-irrelevant difficulty for ELLs on math assessments in particular. The language proficiency needed to comprehend math items contributes to noise in the measurement of ELL students’ math content knowledge. As such, language demand has been the focus of several studies that investigated the performance gap between ELLs and non-ELLs.

Language Demand and the Performance Gap

The link between language demand and ELL test performance is well documented. For example, Bailey (2005) found that the performance gap between ELLs and non-ELLs was higher on assessments with greater language demand. Similarly, in a large study of K–12 students, Abedi (2002) found a significant performance gap between ELL and non-ELL students. Students’ language proficiencies had a stronger impact on test scores than other background variables, such as family income or parental education. Other studies provide evidence that language demand on content assessments contributes to measurement error for ELLs (Abedi & Lord, 2001; Solano-Flores & Trumbull, 2003).

Abella, Urrutia, and Shneyderman (2005) demonstrated the influence of language proficiency in assessing math content. They administered the same math test to ELLs in English and in the students’ native language of Spanish. ELL students performed better on the math tests administered in Spanish. Research has also shown that, when the language on test items is modified to be less complex, the gap between ELLs and other students can be reduced (Abedi et al., 2000).

Language clearly affects ELL students' test performance and can adversely influence the measurement of their content knowledge. Test developers are particularly interested in identifying the specific linguistic features that are problematic for ELLs. If these features are known, test items can be written with more universally designed language, and ELLs will have a better chance of accessing an item's content.

Linguistic Complexity

Linguistic complexity is the amount of academic language, or language load, on a test. Heavy language load can reduce reading speed, increase cognitive demand, and hinder students' understanding of test items (Abedi, 2004; Butler, Bailey, Stevens, Huang, & Lord, 2004; Carpenter, 1980). In essence, linguistic complexity can impede ELL students' abilities to display their content knowledge.

Researchers have identified several features of linguistic complexity that commonly appear on math content assessments and influence ELL students' performance. These features range from item length in words or sentences to the types of words and grammatical structures in an item. There are numerous grammatical features of linguistic complexity, and these features involve everything from phrasing (e.g., noun phrases, prepositional phrases) to verb forms (e.g., passive voice) and sentence type (e.g., compound, complex, compound/complex).

The findings from this body of research indicate that math tests include many features of linguistic complexity (Abedi & Lord, 2001; Bailey, 2005; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006). Furthermore, some studies have shown a link between differential performance of ELLs and non-ELLs and the complexity of language within an item (Haag, Heppt, Stanat, Kuhl,

& Pant, 2013; Mahoney, 2008; Martiniello, 2008, 2009; Wolf & Leon, 2009). Although these studies linked language demand to ELL students' performance on math items, the results from these studies are inconclusive. It is unclear, for example, what specific aspects of language are problematic for ELLs.

The present study expands on previous research by investigating specific item characteristics to explain differential item functioning (DIF) against ELLs in math tests. In this study, statistical methods are used to predict the magnitude of the DIF statistics from several continuous and discrete variables. These variables include a diverse set of linguistic complexity features, as well as other features common in math items (e.g., visual features such as charts and graphs). Additionally, this study examines large sets of math items from assessments administered to Grade 4 and Grade 8 students. To help reduce the potentially confounding effect of background variables, the ELL and non-ELL comparison groups are controlled on socioeconomic status (SES) and native language.

The aim of this study is to reveal potential sources of measurement error on math tests for ELLs. Results from this study will add to the evidentiary argument for inferences made about ELLs' math test scores. Information from this study can help guide item writers and test developers in the construction of more linguistically accessible assessments for students who are still learning English.

The following literature review provides a summary of linguistic complexity, DIF studies involving ELLs, studies that investigated sources of DIF such as linguistic complexity and non-language features, and relevant background factors within the ELL group.

Chapter II: Literature Review

It is clear that English language learners (ELLs) do not perform as well as non-ELLs on content assessments, including math assessments. However, the reasons for this differential performance are less clear. Specific characteristics of math items, such as the language demand and the inclusion of visual aids, have been implicated as explanatory variables. Background characteristics such as a student's native language and socioeconomic status might also differentially influence an ELL student's test performance. The following section summarizes these bodies of research, identifies limitations and gaps in the literature, and builds upon this information to specify research questions for the present study.

Academic Language

On most standardized assessments in the United States, students need some degree of English language proficiency to comprehend the test items. It is often unclear if ELL students' test scores indicate their content knowledge, language proficiency, or a combination of the two. The confounding of content knowledge with language ability is exacerbated on test items with large amounts of academic language.

Academic language has been set apart from other, more basic forms of communication. Cummins (1980) distinguished between basic interpersonal communication skills (BICS) and cognitive/academic language proficiency skills (CALPS). BICS describes language used in ordinary social settings in particular contexts. For example, students use BICS when they interact with others on the playground. CALPS, on the other hand, is language used in academic settings. Bailey (2007) is careful to point out that CALPS is not necessarily more sophisticated or more cognitively demanding than BICS. Instead, CALPS and BICS are distinguished by the contexts in which they are typically learned. BICS is a context-embedded form of communication that develops within social interactions. In these social interactions, communicators receive feedback from one another. In contrast, CALPS is context-reduced and develops in academic contexts. Within these academic contexts, "…students receive fewer opportunities to negotiate meaning or to use contextual cues" (Bailey, 2007, p. 10).

Academic language is important across content areas, and the acquisition of CALPS is essential for academic success (Cummins, 1980). Compared with their native English-speaking classmates, ELL students have more difficulty acquiring and demonstrating academic language skills. It can take non-native English speakers four to seven years to develop the language proficiency needed across various academic contexts (Hakuta et al., January 2000). This disadvantages ELL students in classroom settings and in assessment situations.

Academic language skills are also associated with students' opportunity to learn. Because of their language deficiencies, ELL students may have more difficulty than their non-ELL classmates acquiring content knowledge. ELLs are disadvantaged because they often face very different learning environments than their non-ELL classmates (Rumberger & Gándara, 2004). For example, ELL students may receive less-rigorous math instruction or may receive instruction from teachers who are less qualified. Reduced opportunity to learn has shown to relate to poorer performance for ELLs on math tests (Abedi & Herman, 2010).

Opportunity to learn is a relevant factor in the performance gap between ELL and non-ELL students. There is also extensive evidence that item-level factors relate to the performance gap. Several studies revealed that the academic language in test items adversely affects ELL students' test performance. It is more difficult for ELLs to demonstrate their content knowledge on tests with high levels of academic language. In other words, the linguistic complexity of content assessments can be a source of construct-irrelevant variance. The following sections summarize this relationship between linguistic complexity and ELL students' test performance.

Linguistic Complexity

Linguistic complexity is a multi-faceted construct that includes dozens of features. Recent research conceptualizes linguistic complexity into categories of descriptive, lexical, and grammatical features (Abedi et al., 2005; Bailey, 2005; Butler et al., 2004; Haag et al., 2013). These features can be challenging for ELL students as they try to comprehend text.

Descriptive features are easily quantifiable. They include the number of words in an item, total number of unique words, number of sentences, average sentence length, and the number of paragraphs. The more words and sentences an item has, the more pieces of information the student must process. This slows reading and can cause confusion as students try to comprehend text; this effect is especially problematic for ELLs (Abedi, Lord, & Plummer, 1997; Butler et al., 2004).

Lexical features relate to the types of words in a test item. These features include general academic vocabulary, long words, words with multiple meanings ("polysemous" words), ambiguous words, and low-frequency words. General academic vocabulary is not associated with a particular content area and, as a result, these words are not typically targeted during classroom instruction. Conversely, content-specific words such as *circumference*, *perpendicular*, and *equation* are introduced and explicated during math instruction. Compared to content-specific vocabulary, general academic vocabulary and low-frequency words (i.e., words not commonly encountered in ordinary contexts) are less likely to be presented in a contextual way. The acquisition of these words often occurs outside of the school setting, and this can adversely affect ELL students. Indeed, research supports the idea that general academic words and low-frequency words pose special challenges for ELL students compared with native English-speaking students (Abedi & Lord, 2001).

The classification of words as general academic vocabulary is dependent on grade level. Consider the grade-level differences between sample words appearing on the Berkeley Unified School District's list of Grade Level Academic Vocabulary. The list includes the words *concurrent* and *alternative* for Grade 8 students. At Grade 4, the words *clarify* and *typically* are considered general academic vocabulary (Tugwell, 2013).

Linguistic complexity also includes many grammatical features. Studies that examine the grammatical complexity of items frequently include counts of passive verb tense (e.g., "The car was driven by Luisa" rather than "Luisa drove the car"). Grammatical features of items also include sentence types (e.g., complex sentences, compound sentences, and complex-compound sentences), conditional clauses, and relative clauses.

Prepositions (e.g., of, on, at) are grammatical features that appear frequently in math contexts. For example, it is common to see questions such as, "What is the sum *of* the two largest angles *in* the figure below?" Prepositional phrases can be quite long and confusing, and they can have very different meanings in math contexts than they do in other contexts. For example, the preposition *by* has multiple purposes: it can mean "beside," or it can be used to indicate method of transportation (e.g., to travel *by* bus). However, in a math context, (e.g., a rectangle is 2 by 5 inches) the preposition *by* is used to indicate the dimensions of a shape. The linguistic complexity features mentioned previously can increase the cognitive load of items, produce barriers to comprehension, and cause students to read more slowly (Abedi, 2004). In a study examining math items across several grade levels, Shaftel et al. (2006) found evidence that descriptive (e.g., number of words), lexical (e.g., ambiguous words), and grammatical (e.g., prepositions) features predicted the difficulty of math items.

Research on item modifications and accommodations provides further evidence of the impact of linguistic complexity on test performance. Several studies support the idea that linguistic modification of items can improve ELL students' performance (Abedi & Lord, 2001; Abedi et al., 1997). Abedi et al. (2000) contended that linguistic modification as an accommodation strategy can help narrow the performance gap between ELL and non-ELL students.

Linguistic complexity in math items. Math tests are often assumed to have lower language demand than other content assessments, such as reading comprehension tests. However, several studies show that math items regularly include features of linguistic complexity. Researchers have also examined the relationship between the linguistic complexity features of math items and ELL students' math test performance. For instance, Carpenter (1980) found that ELL students performed much worse on math word problems than they performed on similar items presented in numeric format.

Several studies analyzed the descriptive features of math items. These studies explored length of the item stem (i.e., the question/problem to be solved along with the setup to the question) and foils (i.e., answer choices). Item length measures included the number of words, number of sentences, and average sentence length. Analyses of item length are important because longer items tend to be more difficult for students to read (Abedi, Leon, Wolf, & Farnsworth, 2008; Abedi et al., 1997; Butler et al., 2004). Lengthy items are especially problematic for ELLs, and item length has shown to contribute to differential performance between ELLs and non-ELLs on math tests (Abedi et al., 2008; Abedi et al., 1997; Haag et al., 2013).

Math assessments typically include items from several content areas such as algebra, geometry, and computation. Even math items that measure computation or algebra skills can be quite lengthy. Across several studies that described math item length, items for Grades 4 and 5 included an average of 18 to 40 words per item and about two to three sentences per item. Descriptive features for Grades 7 to 10 were higher, with an average of 30 to 60 words per item and about two to four sentences per item (Lee & Randall, 2011; Shaftel et al., 2006; Wolf & Leon, 2009).

Across these math content areas, items have shown to be complex in terms of descriptive features. On one statewide Grade 8 math assessment, the 19 items within the "Number and Computation" content area had between 5 and 46 words and between one and five sentences. However, there was considerable variation in item length between math content areas. The algebra math items ranged in length from 7 to 82 words and from one to six sentences (Loughran & Skorupski, 2014).

Mathematics also inherently involves specific, technical language. Schleppegrell (2007) contends that math has its own complex language with "characteristic patterns of vocabulary and grammar" (p. 142). For example, students must learn content-related words such as *parallel* and *fraction*. Students must also acquire new, math-related meanings for words they might already know. Schleppegrell provides two common examples: *borrow* and *product*. These words have different meanings within a math context and outside of a math context.

Words with multiple meanings, or polysemous words, are particularly troublesome for ELL students. ELLs face the task of learning multiple definitions for these words and then determining each word’s meaning from context. As Schleppegrell states, “Learning the new vocabulary that is centrally mathematical may be easier than learning the technical meanings for words that students already know in other contexts” (2007, p. 142–143). As one study showed, polysemous words caused native Spanish-speaking ELLs to misinterpret the meaning of some math items, and this negatively affected their test scores (Martiniello, 2008).

Other lexical features of linguistic complexity are common on math tests. Bailey (2005) analyzed the language demand of math items on a standardized Grade 11 assessment. This study revealed that more than 60% of the math items contained non-math-related general vocabulary. In a study that examined the linguistic complexity of math tests across three states and four grades (Grades 4, 5, 7, and 8), Wolf and Leon (2009) found that math tests contained a range of general academic vocabulary. Furthermore, academic vocabulary was the most common feature of linguistic complexity on the math tests.

The presence of non-math-related academic vocabulary on math tests is especially problematic for ELLs. Students tend to learn math-related vocabulary in the context of math instruction. However, the acquisition of general academic vocabulary, especially low-frequency words, can be more elusive. Low-frequency words are less common and less context-embedded than high-frequency words. In the development of test items, researchers recommend using high-frequency words to avoid unfairly disadvantaging ELLs (Abedi & Lord, 2001; Kopriva, 2000).

Grammatical complexity is another category of language demand that commonly appears on math items. Linguistic analyses of math items revealed many instances of prepositional phrases, passive voice, complex verbs, and long noun phrases in the items (Martiniello, 2008; Shaftel et al., 2006). Schleppegrell (2007) provides one math-related example of multiple prepositional phrases embedded in a lengthy noun phrase: “the volume of a rectangular prism with sides 8, 10, and 12 cm" (p. 143). These types of phrases are ubiquitous on math tests. One study revealed that as many as one half to two thirds of Grade 11 math test items were rated as having these types of complex grammatical structures (Bailey, 2005).

Test items that contain prepositional phrases, passive voice verb tense, or other complex verb tenses can be especially confusing for non-native English speakers. These complex grammatical features might be related to differential performance between ELLs and non-ELLs on math tests. Martiniello (2008) analyzed math items on which non-ELLs performed better than ELLs after each group was matched on math ability. She found that these items shared several characteristics, including long noun phrases with embedded prepositional phrases and sentences with multiple clauses.

Although there has been an abundance of research on math test linguistic complexity, the research has not always produced consistent results. Findings reveal great variation in terms of the types of linguistic complexity on math items and the effects of linguistic complexity on test scores. Shaftel et al. (2006) found that number of grammatical features (e.g., prepositions, pronouns, and complex verbs) in an item increased item difficulty. However, the impact of these language features did not seem to differentially affect ELL students. In the same study, researchers identified many ambiguous words on math assessments, but these words did not have a direct relationship to item difficulty for Grade 7 students.

Additionally, Lee and Randall (2011) found relatively low lexical and grammatical complexity on math items. In the same study, several individual features of linguistic complexity were seldom found on the items (e.g., passive voice, conditional clauses, relative clauses, nominalizations). In another study, the researchers were similarly unable to find many instances of passive voice verbs, compound sentences, and clause connectors on Grade 3 math test items (Haag et al., 2013).

Linguistic complexity and grade level. One factor that influences the amount of linguistic complexity on math items is grade level. In general, math tests at higher grade levels tend to have higher language demand (Wolf & Leon, 2009). This pattern emerges for descriptive, lexical, and grammatical complexity features. As mentioned in the previous section, Grade 7 and Grade 8 tests tend to have more words per item, more unique words per item, and more sentences per item than Grade 4 or Grade 5 tests. These results have been documented across states. However, the research is inconclusive about the significance of these grade-level differences (Abedi et al., 2008; Shaftel et al., 2006).

Lexical features also differ by grade. Compared with math tests for Grades 4 and 5, Grade 8 math tests had higher language load as evidenced by more academic vocabulary, higher number of unique words, and a higher proportion of academic words to total unique words (Abedi et al., 2008). Grammatical features are also more likely to appear on math tests at higher grades. This includes higher numbers of prepositional phrases and relative pronouns (Shaftel et al., 2006). Abedi et al. (2008) found that tests for higher grades had more complex sentence structures than tests for lower grades. Similar results have been found across states and content areas.

Clearly, linguistic complexity features vary across tests and grade levels. Even so, there is a great deal of evidence that math tests often have unnecessarily high language demand. Because ELLs are still learning English, it is reasonable to believe that they are more heavily impacted than non-ELLs by an assessment's language demand. Several researchers have confirmed this assertion through DIF analyses. This body of research is summarized in the next section.

Differential Item Functioning

Differential item functioning (DIF) studies are common in ELL research. DIF procedures compare the item-level performance of two groups (e.g., ELLs and non-ELLs) after those groups are matched on ability. For example, if ELLs and non-ELLs are matched on math ability, members from each group at each ability level should have the same probability of success on each math item. A test item shows evidence of DIF if, in the previous example, ELLs and non-ELLs demonstrate different probabilities of success on an item after they are conditioned on ability. When an item displays DIF, there is evidence that the item measures something other than the intended construct.

The matching criterion in a DIF study might be an observed variable (i.e., total test score) or a latent ability measure such as the item response theory (IRT) ability parameter. The comparison groups are called the *focal* and *reference* groups. The focal group (e.g., ELLs) is a collection of examinees that are expected to perform less well than the reference group (e.g., non-ELLs) after being matched on ability or total score.

Results from a DIF study can indicate statistical, differential performance between subgroups on a particular item. Upon finding evidence of DIF, researchers often want to explore the possibility of item bias. The distinction between DIF and item bias is important. DIF is "…a necessary but not sufficient condition for item bias," and item bias occurs when a test item "unfairly favors one group or another" (Clauser & Mazor, 1998, p. 31). Thus, items that are flagged as showing significant DIF are not necessarily biased items.

Item bias is determined by experts' judgments of item content, not by the statistical DIF study. These judgments take into account the DIF study results in addition to other, qualitative information (e.g., the presence of cultural, gender, racial, or other types of bias in the content of the test item). It is therefore important for test developers to identify sources of DIF in test items.

DIF analyses can yield several results. The analysis may indicate that an item: 1) does not show evidence of DIF or shows negligible DIF, 2) shows evidence of uniform DIF, or 3) shows evidence of non-uniform DIF. Uniform DIF means that one group is favored over the other group at every level of ability. As an example, a DIF analysis might show that non-ELLs have higher probability of success on a math item than do ELLs, and this effect is evident at low, medium, and high levels of math ability. With non-uniform DIF, the group that is favored is contingent upon ability level. For example, an item might favor non-ELLs over ELLs at low levels of math ability but favor ELLs over non-ELLs at high levels of math ability.

There are numerous statistical methods for detecting DIF. A full description of the multitude of DIF studies is beyond the scope of this paper, however. DIF procedures are generally classified as being parametric or non-parametric. Non-parametric methods typically condition the comparison groups on some observable measure of ability such as total test score. The most ubiquitous non-parametric DIF method is Mantel-Haenszel (MH), and many variations of MH have been developed. Other non-parametric methods include the standardization method, SIBTEST, and logistic regression.

Parametric DIF methods involve IRT analyses, which calculate item parameters (i.e., item difficulty, item discrimination, and item pseudo-guessing) that are independent of the sample of examinees. IRT analyses also allow for the calculation of examinee ability (i.e., the matching criterion), which is not dependent on the sample of test items. The following studies identified DIF against ELLs on math tests using a variety of DIF analyses. Many of these studies investigated linguistic complexity as the source of DIF.

DIF against ELLs on math tests. In a study that examined DIF between ELLs and non-ELLs across several grades, Abedi et al. (2008) found as many as 14% of math items showed evidence of DIF for Grade 7. For Grade 4, the number of DIF items was lower, with as many as 10% of the items displaying DIF. This study employed the use of three DIF methods: IRT-based likelihood-ratio, logistic regression, and MH.

Another study used both standardization and MH DIF procedures to identify DIF on a Grade 4 math test. Out of the 39 math items, 10 items showed evidence of DIF in favor of non-ELLs. Of those 10 items, nine items displayed “slight/moderate” DIF, and one item showed evidence of “large” DIF (Martiniello, 2008).

Wolf and Leon (2009) used logistic regression DIF analyses with math tests at Grades 4, 5, 7, and 8. They found that more math items displayed evidence of DIF when low English proficiency and high English proficiency ELLs were compared. The researchers generally identified more DIF items at higher grades, and this result was true across several comparison groups (e.g., ELLs vs. non-ELLs, accommodated ELLs vs. non-accommodated ELLs, low- vs. high-proficiency ELLs). These results are consistent with previous research that shows higher linguistic complexity on math tests at higher grade levels.

DIF against ELLs is not always evident on math tests, however, even when items represent a wide range of linguistic complexity. In one study, no evidence of DIF was found in favor of non-ELLs on a standardized math assessment. The math items in the study varied widely in terms of their descriptive and grammatical language features (Mahoney, 2008). The author did acknowledge several limitations to her study, including small sample size, some item misfit in the IRT analysis, and the possibility that English language proficiency variation within the ELL group might confound the results of the DIF study.

Similarly, Young et al. (2008) found that a vast majority of math items functioned similarly for non-ELL students and non-accommodated ELLs, even though ELL students tended to perform worse than non-ELLs on the test. These results may indicate true differences in math ability between ELLs and native English speakers rather than differential item performance.

Linguistic complexity as a source of DIF. For studies that found evidence of DIF in favor of non-ELLs on math tests, there is a great deal of evidence of linguistic complexity as the source of the DIF. In one type of study, researchers identified items that displayed DIF against ELLs and then described the language features in these items. One of these studies used IRT, logistic regression, and MH methods to identify DIF on math tests for Grades 4, 5, 7, and 8. The tests included between 42 and 60 items, and an average of three items per test displayed DIF against ELLs. The items flagged for DIF included more words, sentences, and general academic vocabulary than non-DIF items. These descriptive and lexical features were more important in explaining DIF than the grammatical features (Abedi et al., 2008).

In another study, 39 math items were qualitatively examined for features of linguistic complexity. Items flagged for DIF included several grammatical features of linguistic complexity, such as long noun phrases and multiple clauses. These items also tended to have polysemous words and non-math-related academic words that are typically learned outside the classroom. The author pointed out that the native Spanish-speaking students in this study were able to draw upon their knowledge of Spanish/English cognates to help them understand the items (Martiniello, 2008).

Another study examined statistical relationships between DIF statistics and linguistic complexity. Results indicated that individual aspects of linguistic complexity were correlated with DIF against ELLs. Significant correlations were found between uniform DIF statistics and total number of words and general academic vocabulary (Wolf & Leon, 2009). These results indicated that the magnitude of uniform DIF was related to linguistic complexity.

Using regression analysis, Martiniello (2009) examined composite linguistic complexity as a predictor of DIF against ELLs on a 39-item, Grade 4 math test. The composite linguistic complexity variable was composed of descriptive, grammatical, and non-math-related lexical elements. In this study, linguistic complexity in math items was related to the magnitude of the DIF statistics. Her findings suggest that non-math-related linguistic complexity does present a source of DIF in favor of non-ELLs. However, Martiniello's study did not examine the ability of individual linguistic complexity features (e.g., number of words, number of prepositional phrases) to predict DIF.

Lee and Randall (2011) identified several items on a statewide, 34-item math test that showed evidence of DIF. However, the math items did not seem to include high levels of language demand, and linguistic complexity features did not predict the magnitude of MH DIF statistics in the regression analysis. It is important to note, however, that item discrimination parameters were smaller for ELLs on items with particular descriptive (i.e., number of words) and lexical features.

Haag et al. (2013) also used multiple regression analyses to predict DIF from individual linguistic complexity features. In their examination of a Grade 3, 56-item math test in Germany, they found no instances of significant DIF. However, they did find that descriptive, lexical, and grammatical features were significant predictors of positive DIF statistics (i.e., DIF in favor of non-ELLs). Lexical and grammatical features explained about 28% of the variance in DIF. Some individual features had unique effects on DIF as well, including total number of words (descriptive) and the number of noun phrases (grammatical) per item. General academic vocabulary was only marginally significant in the prediction of DIF. Because the study included Grade 3 students, it is likely that the academic language was not as complex or varied as it would have been at higher grades. Additionally, this study examined math items written in German, a language that may include different lexical and grammatical structures than English.

Schematic Representations in Math Items

Another factor that can influence ELL students' performance on math tests is the presence of nonlinguistic elements in the items. Math items often include charts, graphs, tables, equations, and other visual elements. These features provide another way for ELL students to extract meaning from an item (Kopriva, 2000). Although linguistic complexity can impede ELL students' understandings of a math item, the presence of charts, tables, graphs, and other visual features in the item may help lessen this effect (Wolf & Leon, 2009).

Martiniello (2009) distinguished between "pictorial" and "schematic" representations. Pictorial representations include concrete visuals, such as pictures of physical objects. Schematic representations, on the other hand, are more helpful as students negotiate the meaning of an item (p. 170). Schematic representations may include visual features or mathematical formulas that help display relationships between variables. Martiniello found that, for ELL students, the impact of an item's linguistic complexity on the DIF statistic is moderated by the presence of these schematic representations in the item.

Visual or schematic features in math items can help attenuate the effects of linguistic complexity on ELL students' performance. The proportion of visual features in items can vary widely by grade, however. For example, Abedi et al. (2008) found that about 40% of the math items on a Grade 4 test included visuals, while only about 28% of the math items on a Grade 8 test included visuals.

ELL Group Heterogeneity

ELL students come from a wide range of language, cultural, and family backgrounds. Any examination of ELL students must consider the diversity within this group. If not properly accounted for, variables such as language and socioeconomic status can confound study results.

ELLs in the United States are a linguistically diverse group; they speak more than 325 languages. Spanish is the most commonly spoken language, and between 76% and 80% of ELLs are native Spanish speakers (Hopstock & Stephenson, 2003; NCELA, 2011). The impact of language features on ELLs' test performance differs as a function of their native languages. Words and phrases are translated very differently from language to language. Additionally, some languages, such as English and Spanish, share many cognates. Cognates are words that have a common origin and similar spelling, meaning, and pronunciation. For example, the English word *family* is very similar to the Spanish word *familia*. Cognates can be helpful for ELLs as they try to determine the meaning of English words (Gomez, 2010; Ramírez, Chen, & Pasquarella, 2013).

Although cognates are helpful for ELL students, false cognates can cause confusion. False cognates are words that appear in both English and a student's native language but have a different meaning in each language. As an example, the noun *pan* appears in both English and Spanish, but this word can either mean "a metal container used for frying food, washing, etc." in English, or "a loaf of bread" in Spanish. In the case of false cognates, a student's native language actually hinders his or her understanding of the word.

Cognates, false cognates, and other linguistic features can vary from language to language and have an impact on ELL students' test performance (Kopriva, 2000). In one study of the linguistic complexity of Grade 4 math test items, the researcher conducted cognitive labs with native Spanish-speaking students. The purpose of the cognitive labs was to gain insight from students about difficulties they had with item comprehension. The results indicated that the presence of high-frequency Spanish-English cognates (e.g., the words *impossible* and *combination*) facilitated students' comprehension of the math items. However, the students were less successful in extracting meaning from low-frequency Spanish-English cognates (Martiniello, 2008).

ELLs also vary with regard to their English proficiency and native language literacy. Both of these factors can influence ELL students' performance on math tests and other content assessments. For example, Abella et al. (2005) found that math test scores related to English language proficiency and native language literacy. Studies that have grouped ELLs on the basis of English proficiency have found performance differences between low-proficiency versus high-proficiency ELL students. Abedi et al. (2008) compared ELLs with non-ELLs and found several math items with DIF against the ELL group. However, the number of items displaying evidence of DIF was much higher within the ELL group. When the researchers compared ELLs with low English proficiency (focal group) to ELLs with high English proficiency (reference group), they found many more DIF items than in the ELL versus non-ELL comparison.

The variation of language proficiency within the ELL group may interfere in the identification of group differences between ELLs and non-ELLs. For example, Mahoney (2008) did not identify DIF against ELLs in math items, and she cited within-group variation in language proficiency as a possible explanation.

As a group, ELLs also differ systematically from non-ELLs in other background variables. ELL students are more likely to come from households in which parents have little formal education (Capps, 2005). Additionally, students with limited English proficiency are more than twice as likely to come from low-SES backgrounds than English proficient students

(Abedi, 2002; Capps, 2005). One study found that as many as 82% of ELLs were from a low SES group compared with 27% of non-ELLs (Martiniello, 2008).

Several linguistic and background factors are confounded with ELL status (Abedi, 2002), and these variables have an impact on ELL students' test performance. For example, Abedi and Lord (2001) found that the average math scores of low-SES students were significantly lower than for high-SES students. When possible, researchers should attempt to control for SES, language, and other factors when comparing ELL students with native English speakers. For situations in which a background variable cannot be controlled, researchers should interpret performance differences between ELLs and non-ELLs with caution.

Summary

The body of research on linguistic complexity and math tests confirms that math items often include high language demand. Math items frequently include descriptive, lexical, and grammatical language features. These linguistic complexity features pose a particular challenge for students who are still learning English. DIF studies indicate that ELL and non-ELL students do not have the same probability of success on math items, even after being matched on math ability, and several studies have implicated linguistic complexity as a source of DIF.

However, the relationship between DIF on math tests and individual linguistic complexity features is still unclear. Several limitations in the aforementioned studies should be addressed in future research. First, very few studies examined statistical relationships between DIF and individual linguistic complexity features. Only a few studies used multiple regression to predict DIF statistics from individual linguistic complexity features. These studies generally did not examine a combination of continuous and discrete predictors.

Additionally, the math items analyzed in some studies did not present the full range of linguistic complexity features. For example, Haag et al. (2013) examined items that included very few instances of passive verbs and compound sentences. One reason for this is the small number of math items included in each study. Across studies, the number of math items on each test ranged from 34 to 60. To better ensure that a study includes the broad range of language features, researchers should include larger samples of math items.

Furthermore, ELLs are a diverse group, and few studies controlled for more than one within-group factor. ELLs are linguistically diverse, and native language affects the way students comprehend text. SES has also shown to influence students' test performance. It is important to control for these factors within the ELL population to avoid the confounding of math ability with background factors.

Evidence of DIF against ELLs on math tests is widespread in the literature. What is less well known is the source of DIF, the importance of ELL background variables, and grade-level differences in math item linguistic complexity and DIF. There is some evidence that linguistic complexity is related to DIF against ELLs. However, the research is inconclusive with regard to the specific linguistic complexity features that lead to differential performance between ELLs and non-ELLs. The present study contributes to this area of research by identifying continuous and discrete linguistic complexity features in math items, assessing items for the presence of schematic representations, and using those features as predictors of DIF statistics in multiple regression analyses. This study also examines grade level, SES, and language background as variables in the analyses.

DIF Detection Using Logistic Regression

There are many types of DIF procedures available. The decision about which DIF analysis to use depends on the choice of a matching criterion, sample size in the reference and focal groups, and the desired DIF statistics. A more in-depth discussion about one particular DIF method, logistic regression, is included below.

Logistic regression. With the general logistic regression model, a binary variable is predicted from sets of continuous or discrete independent variables. The goal of logistic regression is to determine the log of odds of being in a particular group (Tabachnick, Fidell, & Osterlind, 2001) with the equation:

$$P(u=1)=\frac{e^{z}}{[1+e^{z}]} \qquad (1)$$

Swaminathan and Rogers (1990) adapted logistic regression as a DIF method (LR DIF). With LR DIF analyses, the probability of success on an item is regressed on total test score (the conditioning variable) and group membership. In the logistic regression equation, *z* is the linear regression model, which includes a slope and intercept, and can include an interaction term. In the logistic regression method of DIF detection described by Swaminathan and Rogers:

$$z=\tau_0+\tau_1\theta+\tau_2 g+\tau_3(\theta g)\,, \qquad (2)$$

in which theta (θ) is examinee ability (i.e., total test score) and is a continuous variable. Group membership is a discrete variable represented by *g*, and θg represents the interaction of ability with group membership. Parameter weights are represented by τ_1 (for ability), τ_2 (for group), and τ_3 (for the interaction between ability and group).

LR DIF provides a relatively straightforward test of significance. Total test score is expected to predict the probability of success on the item. Therefore, τ_1 is always expected to be significant. However, if group membership is a significant predictor of success on the item, this provides evidence of uniform DIF (i.e., $\tau_2 \neq 0$ and $\tau_3 = 0$). Where the reference group is coded "1" and the focal group is coded "0," positive values of τ_2 indicate that the item favors the reference group, while negative values indicate an item that favors the focal group.

The significance of a LR DIF procedure is determined by examining model fit using a chi-square test. The test for uniform DIF is chi-square distributed with one degree of freedom and compares the initial model (which includes only ability) to the model with the grouping variable. The test for non-uniform DIF is chi-squared distributed with two degrees of freedom and compares the second model (which includes the grouping variable) to the third model, which has the interaction between group and ability.

The ability to detect non-uniform DIF gives LR DIF procedures a significant advantage over other non-parametric methods. The significance of the interaction term in the prediction of probability of success on the item indicates non-uniform DIF (i.e., $\tau_3 \neq 0$). A positive value of τ_3 indicates that the item favors the reference group at higher ability levels and favors the focal group at lower ability levels. A negative τ_3 indicates that the item favors the focal group at higher ability levels and favors the reference group at lower ability levels (Jodoin & Gierl, 2001).

In a simulation study involving multiple sample sizes and test lengths, logistic regression was as effective as the Mantel-Haenszel (MH) method in the detection of uniform DIF. However, the same study revealed that the MH method was unable to detect any of the items with non-uniform DIF, while the logistic regression DIF procedure was quite successful, especially with larger sample sizes and longer tests (Swaminathan & Rogers, 1990). There is further support for logistic regression as a better method than MH in the detection of non-uniform DIF (Clauser, Nungester, Mazor, & Ripkey, 1996; Hidalgo & LÓPez-Pina, 2004).

Effect size. Another advantage of LR DIF is that it provides a straightforward estimate of effect size. Effect size is the degree to which group membership influences the probability of success on the item (De Ayala, 2009). Since the significance test in logistic regression is influenced by sample size, it is important to include an effect size measure. Very large sample sizes, for example, could yield significant results when the amount of DIF is actually quite small.

Jodoin and Gierl (2001) point out that, while the coefficients τ_2 and τ_3 could be used as effect sizes, these values are dependent on the way reference and focal groups are coded. A more easily interpretable measure of effect size for dichotomously scored items is the Nagelkerke pseudo R^2 (Zumbo, 1999; Zumbo & Thomas, 1997). It is important to note that the Nagelkerke pseudo R^2 is not interpreted the same as R^2 in ordinary least squares regression (OLS) analyses. In OLS regression, R^2 is the variance explained in the dependent variable by the combination of independent variables. The Nagelkerke pseudo R^2 cannot be interpreted as variance explained, but can be used in combination with other information (e.g., standardized regression coefficients, significance of the model) to interpret logistic regression outcomes (Peng, Lee, & Ingersoll, 2002). Effect size can be assessed by examining the change in R^2 (ΔR^2) from one step in the LR DIF analysis to the next. Zumbo suggests that Nagelkerke pseudo R^2 effect sizes be calculated for uniform DIF as well as a “simultaneous test of uniform and non-uniform DIF” (p. 27).

Zumbo and Thomas (1997) initially provided guidelines on what constitutes negligible DIF ($\Delta R^2 < 0.13$), moderate DIF ($0.13 \leq \Delta R^2 \leq 0.26$), and large DIF ($\Delta R^2 > 0.26$). Alternate guidelines for ΔR^2 were proposed by Jodoin and Gierl (2001). They classified negligible DIF ($\Delta R^2 < 0.035$), moderate DIF ($0.035 \leq \Delta R^2 \leq 0.070$), and large DIF($\Delta R^2 > 0.070$). Using their classification system, they found that many more of the moderate DIF items were actually identified. In a simulation study that compared LR DIF and MH procedures, Hidalgo and LÓPez-Pina (2004) also found better DIF detection rates using the criteria proposed by Jodoin and Gierl.

Overall, logistic regression has many advantages as a DIF detection procedure. LR DIF procedures have proven to be successful at identifying uniform as well as non-uniform DIF. LR DIF is also a relatively straightforward procedure and allows for tests of significance as well as calculation of effect size. As an effect size measure, the Nagelkerke R^2 is easily calculated in statistical packages such as SPSS. Logistic regression also has smaller sample size requirements than other DIF detection methods (e.g., parametric methods). For dichotomously scored items, about 200 examinees are required per group for LR DIF analyses (Zumbo, 1999). Given these advantages, the present study will employ the use of LR DIF detection.

Chapter III: Methods

The following section describes the research questions, instrument, sample, linguistic complexity coding protocol, and the statistical methods used to evaluate the data.

Research Questions

The present study aims to build on previous research in the areas of linguistic complexity in math items and DIF between ELLs and non-ELLs. The data and methods were chosen to answer the following research questions:

1. Do math items show evidence of differential item functioning (DIF) favoring non-ELLs over native Spanish-speaking ELLs, after controlling for socioeconomic status (SES)?
2. Do individual linguistic complexity features predict the DIF statistics?
3. Does the presence of schematic representations in math items attenuate the DIF statistics?
4. Are there differences between Grade 4 and Grade 8 on Research Questions 1–3?

Instrument

The study includes 2010 data from a standardized math test administered in a Midwestern state. To address the fourth research question (comparing data from two different grades), one set of data was examined for Grade 4, and a second set of data was examined for Grade 8. Math items at the higher grade level are expected to include more complex lexical and grammatical features (e.g., more non-math-related academic words, complex sentence structures) than math items at the lower grade level.

For each grade, there were two test forms. For Grade 4, Form 4-1 included 59 items and Form 4-2 included 50 items for a total of 109 items. For Grade 8, Form 8-1 included 61 items and Form 8-2 included 58 items for a total of 119 items. Math items for the Grade 4 and Grade 8 assessments were multiple-choice with four answer choices on each item. The tests included math items from four major content areas: numbers and computation, data, geometry, and algebra. Table 1 shows item specifications for Grade 4 and Grade 8 test forms.

Table 1

Item Specifications for Grade 4 and Grade 8 Math Tests

		Items at each benchmark (%)	
Standard	Benchmark	Grade 4	Grade 8
Numbers and computation	Number sense	0	7.1
	Number systems & their properties	12.3	11.8
	Computation	20.5	12.9
Data	Statistics	15.1	7.1
	Probability	0	12.9
Geometry	Geometric figures and their properties	5.5	11.8
	Geometry from an algebraic perspective	5.5	9.4
	Measurement and estimation	16.5	0
	Transformational geometry	5.5	0
Algebra	Functions	12.3	8.2
	Variables, equations, & inequalities	6.8	14.1
	Models	0	4.7

All items were dichotomously scored and scaled using classical test theory (CTT). Math items could include text as well as schematic representations such as charts, graphs, tables, and equations. Both text and visual representations could be included in the item stem, item foils, or both the item stem and foils.

Sample

To improve the interpretability of results, several groups of students were removed from the sample. First, disability status can affect the interpretability of results, so students with disabilities (SWD) were removed from the sample. Secondly, the present study involves a comparison of students who are still learning English (i.e., ELLs) versus students who are not ELLs. Some students in the dataset were categorized as “monitored” ELLs. These students benefitted from ESOL programs in the past, but they had twice passed the statewide English language proficiency test. Because students in the monitored category do not meet the definition of ELL, these students were removed from the analyses. Thirdly, a student’s native language can influence the way he or she understands language features. For example, Spanish-speaking students may recognize many cognates in English that Chinese-speaking students do not recognize. To reduce the within-group ELL variability and thus make the study results more interpretable, the present study included only ELLs from Spanish-speaking backgrounds. Native Spanish-speaking ELLs were chosen because they represent the largest language group in the United States. In the United States about 80% of ELL students speak Spanish as their native language (U.S. Department of Education, 2007-08). This statistic is reflected in the math assessment data used in this study. In the Grade 4 sample, about 83% of the ELLs were native Spanish-speakers, and in the Grade 8 sample, about 82% of the ELLs were native Spanish-speakers.

The datasets also included information about students who required free or reduced lunch. This variable was used as a proxy for SES. The SES variable included three levels: 1) student receives free lunch, 2) student receives reduced lunch, and 3) student does not receive free or reduced lunch.

The original and reduced samples are discussed in more detail below. This information is organized by grade and test form. This study included data from two test forms of the math assessment for each grade; the forms are identified by the grade level and form number (e.g., Form 4-1 for Grade 4, Form 1).

Grade 4. Form 4-1 initially included 21,368 non-ELLs and 1,973 ELLs for a total sample of 23,341 students. After SWDs, monitored ELLs, and non-Spanish-speaking ELLs were removed, 20,080 non-ELL students and 1,506 ELL students remained in the sample.

On Form 4-2, there were initially 7,797 non-ELLs and 680 ELLs for a total sample of 8,477 students. After SWDs, monitored ELLs, and non-Spanish-speaking ELLs were removed, 7,312 non-ELL students and 540 ELL students remained in the sample.

Grade 8. Form 8-1 initially included 15,249 non-ELLs and 856 ELLs for a total sample of 16,105 students. After SWDs, monitored ELLs, and non-Spanish-speaking ELLs were removed, 14,535 non-ELL students and 622 ELL students remained in the sample.

On Form 8-2, there were initially 15,266 non-ELLs and 829 ELLs for a total sample of 16,095 students. After SWDs, monitored ELLs, and non-Spanish-speaking ELLs were removed, 14,562 non-ELL students and 618 ELL students remained in the sample.

Item Coding

Linguistic complexity features. A preliminary review of items revealed that nearly every item on the 2010 Grade 8 math assessment contained at least one feature of linguistic complexity. Many items on the 2010 Grade 4 math assessment contained at least one feature of linguistic complexity as well. Test items often included more than one feature of linguistic complexity and features from more than one linguistic complexity group (i.e., descriptive, lexical, grammatical).

Previous research reveals dozens of linguistic complexity features that can appear in math items. Some features identified in the research rarely occurred in the math items for this study, so these features were not included in the linguistic complexity coding protocol. While the linguistic complexity coding scheme developed for this study does not contain an exhaustive list of features, the coding protocol was based on the preliminary analysis of items as well as previous research regarding language features in math items. The protocol also built on coding schemes from studies that examined linguistic complexity in math items (Haag et al., 2013; Martiniello, 2009).

The initial item coding protocol included 12 features of linguistic complexity that were categorized into three groups: descriptive features (*number of words* and *number of sentences*), grammatical features (*noun phrases, sentence type, prepositional phrases, dependent adjective clauses, modals,* and *passive verb tense*), and lexical features (*general academic words, multiple-meaning words, colloquial words/phrases,* and *Spanish-English cognates*).

Item coding pilot test. The item coding protocol was pilot tested with three experts using the Grade 4 and Grade 8 math items. The first two experts have backgrounds in both English and mathematics and extensive experience editing and reviewing test items. The third expert has more than five years of experience teaching English as a second language (ESL) and developing English proficiency assessments.

The item coding protocol was edited based on the results from the item coding pilot test. The item feature *noun phrases* was omitted from the protocol because of its extensive overlap with other item features (e.g., prepositional phrases). The item feature *colloquial phrases* was also omitted because there were very few instances of this feature. Other item features, such as *adjective clauses* and *number of sentences* were clarified as a result of the item coding pilot test.

The final item coding protocol, which includes 10 features of linguistic complexity, is included in Appendix A.

Three item coders were trained to use the item coding protocol. The item coders were experts in English language arts (ELA), linguistics, and math. The first item coder was a doctoral student in research, evaluation, measurement, and statistics who also holds a master's degree in English language studies and has extensive experience coding texts. The second item coder was a bilingual (English/Spanish) doctoral student in psychology with several years' experience teaching English in Chile and Spain. The third item coder was a high school math teacher with a strong background in English grammar. All item coders had experience developing and/or editing test items for large-scale assessments.

The item coder who worked as a high school math teacher identified words in the items that were related to math content. These math content words were removed from other coders' counts of *general academic words*. Item coders counted the number of occurrences for each linguistic complexity feature. Item coders also recorded whether an item was primarily schematic or primarily text-based. When item coders had completed approximately 20 items, the researcher compared the three coders' codes for any discrepancies. Major discrepancies were identified on *adjective clauses* and *passive verb tense*. The researcher met with all three item coders to clarify the operationalization of these item features. After the clarification, the item coders displayed higher agreement on these variables.

When all item coding was complete, the researcher noted any discrepancies among item coders. To complete the item coding process, all three coders met with the researcher to reach consensus about discrepancies in their codes. During the meeting, the only item feature on which item coders did not always reach consensus was *general academic words*.

For instances in which there was not 100% agreement on general academic words, the researcher developed additional criteria to designate general academic words: 1) at least two of the three item coders identified the word as being a general academic word for that grade level, and 2) the word was at or above the particular grade level according to the EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies (Taylor et al., 1989) or the Berkeley Unified School District (BUSD) Grade Level Academic Vocabulary Manual (Tugwell, 2013), or 3) if the grade level for a particular word could not be determined using the sources in the second criterion, the word had to include at least three syllables or be derived from a root word. Examples of general academic words for Grade 4 included *amusement* and *marigolds.* At Grade 8, examples of general academic words included *capacity*, *manufacture*, and *merchandise*.

Schematic representations. The effects of linguistic complexity in math items on ELL students can be attenuated by schematic representations in the items (Martiniello, 2009). The present study adopts Martiniello's definition of schematic representations, as well as her coding scheme for determining if an item is primarily text-based or primarily schematic. Items were coded for schematic representations on a dichotomous scale with "0" indicating that the item was primarily text-based and "1" indicating that the item was primarily schematic. Items considered primarily text-based required students to rely, at least partially, on language to successfully answer the item. Items considered primarily schematic could be answered by using information in a chart, graph, table, equation, collection of symbols, etc.

DIF Methods

To answer the first research question, DIF analyses were conducted for the Grade 4 and Grade 8 math items. The focal group consisted of native Spanish-speaking ELLs, and the reference group consisted of non-ELLs.

A preliminary analysis of math items for Grade 8 was conducted prior to this study. This preliminary analysis involved 86 math items from the 2012 administration of the same statewide assessment used in the present study. Results showed that several items demonstrated uniform as well as non-uniform DIF with ELLs as the focal group. Because the 2010 math items are expected to be psychometrically similar to the 2012 math items, the DIF methods selected for the present study allow for the detection of both uniform and non-uniform DIF.

The datasets for the present study were initially calibrated using CTT. Total scores were calculated as the matching criterion. DIF analyses were conducted using the logistic regression DIF (LR DIF) procedure. LR DIF allows for the detection of both uniform and non-uniform DIF through a stepwise procedure. The lunch status variable was included in the study as a proxy for SES. Lunch status (i.e., no lunch support, reduced lunch, free lunch) was entered in the first block of the analysis to control for students' SES. Math total score was entered in the second block, group membership in the third block, and the interaction of total score and group in the final block. Effect sizes were determined by examining Nagelkerke pseudo R^2.

Multiple Regression Analyses

The relationship of linguistic complexity features to the LR DIF statistics was analyzed by generating scatterplots and a correlation matrix. To answer the second and third research questions, the DIF statistics were regressed on various linguistic complexity features and the schematic representation variable using the stepwise regression procedure.

Grade Level Differences

The fourth research question addresses grade-level differences in the above analyses. The number of items demonstrating DIF and the degree of DIF in those items was compared for

Grade 4 and Grade 8. Additionally, the multiple regression analyses were compared to see which linguistic complexity features were significant predictors of DIF for each grade.

Chapter IV: Results

The following section describes sample demographics and test scores, item features, results from the LR DIF procedures, and results from the multiple regression analyses associating DIF effect size with item features.

Sample Demographics and Test Scores

At Grades 4 and 8, there was a balance of male and female students. The majority of non-ELLs (between 76.7% and 77.3%) were White. In the ELL group, only native Spanish-speakers were retained for this study. At Grades 4 and 8, each ELL sample was at least 96% Hispanic (Tables 2 and 3).

Table 2

Demographics for Grade 4 (Percent)

	Form 4-1			Form 4-2		
	Total	ELL	Non-ELL	Total	ELL	Non-ELL
Gender						
Male	49.3	50.3	49.2	48.4	48.9	48.3
Female	50.7	49.7	50.8	51.6	51.1	51.7
Race						
American Indian	1.2	0.7	1.2	1.3	0.6	1.3
Asian	1.7	0.2	1.8	1.7	0.2	1.8
Black	6.6	0.0	7.0	6.5	0.0	7.0
Hispanic	14.7	97.0	8.6	14.5	98.0	8.3
White	71.8	2.0	77.0	72.0	1.2	77.3
Missing or multi-ethnic	4.0	0.1	4.4	4.0	0.0	4.3

Table 3

Demographics for Grade 8 (Percent)

	Form 8-1			Form 8-2		
	Total	ELL	Non-ELL	Total	ELL	Non-ELL
Gender						
Male	49.0	49.8	48.9	49.1	50.0	49.1
Female	51.0	50.2	51.1	50.9	50.0	50.9
Race						
American Indian	1.2	0.3	1.2	1.2	0.3	1.2
Asian	2.0	0.3	2.1	2.1	0.5	2.1
Black	6.4	0.0	6.7	6.6	0.0	6.9
Hispanic	12.9	96.1	9.4	12.7	97.2	9.1
White	73.8	3.1	76.7	73.7	1.8	76.7
Missing or multi-ethnic	3.7	0.2	3.9	3.7	0.2	4.0

Tables 4 and 5 display total test score statistics by group, and the number of students in each lunch support category (i.e., the proxy for SES) is indicated. The highest SES group is indicated as “no lunch support,” the next highest is “reduced lunch,” and the lowest SES group consists of students who receive “free lunch.”

ELL students were much more likely to be in the lower SES groups than non-ELL students. Across the two test forms for Grade 4, about 93.3% of the ELL students were in the lower-SES groups (i.e., received either reduced lunch or free lunch) compared with 38.4% in the non-ELL group. Results were similar for Grade 8, in which 93.9% of ELL students were in the lower-SES groups compared with 34.5% in the non-ELL group. These results are consistent with other studies that found ELL students were more likely to come from low-SES backgrounds (Abedi, 2002; Capps, 2005; Martiniello, 2008).

Table 4

Math Total Scores for Grade 4 by Group and Lunch Status

	Form 4-1					Form 4-2			
Group	Min	Max	*M*	*SD*	Group	Min	Max	*M*	*SD*
All students (*N* = 21,586)	6	59	48.67	7.45	**All students** (*N* = 7,852)	11	50	41.04	6.21
No lunch support (*N* = 12,499)	6	59	50.38	6.43	No lunch support (*N* = 4,504)	18	50	42.62	5.30
Reduced lunch (*N* = 2,096)	15	59	48.06	7.36	Reduced lunch (*N* = 758)	16	50	40.36	6.17
Free lunch (*N* = 6,991)	11	59	45.79	8.23	Free lunch (*N* =2,590)	11	50	38.50	6.77
ELLs (*N* = 1,506)	11	59	44.88	8.49	**ELLs** (*N* = 540)	13	49	38.05	6.87
No lunch support (*N* = 97)	25	58	46.88	7.28	No lunch support (*N* = 40)	29	49	40.12	5.64
Reduced lunch (*N* = 152)	15	58	45.95	8.22	Reduced lunch (*N* = 45)	22	49	39.40	6.98
Free lunch (*N* = 1,257)	11	59	44.60	8.58	Free lunch (*N* = 455)	13	49	37.74	6.92
Non-ELLs (*N* = 20,080)	6	59	48.95	7.29	**Non-ELLs** (*N* = 7,312)	11	50	41.26	6.09
No lunch support (*N* = 12,402)	6	59	50.40	6.42	No lunch support (*N* = 4,464)	18	50	42.64	5.29
Reduced lunch (*N* = 1,944)	15	59	48.22	7.26	Reduced lunch (*N* = 713)	16	50	40.42	6.12
Free lunch (*N* = 5,734)	11	59	46.05	8.13	Free lunch (*N* = 2,135)	11	50	38.66	6.72

Across test forms and grades, lower-SES students performed less well on the math tests, ELL students had lower test scores than non-ELL students, and the group with the lowest math test scores was comprised of low-SES ELL students.

Table 5

Math Total Scores for Grade 8 by Group and Lunch Status

Form 8-1					Form 8-2				
Group	Min	Max	*M*	*SD*	Group	Min	Max	*M*	*SD*
All students (*N* = 15,157)	2	61	46.08	10.38	**All students** (*N* = 15,180)	2	58	43.21	9.64
No lunch support (*N* = 9,538)	11	61	48.59	9.13	No lunch support (*N* = 9,580)	7	58	45.59	8.51
Reduced lunch (*N* = 1,421)	12	61	44.55	10.08	Reduced lunch (*N* = 1,378)	12	58	41.34	9.49
Free lunch (*N* = 4,197)	2	61	40.88	11.08	Free lunch (*N* = 4,222)	2	58	38.41	10.19
ELLs (*N* = 622)	9	60	36.99	10.92	**ELLs** (*N* = 618)	8	57	34.67	10.45
No lunch support (*N* = 38)	13	59	39.24	10.94	No lunch support (*N* = 37)	13	56	34.92	10.04
Reduced lunch (*N* = 67)	13	57	38.88	10.86	Reduced lunch (*N* = 46)	18	54	37.02	9.47
Free lunch (*N* = 517)	9	60	36.57	10.90	Free lunch (*N* = 535)	8	57	34.45	10.09
Non-ELLs (*N* = 14,535)	2	61	46.47	10.17	**Non-ELLs** (*N* = 14,562)	2	58	43.57	9.46
No lunch support (*N* = 9,501)	11	61	48.63	9.10	No lunch support (*N* = 9,543)	7	58	45.63	8.48
Reduced lunch (*N* = 1,354)	12	61	44.83	9.96	Reduced lunch (*N* = 1,332)	12	58	41.49	9.45
Free lunch (*N* = 3,680)	2	61	41.48	10.97	Free lunch (*N* = 3,687)	2	58	38.99	10.08

Differential Item Functioning Analyses

To answer the first research question (*Do math items show evidence of DIF favoring non-ELLs over native Spanish-speaking ELLs, after controlling for SES?*), LR DIF analyses were conducted on all Grade 4 and Grade 8 math items. Full results from the LR DIF analyses for Grades 4 and 8 are included in Appendices B and C, respectively.

In Step 1 of each LR DIF analysis, the lunch (i.e., SES proxy) and math total score variables were entered as predictors of each dichotomous item score. In Step 2 the group variable (i.e., ELL or non-ELL) was added as a predictor to test for uniform DIF. In Step 3, the interaction between group and score was added as a predictor of each item score to test for non-uniform DIF. After controlling for lunch status and math total score, several items were flagged as having significant uniform DIF and/or significant non-uniform DIF (Table 6).

Table 6

Items Showing Significant (p < .05) DIF across Grades and Test Forms

			Uniform DIF			Non-uniform DIF	
Grade	Test form	Total items	Items favoring Non-ELLs	Items favoring ELLs	Total (%)	Total (%)	Total items that showed uniform DIF, non-uniform DIF, or both (%)
4	4-1	59	14	16	30 (51%)	29 (49%)	43 (72.9%)
	4-2	50	8	7	15 (30%)	11 (22%)	20 (40.0%)
8	8-1	61	8	11	19 (31%)	30 (49%)	40 (65.6%)
	8-2	58	7	8	15 (26%)	26 (45%)	32 (55.2%)

Grade 4. Results for Grade 4 show that 63 out of 109 items showed uniform DIF, non-uniform DIF, or both uniform and non-uniform DIF. After controlling for the SES proxy variable (i.e., lunch status) and the matching criterion (i.e., total test score) the group variable significantly impacted the probability of success on 45 of the 109 total items at the .05 level. About the same number of items favored non-ELLs (22 items) as ELLs (23 items). Additionally,

40 of the 109 total items displayed non-uniform DIF. On these items, the interaction between the group variable and total score was a significant predictor of item score at the .05 level. Total DIF (Appendices B and C) was calculated by adding the change in the Nagelkerke R^2 (ΔR^2) statistic from Step 1 to Step 2 to the ΔR^2 from Step 2 to Step 3. For all items, the magnitude of uniform, non-uniform, and total DIF was low.

The primary purpose of this study is to predict uniform DIF, and this will be the focus of the following discussion. Uniform DIF statistics, as shown in the stem-and-leaf plots in Figures 1 and 2, were calculated by taking the difference in the Nagelkerke pseudo-R^2 value at the second block (total score) and the third block (group variable). To aid interpretation of the uniform DIF, statistics favoring non-ELLs are positive, and statistics favoring ELLs were made negative. For Grade 4, the uniform DIF statistics ranged from -0.0027 to 0.0115 (M = 0.0001, SD = 0.0015). The plot in Figure 1 displays the distribution of uniform DIF for Grade 4.

Frequency	Sign	Stem	Leaf
4 outliers: -.0021, -.0023, -.0026, -.0027			
1	-	.001	6
2	-	.001	4 4
3	-	.001	2 3 3
1	-	.001	1
6	-	.000	8 8 9 9 9 9
5	-	.000	6 6 6 6 7
4	-	.000	4 4 5 5
14	-	.000	2 2 2 2 2 2 2 3 3 3 3 3 3 3
9	-	.000	1 1 1 1 1 1 1 1 1
29		.000	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4	+	.000	2 2 2 3
7	+	.000	4 4 4 4 4 5 5
3	+	.000	6 6 7
1	+	.000	9
5	+	.001	0 0 1 1 1
2	+	.001	2 3
1	+	.001	5
2	+	.001	6 7
1	+	.001	9
5 outliers: .0024, .0029, .0034, .0046, .0115			

Figure 1. Stem and leaf plot of uniform DIF statistics for Grade 4. Positive values favor non-ELLs.

Grade 8. Results from Grade 8 also revealed several items with uniform DIF, non-uniform DIF, or both. For Grade 8, 72 out of the 119 items displayed one or both types of DIF. After controlling for lunch status and total test score, 34 of the 119 items displayed uniform DIF. Of these items, 15 favored the reference group (non-ELLs), and 19 items favored the focal group (ELLs). Furthermore, 56 of the 119 items displayed non-uniform DIF. On these items, the interaction between the group variable and total score was significant at the .05 level.

As with Grade 4, the magnitude of uniform, non-uniform, and total DIF for Grade 8 items was low. Uniform DIF statistics ranged from -0.0035 to 0.0018 (*M* = -0.0001, *SD* = 0.0007). The stem-and-leaf plot in Figure 2 displays the distribution of uniform DIF for Grade 8.

Frequency	Sign	Stem	Leaf

14 outliers: -.0008, -.0008, -.0008, -.0008, -.0010, -.0012, -.0012, -.0013, -.0013, -.0014, -.0015, -.0018, -.0024, -.0035

2	-	.000	6 6
2	-	.000	5 5
3	-	.000	4 4 4
5	-	.000	3 3 3 3 3
10	-	.000	2 2 2 2 2 2 2 2 2 2
15	-	.000	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
33		.000	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
12	+	.000	1 1 1 1 1 1 1 1 1 1 1 1
4	+	.000	2 2 2 2
7	+	.000	3 3 3 3 3 3 3
1	+	.000	4
2	+	.000	5 5

9 outliers: .0006, .0007, .0009, .0010, .0012, .0013, .0013, .0017, .0018

Figure 2. Stem and leaf plot of uniform DIF statistics for Grade 8. Positive values favor non-ELLs.

Item Features

Tables 7 and 8 show descriptive statistics for the item features and uniform DIF effect sizes for Grade 4 and Grade 8, respectively. Grade 8 items were generally more linguistically complex than Grade 4 items. On average, Grade 8 items had more words and sentences, more compound/complex sentences, more words in prepositional phrases, adjective clauses, modals, and more instances of passive verb tense than Grade 4 items. On average, Grade 4 items had slightly more general academic words and non-math-related, multi-meaning words. On average, Grade 8 test items had more Spanish-English cognates (non-math-related). Grade 4 and Grade 8 test forms had about the same amount of primarily schematic items (33% and 34%, respectively).

Table 7

Descriptive Statistics for Item Features and DIF Statistics for Grade 4 (N_{items} = 109)

Variable	Min	Max	*M*	*SD*
Descriptive features				
1. Number of words	5	86	26.30	17.03
2. Number of sentences	1	9	2.37	1.75
Grammatical features				
3. Sentence type*	0	1	0.51	0.50
4. Number of words in prepositional phrases	0	30	8.18	6.77
5. Adjective clauses	0	4	0.16	0.51
6. Modals	0	4	0.23	0.58
7. Passive verb tense**	0	1	0.42	0.49
Lexical features				
8. General academic words (not math-related)	0	7	0.75	1.13
9. Multi-meaning words (not math-related)	0	5	1.14	1.18
10. Spanish-English cognates (not math-related)	0	2	0.33	0.61
Schematic features				
11. Schematic representations***	0	1	0.33	0.47
Uniform DIF				
12. DIF effect size (ΔR^2)	-.0027	.0115	.0001	.0015

*Sentence type: 0 = simple sentences, 1 = some compound, complex, or compound/complex sentences

**Passive verb tense: 0 = no passive verb tense, 1 = item includes some passive verb tense

***Schematic representations: 0 = primarily text/pictorial, 1 = primarily schematic

Table 8

Descriptive Statistics for Item Features and DIF Statistics for Grade 8 (N_{items} = 119)

Variable	Min	Max	*M*	*SD*
Descriptive features				
1. Number of words	5	82	30.49	15.72
2. Number of sentences	1	9	2.65	1.39
Grammatical features				
3. Sentence type*	0	1	0.61	0.49
4. Number of words in prepositional phrases	0	48	11.86	9.00
5. Adjective clauses	0	5	0.24	0.66
6. Modals	0	4	0.34	0.70
7. Passive verb tense**	0	1	0.57	0.49
Lexical features				
8. General academic words (not math-related)	0	5	0.61	1.00
9. Multi-meaning words (not math-related)	0	4	1.00	1.15
10. Spanish-English cognates (not math-related)	0	5	0.57	0.89
Schematic features				
11. Schematic representations***	0	1	0.34	0.47
Uniform DIF				
12. DIF effect size (ΔR^2)	-.0035	.0018	-.0001	.0007

*Sentence type: 0 = simple sentences, 1 = some compound, complex, or compound/complex sentences
**Passive verb tense: 0 = no passive verb tense, 1 = item includes some passive verb tense
***Schematic representations: 0 = primarily text/pictorial, 1 = primarily schematic

Correlations among item features. Correlations among the item characteristics and DIF statistics are shown in Table 9 (for Grade 4) and Table 10 (for Grade 8). At both grades, the descriptive features were highly correlated with several grammatical and lexical features. This indicates that items with more words and sentences also tended to have more complex grammatical structures and complex words. However, for Grade 4, one grammatical feature, *passive verb tense*, was not significantly associated with item length in words or sentences. At both Grades 4 and 8, *passive verb tense* was significantly correlated with fewer linguistic complexity variables than any other item feature.

As expected, most language features at Grades 4 and 8 were negatively associated with the feature *schematic representations.* This indicates that primarily schematic items tended to have fewer linguistic complexity features than non-schematic items. However, at Grade 4, *passive verb tense* showed a significant positive relationship to *schematic features*. Schematic items at Grade 4 tended to include the phrase "is shown below" in the item stem (e.g., "A coordinate plane *is shown below*."). The item coders agreed that this phrase is an example of passive verb tense; they coded items with this phrase accordingly.

Item features associated with uniform DIF. On the Grade 4 math tests, only one grammatical feature, *adjective clauses*, was significantly correlated with uniform DIF (r = .53, p < .01). The positive correlation indicates that items with more adjective clauses had higher DIF against ELLs. Item 4.2.29 had the most adjective clauses (one in each of the four answer choices), and it also had the largest uniform DIF against ELLs on the Grade 4 test (ΔR^2 = 0.0115).

For Grade 8 four features, two grammatical features and two lexical features, were significantly related to uniform DIF: *number of words in prepositional phrases* (*r = -.18, p < .05*), *adjective clauses* (r = -.31, p < .01), *multi-meaning words not related to math* (*r = -.25, p < .01*)*, and Spanish-English cognates not related to math* (r = -.20, p < .01). The directions of the relationships are all negative, indicating that items with more of these features tended to have lower uniform DIF statistics (i.e., DIF *in favor of* ELLs). Items with more Spanish-English cognates (i.e., words that are similar in Spanish and English) tended to favor the native Spanish-speaking students over non-ELLs. This suggests that the ELL students were able to draw on their native language to better comprehend the math items. However, the results for prepositional phrases, adjective clauses, and multi-meaning words contradict the expectation that higher linguistic complexity in the items would be related to larger DIF against ELLs. The results suggest that items with more of these features tended to actually favor ELL students.

Multiple Regression Analyses

To answer the second and third research questions, the uniform DIF effect size (ΔR^2) was regressed on individual item characteristics to examine the relationship between DIF effect size and the item characteristics. The exploratory stepwise regression procedure was used to regress the amount of uniform DIF on each of the 11 item features.

Grade 4. An investigation of the item-level features showed that two Grade 4 items were multivariate outliers. With 11 item feature variables entered in the prediction of DIF, Item 4.2.29 and Item 4.2.18 had Mahalanobis distance values that exceeded the critical value of 31.62 ($p <$ .001). Item 4.2.29, for example, had more adjective clauses than any other item. A different adjective clause appeared in each of the four answer choices for this item. However, this item had an average number of words and sentences. This item also had the largest uniform DIF against ELLs of all the Grade 4 math items ($\Delta R^2 = 0.0115$). Multiple regression results are reported with these two items included (for a total of 109 items), and again with these two items omitted (for a total of 107 items).

Table 9

Correlations of the Item Features and Uniform DIF Statistics for Grade 4 (N_{items} = 109)

Variable	1	2	3	4	5	6	7	8	9	10	11	12[1]
1. Number of words	---	.87**	.41**	.79**	.24*	.27**	.14	.63**	.40**	.32**	-.39**	-.06
2. Number of sentences		---	.30**	.64**	.04	.26**	.09	.64**	.19*	.23*	-.27**	-.11
3. Sentence type			---	.39**	.29**	.17	.39**	.24*	.22*	.23*	-.06	.00
4. Words in prepositional phrases				---	.16	.14	.08	.48**	.44**	.29**	-.26**	-.10
5. Adjective clauses					---	.16	-.04	.12	.19*	.25**	-.18	.53**
6. Modals						---	.09	.06	.04	-.07	-.05	.11
7. Passive verb tense							---	-.03	-.07	-.07	.27**	-.13
8. General academic words								---	.32**	.34**	-.33**	-.12
9. Multi-meaning words									---	.23*	-.35**	-.13
10. Spanish-English cognates										---	-.22*	-.04
11. Schematic representations											---	-.14
12. Uniform DIF (ΔR^2)												---

[1] Positive values indicate that items with more of the language feature tend to display DIF against ELLs.
*Significant at the .05 level.
**Significant at the .01 level.

Table 10

Correlations of the Item Features and DIF Statistics for Grade 8 (N_{items} = 119)

Variable	1	2	3	4	5	6	7	8	9	10	11	12[1]
1. Number of words	---	.73**	.45**	.79**	.34**	.32**	.24**	.44**	.48**	.39**	-.34	-.08
2. Number of sentences		---	.29**	.43**	.05	.30**	.07	.28**	.34**	.16	-.31**	.01
3. Sentence type			---	.21*	.07	.28**	.08	.23*	.35**	.20*	-.20*	.09
4. Words in prepositional phrases				---	.27**	-.01	.20*	.36**	.35**	.42**	-.12	-.18*
5. Adjective clauses					---	.18	-.03	.04	.34**	.33**	-.04	-.31**
6. Modals						---	.08	-.01	.30**	.08	-.11	.10
7. Passive verb tense							---	.14	.13	.02	.04	.07
8. General academic words								---	.19*	.44**	-.08	-.03
9. Multi-meaning words									---	.32**	-.08	-.25**
10. Spanish-English cognates										---	.10	-.20**
11. Schematic representations											---	-.16
12. Uniform DIF (ΔR^2)												---

[1] Positive values indicate that items with more of the language feature tend to display DIF against ELLs.
*Significant at the .05 level.
**Significant at the .01 level.

With all 109 Grade 4 items included, the stepwise regression procedure identified two item features that significantly predicted DIF. Table 11 displays the unstandardized (B) and standardized (*β*) regression coefficients, R^2, adjusted R^2, and the squared semi-partial correlations after these two variables were entered.

Table 11

Grade 4 Regression Analysis for Item Features Predicting Uniform DIF[1]

(N_{items} = 109)

Predictors	*B*	*SE B*	*β*	sr^2
Intercept	.000	.000		
Adjective clauses	.002	.000	.578***	.321
Multi-meaning words	.000	.000	-.239**	.055
Adjusted R^2	.325			
Unadjusted R^2	.338			

[1] Positive coefficients indicate that items with more of the language feature exhibit DIF against ELLs
**$p < .01$. ***$p < .001$.

The number of adjective clauses ($\beta = .58$, $p < .001$) and the number of multi-meaning, non-math related words ($\beta = -.24$, $p < .01$) were significant predictors of DIF. Items with more adjective clauses tended to favor non-ELLs. However, items with more multi-meaning words tended to favor ELLs.

The number of adjective clauses and the number of multi-meaning words together accounted for about 34% of the variance in DIF ($R^2 = .34$, adjusted $R^2 = .33$, $p < .001$). Scatterplots for these two linguistic features and uniform DIF are displayed in Figures 3 and 4.

However, after removing the two items that were multivariate outliers, the stepwise regression analysis failed to identify any item features that significantly contributed to the variance in uniform DIF.

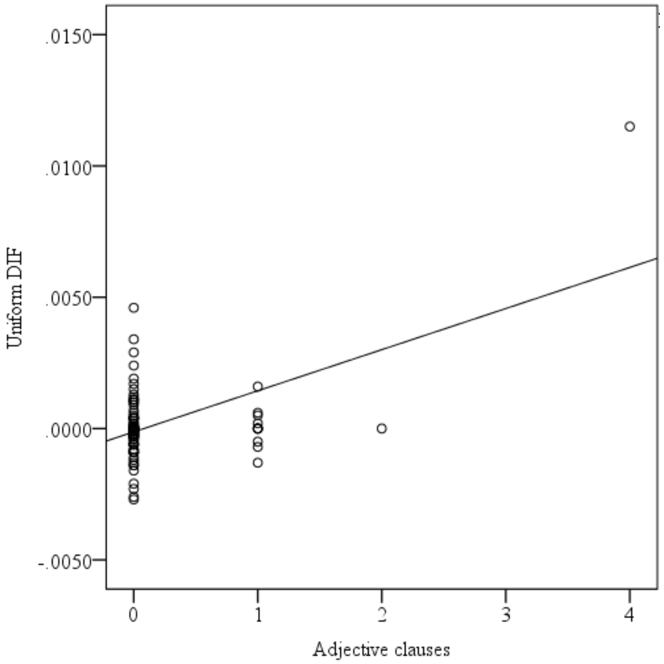

Figure 3. Relationship between uniform DIF statistics with adjective clauses for Grade 4 (R^2 = .28).

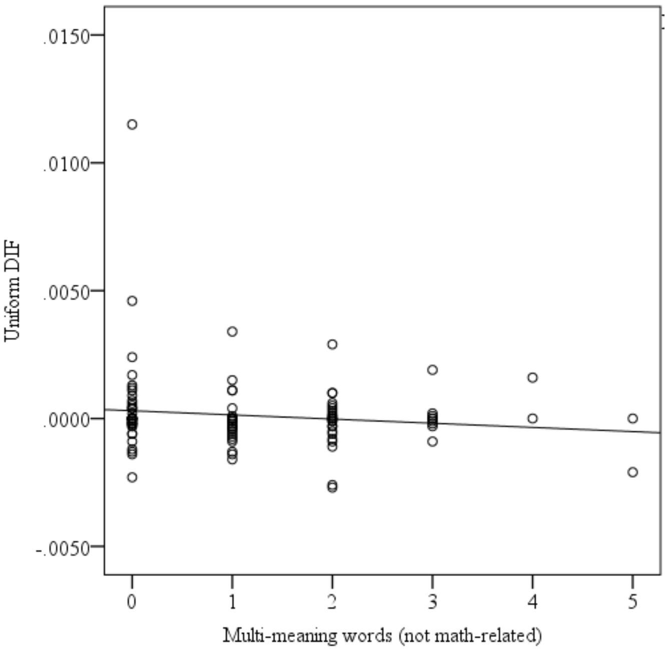

Figure 4. Relationship between uniform DIF statistics with multi-meaning words for Grade 4 (R^2 = .02).

Grade 8. Outlier analyses did not reveal any multivariate outliers in the Grade 8 math items. The multiple regression analysis was conducted with all 119 items included. Table 12 shows the estimated regression models. The stepwise regression procedure revealed two item features, *adjective clauses* (β = -.31, p < .001) and *primarily schematic* (β = -.18, p < .05) that were significant predictors of uniform DIF. Scatterplots for these two linguistic features and uniform DIF are displayed in Figures 5 and 6. Contrary to the expectations outlined in the literature review, items with *more* adjective clauses tended to display uniform DIF statistics *in favor of* ELLs. Additionally, items coded as being primarily schematic tended to favor ELLs. Together, these two item features accounted for about 13% of the variance in uniform DIF (R^2 = .13, adjusted R^2 = .11, p < .001).

Table 12

Grade 8 Regression Analysis for Item Features Predicting Uniform DIF[1]

(N_{items} = 119)

Predictors	*B*	*SE B*	*β*	*sr*^{2}
Intercept	.000	.000		
Adjective clauses	.000	.000	-.313***	.098
Primarily schematic	.000	.000	-.176*	.031
Adjusted R^2	.110			
Unadjusted R^2	.125			

[1] Negative coefficients indicate that items with more of the language feature exhibit DIF in favor of ELLs
*p<.05. ***p<.001.

These results indicate that individual language features may predict DIF. However, these results are not consistent across Grades 4 and 8. These results, their implications, and directions for future research are discussed in the following section.

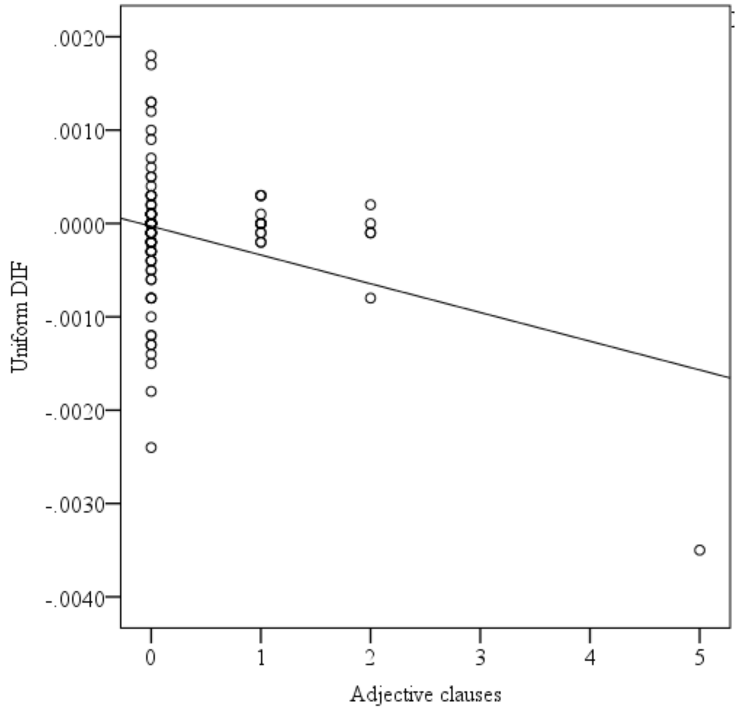

Figure 5. Relationship between uniform DIF statistics with adjective clauses for Grade 8 (R^2 = .09).

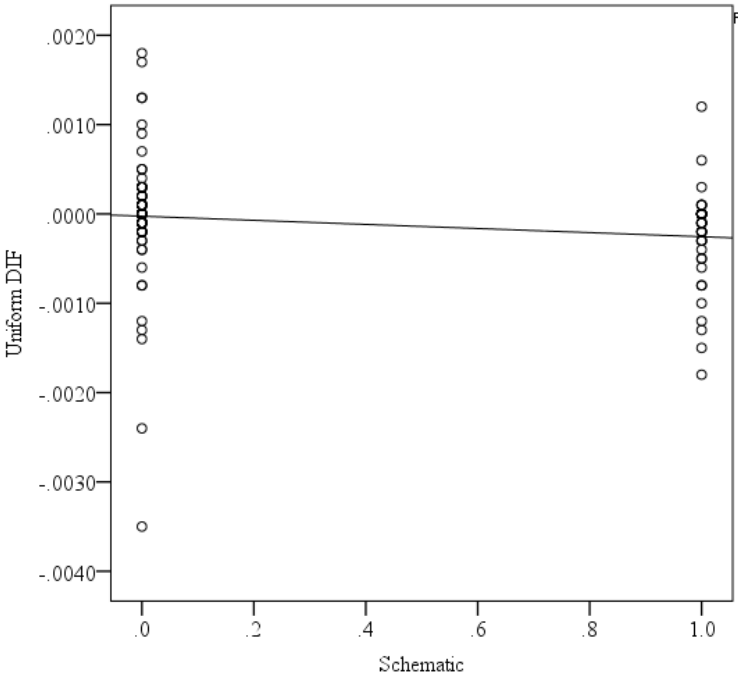

Figure 6. Relationship between uniform DIF statistics with primarily schematic item feature for Grade 8 (R^2 = .03).

Chapter V: Discussion

This study aimed to understand DIF between ELLs and non-ELLs by examining individual item features in a statistical model. This study contributed to previous research by identifying linguistic complexity features that relate to DIF for native Spanish-speaking ELL students. The methodology used in this study included more items in the multiple regression analyses than other previous studies. Additionally, this study controlled for the potentially confounding factor of SES in the DIF analysis and examined a subgroup of the ELL population (native Spanish-speaking students). Results are discussed below as they relate to each research question, and grade level differences are addressed throughout the discussion.

Several items showed evidence of DIF between native Spanish-speaking ELLs and non-ELLs, even after controlling for the SES-proxy variable (i.e., lunch status). For Grade 4, the number of items favoring non-ELLs and the number of items favoring ELLs was about the same. It is also important to note that, between the two test forms for Grade 4, there was substantial variation with regard to the number of items flagged for DIF. On Form 4-1, nearly 73% of the items were flagged for either uniform or non-uniform DIF, but on Form 4-2, only 40% of the items were flagged. This result is likely due to the difference in sample size on the two forms. Form 4-1 included scores for over 21,000 students, while Form 4-2 included scores for about 7,800 students. It is likely that the high number of significant DIF items on Form 4-1 can be attributed to an increase in statistical power from the large sample size. This is also evident in the small magnitudes of the total DIF statistics, which ranged from .0000 to .0142 for Grade 4.

For Grade 8, the number of items flagged for significant DIF was more balanced between Forms 8-1 and 8-2. Furthermore, for Grade 8, slightly more items favored ELLs than non-ELLs. The DIF magnitudes for Grade 8 items were all quite small, and sample sizes for the two Grade 8 forms were quite large (Form 8-1 and Form 8-2 each included scores for over 15,000 students, and the reference and focal group sample sizes were unbalanced). These limitations are discussed in more detail in the section below.

This study supported previous research by finding that math items at both grades included many features of linguistic complexity. As expected, math items at Grade 8 tended to include more linguistic complexity features than math items at Grade 4. This supports findings by Abedi et al. (2008) and Shaftel et al. (2006).

Even though math items often included many descriptive, lexical, and grammatical features, only a few of these features predicted DIF against ELL students. Furthermore, the results were different for Grade 4 and Grade 8. For Grade 4, only two variables, *adjective clauses* and *multi-meaning words*, were found to predict variation in uniform DIF. Items with more adjective clauses were associated with uniform DIF favoring non-ELLs, and items with more non-math-related multi-meaning words were associated with uniform DIF favoring ELL students. These two variables accounted for about 34% of the variance in DIF. These results lead to a few possible conclusions. First, the presence of multiple adjective clauses in an item might cause a barrier to comprehension that unfairly disadvantages ELL students. However, to our knowledge, no other studies support a statistical relationship between adjective clauses and DIF against ELLs.

The second significant predictor of DIF is surprising; it was expected that items with more non-math-related multi-meaning words would be more difficult for students who are still learning English, and these items would present uniform DIF against ELLs. However, results indicate that, after ELLs and non-ELLs were matched on total math score, ELL students performed better than non-ELLs on items with more polysemous words.

A second possible conclusion for the Grade 4 findings, however, is that statistical significance was identified because of one outlier item. It is important to note that, when the item with an outlier value for *adjective clauses* was removed from the multiple regression analysis, the step-wise regression analysis failed to identify any item-level features that accounted for a significant portion of the variance in DIF. Since statistical significance was established due to a single item, it is recommended that follow-up studies be conducted to better establish the relationship between these grammatical features and DIF between Spanish-speaking ELLs and non-ELLs.

The grammatical feature *adjective clauses* was also identified as a significant predictor of DIF at Grade 8. However, items with more adjective clauses tended to favor ELL students at this grade. This result is surprising. It was expected that the presence of more adjective clauses in an item would increase the language load and thus disadvantage students who are non-native English speakers.

For Grade 8, items that were coded as being primarily schematic tended to show uniform DIF in favor of ELL students. This result relates to Martiniello's (2009) study in which schematic features reduced the impact of linguistic complexity on DIF against non-native English speaking students. However, unlike Martiniello's study, the presence of schematic features as a moderator of linguistic complexity on DIF against ELLs could not be examined. For Grade 8, none of the language features predicted uniform DIF against ELLs.

The Grade 8 items featured a wide variety of schematic features. Items coded as being primarily schematic were coded this way because a student could rely on a chart, graph, and/or equation to answer the item without having to rely on language skills. These schematic features can help ELL access math content in the items. Furthermore, coding an item as *primarily* *schematic* may also be an indication of that item's guessability. Some of the schematic items had only one plausible answer choice, as long as the student was math-savvy. For instance, several of these items had equations/numbers as answer choices, and a math-savvy student could easily rule out all but one answer choice. It is possible that on primarily schematic items, native English-speaking students focused on both text and schematic components of the item, while ELL students ignored the text and attended to only the schematic components. This may have reduced the difficulty of these math items for ELLs compared to non-ELLs.

Limitations and Future Research

One limitation in this study was the unbalanced number of students in the reference and focal groups for both grades. Conditions were especially unbalanced at Grade 8, where the two reference groups each had about 14,500 students, and the focal groups each had about 620 students. Future studies should attempt to balance reference and focal groups to alleviate potential problems with DIF detection as identified in Jodoin and Gierl (2001).

Furthermore, while the DIF analyses identified several items with significant DIF for both grades, these findings could reflect the large sample sizes. Additionally, for the LR DIF detection procedure, Jodoin and Gierl (2001) recommended that the chi-square test be used in conjunction with an effect size measure. The present study did report statistical significance and the Nagelkerke pseudo-R^2 statistic, but the examination of effect sizes for DIF showed that these effect sizes were quite small.

This study examined uniform DIF against ELLs but did not include an examination of sources of non-uniform DIF. To better identify the pattern and sources of DIF against ELLs, future research should include the examination of non-uniform DIF.

Item coding. Another potential limitation in this study relates to the item codes for the various linguistic complexity features. Previous research provided some evidence of the statistical relationship between DIF and specific linguistic features. The present study also revealed relationships between a few language features and DIF. However, most language features examined were not statistically related to DIF, and a few language features (multi-meaning words at Grade 4 and adjective clauses at Grade 8) related to DIF in favor of ELLs, results that are counterintuitive to the research base on language demand and ELL students.

One particularly surprising finding was that the presence of more general academic words in the math items was not related to higher DIF against ELLs. This result fails to support findings from Haag et al. (2013), who found a marginal relationship between this lexical feature and DIF. A potential limitation in the present study that could have impacted this result was low inter-rater agreement about which words were considered “general academic vocabulary.” For the Grade 4 items, two of the three item coders agreed on about 66% of their general academic words designations. At Grade 8, two of the three item coders agreed on about 74% of the words identified as general academic words.

There were a few issues related to the identification of general academic words. First, there are several definitions of *general academic vocabulary*. These definitions were provided to the three item coders, but the coders still found it difficult to identify these words in the math items. The researcher facilitated a meeting with the three item coders in which they reconciled many of their disagreements regarding general academic words. However, the item coders did not feel that they had the grade-level expertise (i.e., teaching experience at the Grade 4 or Grade 8 level) to adequately identify all of general academic words. Therefore, during the item coding process, other sources of information (e.g., the EDL Core Vocabularies, BUSD Grade-Level

Academic Vocabulary) were accessed to make final decisions about which words were general academic and which words were not. To encourage higher inter-rater reliability in the identification of general academic words, future studies should 1) designate a special team of item coders comprised of classroom teachers, linguistics, and specialists in ESL to identify these words, and 2) provide better operationalization for this variable.

Another potential limitation with regard to item coding was the identification of passive voice in the math items. For Grade 4, items coded as having passive verb tense usually had the same phrase: "…as shown below." This phrase preceded a table, graph, or other visual feature in the item. Often, the visual feature that this phrase introduced was a schematic feature; this is evident given the significant, positive correlation between passive verb tense and schematic representations at Grade 4. Of the Grade 4 items coded as having passive verb tense, a majority of those items included the same phrase "as shown below." There was little variation in the types of passive verb tense phrases, so this may account for the lack of a relationship between this linguistic complexity feature and DIF against ELLs.

It is also surprising that the presence of Spanish-English cognates in items did not attenuate uniform DIF against ELLs. Cognates are words that have similar spelling and meaning in the two languages, so these words may reduce the linguistic demand for native Spanish-speaking students and thus reduce DIF against native Spanish-speaking ELLs. It was expected that cognates could help native Spanish-speaking ELL students understand the content of math items, an effect identified in Martinello (2008). While cognates were significantly and negatively correlated with uniform DIF (indicating that items with more Spanish-English cognates tended to have lower DIF against native Spanish speaking students), this item feature was not a significant predictor of uniform DIF. A possible reason for this finding is that this study employed the expert judgment of only one bilingual English-Spanish item coder. Future studies should include several coders who can provide their expertise in identifying cognates.

Practical Implications

Results from this study can help guide other researchers in identifying language features that can disadvantage ELL students. The results of this study have implications for item writing, item review, and the continuing analysis of math tests for evidence of validity for ELLs. Additionally, the methodology used in this study (i.e., the prediction of total DIF from individual item features) can be applied more broadly to other reference and focal groups, individual item features, and content assessments. In continuing this line of research, future studies should include more balanced reference and focal groups, continue to improve the operationalization of item-level features, and incorporate additional DIF detection methods.

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational assessment, 8*(3), 231-257.

Abedi, J. (2004). The no child left behind act and English language learners: Assessment and accountability issues. *Educational Researcher, 33*(1), 4-14.

Abedi, J., Bailey, A. L., Butler, F., Castellon-Wellington, M., Leon, S., & Mirocha, J. (2005). The Validity of Administering Large-Scale Content Assessments to English Language Learners: An Investigation from Three Perspectives. CSE Report 663. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST).*

Abedi, J., & Herman, J. (2010). Assessing English language learners' opportunity to learn mathematics: Issues and limitations. *The Teachers College Record, 112*(3).

Abedi, J., Leon, S., Wolf, M. K., & Farnsworth, T. (2008). Detecting test items differentially impacting the performance of ELL students. In M. K. Wolf, J. Herman, J. Kim, J. Abedi, S. Leon & N. Griffin (Eds.), *Providing validity evidence to improve the assessment of English Language Learners* (pp. 55-80). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219-234.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16-26.

Abedi, J., Lord, C., & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance*: 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.

Abella, R., Urrutia, J., & Shneyderman, A. (2005). An examination of the validity of English-language achievement test scores in an English language learner population. *Bilingual Research Journal, 29*(1), 127-144.

Bailey, A. L. (2005). Language analysis of standardized achievement tests: Considerations in the assessment of English language learners. In J. Abedi, A. Bailey, F. Butler, M. Castellon-Wellington, S. Leon & J. Mirocha (Eds.), *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (pp. 79-100). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Bailey, A. L. (2007). *The language demands of school: Putting academic language to the test*: Yale University Press.

Bailey, A. L., & Butler, F. A. (2003). *An evidentiary framework for operationalizing academic language for broad application to K-12 education: A design document*: 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.

Butler, F. A., Bailey, A. L., Stevens, R., Huang, B., & Lord, C. (2004). Academic English in Fifth-grade Mathematics, Science, and Social Studies Textbooks.

Capps, R. (2005). The new demography of America's schools: Immigration and the no child left behind act.

Carpenter, T. P. (1980). Solving Verbal Problems: Results and Implications from National Assessment. *Arithmetic Teacher, 28*(1), 8-12.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31-44.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripkey, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33*(2), 202-214.

Cummins, J. (1980). The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue. *Tesol Quarterly*, 175-187.

De Ayala, R. J. (2009). *The theory and practice of item response theory*: Guilford Press.

Gomez, C. L. (2010). Teaching with Cognates. *Teaching Children Mathematics, 16*(8), 470-474.

Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H. A. (2013). Second language learners' performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction, 28*, 24-34.

Hakuta, K., Butler, Y. G., & Witt, D. (January 2000). How long does it take English learners to attain proficiency? Stanford University: The University of California Linguistic Minority Research Institute.

Hidalgo, M. D., & LÓPez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903-915.

Hopstock, P. J., & Stephenson, T. G. (2003). Native languages of LEP students (Descriptive study of services to LEP students and LEP students with disabilities, Special Topic Report No. 1). Washington, D.C.: U.S. Department of Education.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329-349.

Kopriva, R. (2000). *Ensuring Accuracy in Testing for English Language Learners*: ERIC.

Lee, M. K., & Randall, J. (2011). Exploring Language as a Source of DIF in a Math Test for English Language Learners.

Loughran, J. T., & Skorupski, W. (2014). *Understanding Differential Item Functioning with the General Linear Model*. Paper presented at the National Council on Measurement in Education (NCME), Philadelphia, PA.

Mahoney, K. (2008). Linguistic influences on differential item functioning for second language learners on the national assessment of educational progress. *International Journal of Testing, 8*(1), 14-33.

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

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational assessment, 14*(3-4), 160-179.

Messick, S. (1990). Validity of test interpretation and use.

NCELA. (2011). What languages do English learners speak? (Vol. FS2011-2): National Clearinghouse for English Language Acquisition.

No Child Left Behind (NCLB) Act of 2001, 20, Pub. L. No. 107-110 § 115, 1425 Stat. (2002).

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96*(1), 3-14.

Ramírez, G., Chen, X., & Pasquarella, A. (2013). Cross-linguistic transfer of morphological awareness in Spanish-speaking English language learners: The facilitating effect of cognate knowledge. *Topics in Language Disorders, 33*(1), 73-92.

Rumberger, R., & Gándara, P. (2004). Seeking equity in the education of California's English learners. *The Teachers College Record, 106*(10), 2032-2056.

Schleppegrell, M. J. (2007). The linguistic challenges of mathematics teaching and learning: A research review. *Reading & Writing Quarterly, 23*(2), 139-159.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational assessment, 11*(2), 105-126.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher, 32*(2), 3-13.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Tabachnick, B. G., Fidell, L. S., & Osterlind, S. J. (2001). Using multivariate statistics.

Taylor, S. E., Frackenpohl, H., White, C. E., Nieroroda, B. W., Browning, C. L., & Birsner, E. P. (1989). *EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies*: Steck-Vaughn Company.

Tugwell, H. (2013). BUSD Grade Level Academic Vocabulary (P. D. Office, Trans.). Berkeley, CA: Berkeley Unified School District.

U.S. Department of Education. (2007-08). Public School, BIE School, and Private School Data Files *Schools and Staffing Survey (SASS)*: National Center for Education Statistics.

US Department of Education National Center for Educational Statistics. (2013). Fast facts: English language learners. 2014, from nces.ed.gov/fastfacts/display.asp?id=96

Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational assessment, 14*(3-4), 139-159.

Young, J. W., Cho, Y., Ling, G., Cline, F., Steinberg, J., & Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. *Educational assessment, 13*(2-3), 170-192.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF). *Ottawa: National Defense Headquarters*.

Zumbo, B. D., & Thomas, D. (1997). A measure of effect size for a model-based approach for studying DIF. *Prince George, Canada: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science*.

Appendix A

Linguistic Complexity Coding Protocol

Descriptive Features	**Description & Examples**
(1) Number of Words	**Count the number of words** in the item stem, question, and foils (i.e., answer choices). Do not include words in visuals (e.g., charts, graphs).
(2) Number of Sentences	**Count the number of sentences** in the item stem, question, and foils.
Grammatical Features	**Description & Examples**
(3) Sentence Type	**For each item, code the sentence type as follows:** 0 = Item includes only simple sentences. 1 = Item has one or more of the following: compound sentences, complex sentences, and/or compound-complex sentences.
(4) Prepositional Phrases	For each item, **count the number of words that are in prepositional phrases**. A prepositional phrase consists of a preposition + a noun or noun phrase. Examples include: • There is a rose bush in front of my house. • He hid the toy under the old blanket. • What is the length of the square?
(5) Dependent Adjective Clauses	For each item, **count the number of dependent adjective clauses**. Dependent adjective clauses (also called relative clauses) begin with a relative pronoun (*who, whom, whose, which, that*) or relative adverb (*when, where*). Like adjectives, adjective clauses modify (give more information about) nouns and pronouns. Examples: • An equation, which is missing some information, is pictured below. • A student, who has ten dollars, wants to know how many pencils he can buy.

(6) Modals	**Count the number of modals**. There are 11 modals: *can, could, will, would, shall, should, may, might, must, ought to*, and *had better*.
(7) Passive Verb Tense	**Code this feature as:** 0 = Item DOES NOT include any passive verb forms. 1 = Item DOES include one or more passive verb forms. In passive verb forms, the subject receives the action of the verb. • Active voice: Maria drove the car. • Passive voice: The car <u>was driven</u> by Maria.

Lexical Features	**Description & Examples**
(8) General Academic Words	**Count the number of general academic words in each item**. **DO NOT INCLUDE MATH-RELATED WORDS**. General academic words generally have the same meaning across content areas. These words are more likely to be encountered in an academic environment. Count only non-math-related academic words that are at or above grade level. • Examples: researcher, infer, debate
(9) Multi-meaning words	**Count the number of multi-meaning words in each item.** Multi-meaning words are words that have more than one common definition. In your count, include multi-meaning words that are NOT math related: **Examples:** • "dove" (a bird) vs. "dove" (verb-past tense form of 'dive') • "scale" (a bathroom scale) vs. "scale" (a fish's scale [or] to scale a fish)
(10) Spanish-English Cognates	**Count the number of non-math-related, Spanish-English cognates in the item stem, question, and foils.** Examples: • family (familia) • restaurant (restaurante)

(11) Schematic Representations	
Code	**Description**
0	Text only/primarily pictorial: The item consists primarily of text (or) text along with a picture of a concrete object. The student must rely on language comprehension to correctly answer the item.
1	Schematic: Item may include text, but the item is “primarily schematic.” The student can rely largely on the schematic feature(s) to answer the item correctly. • Schematic representations “are images or symbols that represent spatial or numerical relationships among objects/variables” (Martiniello, 2009, p. 170) • Schematic items: “include a combination of text and nonlinguistic representations depicting mathematical relationships in the form of visual-spatial patterns or algebraic expressions” (Martiniello, 2009, p. 170). Examples: images that represent spatial/mathematical relationships among objects or symbols (e.g., equations, diagrams, tables)

Appendix B

Grade 4 Logistic Regression Differential Item Functioning (DIF) Results

Grade 4 Logistic Regression Differential Item Functioning (DIF) Results

	Nagelkerke ΔR^2				Logistic Regression DIF Coefficients[1]					
Item	Lunch & score	Group	Score X group	Total DIF	τ_0	τ_1	τ_2	τ_3	τ_4	τ_5
4.1.1	0.2292	0.0002	0.0002	0.0004	-4.9807***	0.0256	0.0085	0.1417***	-0.1049	-0.0185*
4.1.2	0.2150	0.0006	0.0005	0.0011	-3.5758***	0.0419	0.0309	0.1376***	-0.2418**	-0.0245**
4.1.3	0.2254	0.0004	0.0000	0.0005	-4.3987***	-0.0982	-0.0176	0.1405***	-0.2032*	-0.0039
4.1.4	0.2558	0.0000	0.0005	0.0005	-4.6113***	0.0912	0.0224	0.1535***	0.0191	-0.0264**
4.1.5	0.0946	0.0011	0.0002	0.0013	-2.2702***	-0.0456	0.1713**	0.084***	0.2589***	-0.0144*
4.1.6	0.1748	0.0000	0.0001	0.0001	-1.8375**	0.0352	-0.0531	0.1378***	0.0719	-0.0092
4.1.7	0.2109	0.0003	0.0005	0.0008	-2.9598***	-0.026	-0.1812*	0.1395***	-0.169	-0.023*
4.1.8	0.2247	0.0005	0.0003	0.0008	-3.4798***	0.0007	-0.151	0.1432***	-0.2215*	-0.0186
4.1.9	0.1385	0.0012	0.0004	0.0016	-2.7123***	0.0482	0.0649	0.1057***	0.2877***	-0.0201**
4.1.10	0.1253	0.0003	0.0020	0.0023	-1.6693***	0.0003	-0.0334	0.0995***	-0.1418*	-0.0419***
4.1.11	0.2741	0.0000	0.0005	0.0005	-4.6594***	0.0152	-0.0592	0.1622***	0.0692	-0.025**
4.1.12	0.0986	0.0009	0.0003	0.0012	-2.5323***	0.0519	-0.137*	0.0833***	0.2352***	-0.0145*
4.1.13	0.1838	0.0003	0.0000	0.0003	-3.7008***	0.073	0.0171	0.1243***	-0.1746*	0.0008
4.1.14	0.1297	0.0004	0.0002	0.0006	-2.6546***	-0.0101	-0.1288	0.1002***	0.17*	-0.0131
4.1.15	0.1238	0.0000	0.0001	0.0001	-1.8873***	0.041	0.1994	0.1116***	-0.0216	0.0089
4.1.16	0.1853	0.0023	0.0011	0.0034	-0.9549	0.1059	0.0008	0.1423***	-0.5509**	-0.0372*
4.1.17	0.3251	0.0009	0.0009	0.0018	-5.5473***	0.0729	0.0506	0.1862***	-0.2929***	-0.0404***
4.1.18	0.3070	0.0001	0.0001	0.0002	-6.4042***	-0.1466	0.0037	0.1743***	-0.0915	0.0153
4.1.19	0.2219	0.0000	0.0002	0.0002	-3.2494***	-0.0679	-0.0857	0.1461***	-0.0189	-0.0157
4.1.20	0.1412	0.0006	0.0009	0.0015	-1.2856**	0.0786	-0.2662*	0.1212***	0.2513*	-0.0269*
4.1.21	0.1513	0.0014	0.0001	0.0015	-1.8719***	0.0415	-0.1422	0.1196***	-0.3989***	-0.0098
4.1.22	0.1851	0.0000	0.0001	0.0001	-3.149***	0.078	0.0369	0.1291***	-0.091	-0.0072

[1] $z = \tau_0 + \tau_1 SES1 + \tau_2 SES2 + \tau_3 \theta + \tau_4 g + \tau_5(\theta g)$; where SES1 = “free lunch,” and SES2 = “no lunch support”
$*p$ < .05. $**p$ < .01. $***p$ < .001.

Grade 4 Logistic Regression Differential Item Functioning (DIF) Results

	Nagelkerke ΔR^2				Logistic Regression DIF Coefficients[1]					
Item	Lunch & score	Group	Score X group	Total DIF	τ_0	τ_1	τ_2	τ_3	τ_4	τ_5
4.1.23	0.1100	0.0000	0.0004	0.0004	-2.4453***	0.0754	-0.017	0.0897***	0.0554	-0.0182*
4.1.24	0.2115	0.0000	0.0001	0.0001	-3.954***	0.0266	-0.0978	0.1357***	0.0003	-0.0074
4.1.25	0.1598	0.0000	0.0000	0.0000	-3.1459***	-0.0217	-0.054	0.1145***	-0.0107	-0.0058
4.1.26	0.2622	0.0002	0.0014	0.0016	-4.2867***	0.1987**	0.1546*	0.1581***	-0.1366	-0.0428***
4.1.27	0.3515	0.0003	0.0002	0.0005	-6.5395***	0.0207	0.0526	0.1961***	-0.1995*	-0.0195
4.1.28	0.1563	0.0001	0.0002	0.0003	-2.613***	-0.0744	-0.0995	0.1157***	0.0794	-0.0151
4.1.29	0.2047	0.0005	0.0001	0.0006	-3.9562***	-0.0765	0.0978	0.1344***	0.1855*	-0.0103
4.1.30	0.0734	0.0004	0.0004	0.0008	-2.1372***	0.1478	0.4963***	0.0788***	0.1676	0.0167
4.1.31	0.0829	0.0000	0.0003	0.0003	-1.1416***	-0.191*	0.1337	0.0838***	0.0027	-0.0159
4.1.32	0.2430	0.0000	0.0000	0.0000	-4.9365***	-0.0435	0.0404	0.1478***	-0.0167	-0.0003
4.1.33	0.1456	0.0000	0.0002	0.0002	-2.6115***	-0.0418	0.0483	0.1098***	0.0204	-0.0157
4.1.34	0.1427	0.0015	0.0006	0.0021	-2.6706***	0.01	0.3149***	0.1103***	0.3212***	-0.0236**
4.1.35	0.2786	0.0013	0.0011	0.0024	-4.3691***	-0.0347	-0.0333	0.1652***	-0.3631***	-0.0395***
4.1.36	0.1805	0.0003	0.0004	0.0007	-3.4932***	0.0214	0.2324***	0.1228***	-0.1716*	-0.0203*
4.1.37	0.2858	0.0000	0.0004	0.0004	-4.8635	0.0250	0.0869	0.1681***	-0.0384	-0.0245
4.1.38	0.1344	0.0014	0.0003	0.0017	-1.2829**	-0.0223	0.3244**	0.1165***	-0.3969**	-0.0145
4.1.39	0.2477	0.0019	0.0015	0.0034	-6.5896***	-0.0751	0.0831	0.1481***	0.4076***	0.0463***
4.1.40	0.1383	0.0002	0.0003	0.0005	-2.2072***	0.0692	-0.0518	0.1108***	0.1444	-0.0146
4.1.41	0.0505	0.0011	0.0000	0.0011	-1.5678***	-0.0895	0.0841	0.059***	0.2477***	-0.0053
4.1.42	0.1211	0.0006	0.0021	0.0027	-4.6448***	0.048	0.3246***	0.1309***	0.2579**	-0.0726***
4.1.43	0.1131	0.0000	0.0001	0.0001	-1.9125***	0.0518	-0.1054	0.0973***	-0.0544	-0.0096
4.1.44	0.1387	0.0002	0.0008	0.0010	-4.3752***	0.0205	0.1499**	0.1129***	0.1151	-0.0339***

[1] $z = \tau_0 + \tau_1 SES1 + \tau_2 SES2 + \tau_3\theta + \tau_4 g + \tau_5(\theta g)$; where SES1 = “free lunch,” and SES2 = “no lunch support”

*p < .05. **p < .01. ***p < .001.

Grade 4 Logistic Regression Differential Item Functioning (DIF) Results

	Nagelkerke ΔR^2				Logistic Regression DIF Coefficients[1]					
Item	Lunch & score	Group	Score X group	Total DIF	τ_0	τ_1	τ_2	τ_3	τ_4	τ_5
4.1.45	0.0972	0.0013	0.0007	0.0020	-1.1477***	-0.0638	-0.0753	0.0879***	-0.3093***	-0.0253**
4.1.46	0.2799	0.0000	0.0009	0.0009	-5.0606***	-0.1963**	-0.0361	0.1673***	-0.0326	-0.0379***
4.1.47	0.2909	0.0011	0.0001	0.0012	-6.8199***	0.1117	0.0172	0.1678***	0.316***	-0.0118
4.1.48	0.1587	0.0004	0.0004	0.0008	-3.049***	-0.0258	-0.0209	0.112***	-0.1837**	-0.0189*
4.1.49	0.2174	0.0001	0.0006	0.0007	-3.8498***	-0.0066	0.0458	0.1381***	-0.0564	-0.0271**
4.1.50	0.2060	0.0005	0.0013	0.0018	-4.0537***	0.0682	0.0808	0.136***	-0.199**	-0.0406***
4.1.51	0.2347	0.0009	0.0009	0.0018	-4.2428***	-0.031	-0.0493	0.1455***	-0.2785***	-0.0325***
4.1.52	0.2161	0.0010	0.0002	0.0012	-4.8521***	-0.0669	-0.0583	0.1367***	0.2989***	0.0118
4.1.53	0.0741	0.0000	0.0001	0.0001	0.4667	-0.183	0.0174	0.0981***	-0.1073	0.005
4.1.54	0.1990	0.0000	0.0006	0.0006	-3.4502***	-0.0334	-0.0896	0.1299***	-0.0613	-0.0249**
4.1.55	0.1126	0.0001	0.0000	0.0001	-3.3844***	0.0698	0.041	0.0886***	-0.0706	0.0025
4.1.56	0.2374	0.0000	0.0005	0.0005	-4.4046***	-0.0743	-0.0673	0.1454***	-0.0606	-0.0233**
4.1.57	0.1613	0.0002	0.0029	0.0031	-3.5226***	-0.0089	0.1359**	0.1258***	0.1148	-0.0622***
4.1.58	0.1321	0.0017	0.0002	0.0019	-1.9552***	-0.0661	0.2998*	0.1145***	0.3721***	-0.0091
4.1.59	0.1307	0.0010	0.0001	0.0011	-3.0606***	-0.051	-0.046	0.1037***	0.2869***	0.0089
4.2.1	0.1037	0.0000	0.0005	0.0005	-2.1801***	0.0385	0.0230	0.1068***	0.0099	-0.0234
4.2.2	0.1051	0.0009	0.0005	0.0014	-0.8026	0.2198	-0.1752	0.1238***	-0.3401	-0.0283
4.2.3	0.2829	0.0004	0.0002	0.0006	-6.2772***	0.0026	-0.0132	0.1999***	0.1813	-0.0207
4.2.4	0.1437	0.0026	0.0002	0.0028	-0.8657	0.2238	0.0381	0.1565***	-0.6312	-0.0251
4.2.5	0.1227	0.0004	0.0012	0.0016	-2.0649***	0.0750	-0.0801	0.1194***	0.1658	-0.0389*
4.2.6	0.0709	0.0012	0.0000	0.0012	1.1904	-0.8074	-0.5408	0.1179***	-0.423	-0.0069
4.2.7	0.2670	0.0000	0.0003	0.0003	-3.6152***	-0.3011	-0.3230	0.2084***	0.0299	-0.0228

[1] $z = \tau_0 + \tau_1 SES1 + \tau_2 SES2 + \tau_3 \theta + \tau_4 g + \tau_5(\theta g)$; where SES1 = “free lunch,” and SES2 = “no lunch support”
*p < .05. **p < .01. ***p < .001.

Grade 4 Logistic Regression Differential Item Functioning (DIF) Results

	Nagelkerke ΔR^2				Logistic Regression DIF Coefficients[1]					
Item	Lunch & score	Group	Score X group	Total DIF	τ_0	τ_1	τ_2	τ_3	τ_4	τ_5
4.2.8	0.1868	0.0001	0.0004	0.0005	-2.8305***	-0.0295	0.0792	0.1619***	-0.0889	-0.0268
4.2.9	0.1803	0.0024	0.0000	0.0024	-7.8409***	-0.1328	0.2191*	0.1704***	0.4878***	-0.0069
4.2.10	0.3067	0.0013	0.0000	0.0013	-7.4093***	-0.1062	0.0318	0.2127***	0.3457**	0.0131
4.2.11	0.1430	0.0002	0.0003	0.0005	-4.7351***	-0.0333	-0.0997	0.1220***	-0.1043	0.0237
4.2.12	0.1346	0.0001	0.0002	0.0003	-3.0905***	0.0371	0.2119	0.1346***	-0.094	0.0180
4.2.13	0.1555	0.0002	0.0022	0.0024	-2.0615***	0.0350	0.1292	0.1394***	-0.1215	-0.0556***
4.2.14	0.1466	0.0034	0.0000	0.0034	-3.6788***	-0.4349**	0.1961	0.1368***	0.483***	0.0000
4.2.15	0.1896	0.0000	0.0040	0.0040	-5.8380***	-0.0562	-0.0560	0.1622***	-0.0273	0.1011**
4.2.16	0.1855	0.0006	0.0013	0.0019	-3.4556***	0.0815	0.0639	0.1515***	-0.2073	-0.0472**
4.2.17	0.0371	0.0001	0.0017	0.0018	-1.5207	0.0625	0.0566	0.0721***	0.1197	0.0509
4.2.18	0.2615	0.0003	0.0001	0.0004	-5.7145***	0.0139	0.0623	0.1891***	-0.1418	-0.0206
4.2.19	0.1554	0.0008	0.0001	0.0009	-2.5175***	0.2290	0.2199	0.1465***	-0.2708	-0.0164
4.2.20	0.1152	0.0009	0.0000	0.0009	-2.6930***	0.1150	0.1132	0.1186***	-0.2827	0.0093
4.2.21	0.2520	0.0016	0.0001	0.0017	-5.4364***	-0.1023	-0.1808	0.1829***	-0.3818***	-0.0147
4.2.22	0.1788	0.0027	0.0018	0.0045	-2.0081**	0.3590**	0.0854	0.1531***	-0.5145***	-0.0568**
4.2.23	0.1307	0.0002	0.0001	0.0003	-2.0494**	0.3040	-0.0111	0.1344***	-0.1354	-0.0168
4.2.24	0.1996	0.0001	0.0012	0.0013	-3.6212***	0.1430	0.2330*	0.1596***	-0.0911	-0.0450**
4.2.25	0.1090	0.0029	0.0000	0.0029	-3.7956***	-0.0935	0.1690	0.1074***	0.4237***	-0.0091
4.2.26	0.2391	0.0001	0.0000	0.0001	-6.1549***	0.1420	0.1624	0.1744***	0.0698	0.0064
4.2.27	0.3767	0.0001	0.0000	0.0001	-8.6861***	-0.1545	-0.0329	0.2568***	-0.0705	-0.0077
4.2.28	0.2179	0.0007	0.0008	0.0015	-4.3554***	-0.1814	-0.2394*	0.1668***	0.2319*	-0.0385*
4.2.29	0.1250	0.0115	0.0027	0.0142	-3.0861**	-0.3851	0.0896	0.1371***	0.9389***	0.0551

[1] $z = \tau_0 + \tau_1 SES1 + \tau_2 SES2 + \tau_3 \theta + \tau_4 g + \tau_5(\theta g)$; where SES1 = “free lunch,” and SES2 = “no lunch support”
*p < .05. **p < .01. ***p < .001.

Grade 4 Logistic Regression Differential Item Functioning (DIF) Results

	Nagelkerke ΔR^2				Logistic Regression DIF Coefficients[1]					
Item	Lunch & score	Group	Score X group	Total DIF	τ_0	τ_1	τ_2	τ_3	τ_4	τ_5
4.2.30	0.2390	0.0004	0.0000	0.0004	-5.9731***	0.0816	0.0087	0.1745***	0.1686	-0.0065
4.2.31	0.3219	0.0006	0.0006	0.0012	-6.0732***	0.0520	-0.0474	0.2232***	-0.2367*	-0.0419*
4.2.32	0.0588	0.0001	0.0000	0.0001	-2.3218***	-0.0452	-0.1306	0.0749***	0.0597	-0.0080
4.2.33	0.2475	0.0001	0.0001	0.0002	-4.0525***	0.4143*	0.2393	0.1954***	-0.1146	-0.0167
4.2.34	0.0922	0.0005	0.0000	0.0005	-1.9779**	-0.0949	0.2390	0.1123***	0.2089	0.0038
4.2.35	0.0295	0.0002	0.0000	0.0002	-2.6376***	-0.0300	-0.0132	0.0574***	-0.0839	-0.0091
4.2.36	0.2413	0.0002	0.0007	0.0009	-2.8531***	0.0626	-0.0304	0.1985***	-0.1509	-0.0370
4.2.37	0.1273	0.0016	0.0011	0.0027	-2.8664***	0.0083	0.3131***	0.1216***	0.3201**	-0.0397**
4.2.38	0.2773	0.0000	0.0002	0.0002	-5.3145***	0.0856	0.2117	0.2019***	0.0857	-0.0201
4.2.39	0.2334	0.0000	0.0000	0.0000	-5.0801***	-0.2835*	-0.2221	0.1743***	0.0313	0.0032
4.2.40	0.2097	0.0003	0.0001	0.0004	-3.7304***	-0.0677	0.0100	0.1720***	0.1427	-0.0147
4.2.41	0.1447	0.0003	0.0002	0.0005	-1.7758*	0.0243	-0.0968	0.1467***	-0.1743	-0.0201
4.2.42	0.3079	0.0008	0.0002	0.0010	-5.8100***	-0.0350	0.0989	0.2163***	-0.2911*	-0.0213
4.2.43	0.2125	0.0001	0.0002	0.0003	-3.4808***	0.0137	-0.0217	0.1735***	-0.0446	-0.0199
4.2.44	0.1793	0.0011	0.0017	0.0028	-3.0613***	0.0593	0.0314	0.1487***	-0.2925**	-0.0519***
4.2.45	0.1495	0.0046	0.0000	0.0046	-3.6654***	0.1949	0.6056***	0.1455***	0.5896***	0.0038
4.2.46	0.1592	0.0021	0.0009	0.0030	-1.9809**	0.0489	-0.0884	0.1419***	-0.4373**	-0.0409*
4.2.47	0.1768	0.0000	0.0002	0.0002	-3.0067***	0.0070	0.0118	0.1555***	0.0205	-0.0177
4.2.48	0.2383	0.0006	0.0005	0.0011	-3.8327***	0.0460	-0.0667	0.1812***	-0.2361	-0.0316
4.2.49	0.1596	0.0000	0.0001	0.0001	-3.6071***	-0.1438	-0.0831	0.1368***	0.0682	-0.0070
4.2.50	0.1871	0.0007	0.0003	0.0010	-5.6251***	0.0941	0.0007	0.1451***	-0.232*	0.0254

[1] $z = \tau_0 + \tau_1 SES1 + \tau_2 SES2 + \tau_3 \theta + \tau_4 g + \tau_5(\theta g)$; where SES1 = “free lunch,” and SES2 = “no lunch support”

$*p < .05$. $**p < .01$. $***p < .001$.

Appendix C

Grade 8 Logistic Regression Differential Item Functioning (DIF) Results

Grade 8 Logistic Regression Differential Item Functioning (DIF) Results

	Nagelkerke ΔR^2				Logistic Regression DIF Coefficients[1]					
Item	Lunch & score	Group	Score X group	Total DIF	τ_0	τ_1	τ_2	τ_3	τ_4	τ_5
8.1.1	0.0685	0.0001	0.0016	0.0017	-1.0473***	0.0589	-0.0101	0.0512***	-0.0938	-0.0341***
8.1.2	0.3595	0.0000	0.0009	0.0009	-5.1141***	-0.042	0.0303	0.1552***	-0.0189	-0.0471***
8.1.3	0.2076	0.0001	0.0000	0.0001	-2.3316***	0.1682	0.1661*	0.1024***	-0.0965	-0.0099
8.1.4	0.1571	0.0000	0.0001	0.0001	-1.761***	0.0336	-0.0143	0.0841***	0.0242	-0.0114
8.1.5	0.2375	0.0000	0.0006	0.0006	-3.1224***	0.122	0.1188	0.1056***	0.0121	-0.0266**
8.1.6	0.2099	0.0002	0.0001	0.0003	-3.3615***	0.0368	0.1401*	0.0965***	0.1553	-0.0109
8.1.7	0.0687	0.0001	0.0007	0.0008	-0.7117*	-0.0047	-0.0199	0.0503***	-0.1029	-0.0224**
8.1.8	0.1863	0.0000	0.0000	0.0000	-2.19***	-0.04	0.0409	0.0944***	0.0258	-0.0078
8.1.9	0.2310	0.0000	0.0001	0.0001	-2.8406***	-0.0387	-0.0096	0.1138***	-0.025	0.0126
8.1.10	0.2041	0.0006	0.0009	0.0015	-2.948***	-0.2961*	-0.2299	0.1071***	0.2495*	0.0352*
8.1.11	0.3295	0.0013	0.0007	0.0020	-3.6623***	-0.1309	0.1519	0.1455***	0.4266***	-0.0319**
8.1.12	0.2333	0.0001	0.0042	0.0043	-2.0766***	0.1789**	0.087	0.1081***	-0.0981	-0.068***
8.1.13	0.2296	0.0018	0.0000	0.0018	-3.1822***	0.1737*	0.0192	0.1026***	-0.487***	0.0011
8.1.14	0.2881	0.0000	0.0007	0.0007	-3.2229***	-0.0847	0.0696	0.1247***	0.0208	-0.0298**
8.1.15	0.1846	0.0001	0.0000	0.0001	-2.0327***	-0.0432	-0.1188	0.0946***	-0.1251	-0.0015
8.1.16	0.2097	0.0001	0.0009	0.001	-2.9321***	0.0024	-0.0157	0.0975***	0.1146	-0.0328***
8.1.17	0.1537	0.0006	0.0011	0.0017	-1.5927**	-0.2584	-0.3282*	0.0974***	-0.2749	0.0437*
8.1.18	0.3481	0.0001	0.0001	0.0002	-4.1162***	-0.0908	-0.0032	0.1518***	0.1303	-0.0067
8.1.19	0.1625	0.0000	0.0011	0.0011	-1.8647***	0.114	0.1083	0.0827***	-0.0225	-0.0304***
8.1.20	0.1768	0.0013	0.0008	0.0021	-1.5615***	-0.0063	0.0106	0.0945***	0.3568***	-0.0274**
8.1.21	0.0862	0.0001	0.0008	0.0009	-0.318	-0.0182	-0.1578*	0.0596***	-0.1013	-0.0246**
8.1.22	0.2793	0.0013	0.0002	0.0015	-2.8462***	0.0574	0.0453	0.1229***	-0.4223***	-0.0183

[1] $z = \tau_0 + \tau_1 SES1 + \tau_2 SES2 + \tau_3\theta + \tau_4 g + \tau_5(\theta g)$; where SES1 = “free lunch,” and SES2 = “no lunch support”
*p < .05. **p < .01. ***p < .001.

Grade 8 Logistic Regression Differential Item Functioning (DIF) Results

	Nagelkerke ΔR^2				Logistic Regression DIF Coefficients[1]					
Item	Lunch & score	Group	Score X group	Total DIF	τ_0	τ_1	τ_2	τ_3	τ_4	τ_5
8.1.23	0.3256	0.0008	0.0001	0.0009	-3.5502***	0.1041	0.0477	0.1388***	-0.3496**	-0.0144
8.1.24	0.2587	0.0001	0.0002	0.0003	-2.6512***	-0.1148	0.1114	0.1229***	0.0515	-0.0157
8.1.25	0.2395	0.0001	0.0002	0.0003	-3.36***	-0.0927	-0.1048	0.1090***	-0.143	0.0146
8.1.26	0.2294	0.0009	0.0005	0.0014	-2.9604***	-0.0023	0.3233*	0.1248***	0.3175*	0.0245
8.1.27	0.3203	0.0005	0.0003	0.0008	-3.7085***	0.0299	-0.1171	0.1472***	-0.2782*	0.0231
8.1.28	0.3415	0.0000	0.0001	0.0001	-4.7123***	0.0251	-0.0412	0.1391***	-0.031	-0.0093
8.1.29	0.1857	0.0000	0.0001	0.0001	-2.8673***	0.0489	-0.0166	0.0906***	-0.0142	0.0088
8.1.30	0.3391	0.0012	0.0000	0.0012	-4.5119***	0.0299	-0.0826	0.1382***	-0.4227***	-0.0116
8.1.31	0.3204	0.0003	0.0018	0.0021	-3.7917***	0.0896	0.0378	0.1355***	0.2058*	-0.055***
8.1.32	0.3425	0.0002	0.0002	0.0004	-3.6442***	-0.0647	0.0998	0.1549***	0.1823	-0.0179
8.1.33	0.2592	0.0000	0.0000	0.0000	-3.0527***	-0.052	-0.0638	0.1208***	-0.0277	0.0043
8.1.34	0.2810	0.0000	0.0006	0.0006	-4.8658***	-0.0828	-0.0313	0.1321***	-0.1336	-0.0361**
8.1.35	0.2076	0.0001	0.0008	0.0009	-2.6151***	0.0856	0.0821	0.0965***	-0.0599	-0.0304***
8.1.36	0.2605	0.0006	0.0011	0.0017	-4.0994***	-0.017	0.0092	0.1198***	-0.2707*	0.0451**
8.1.37	0.2541	0.0001	0.0020	0.0021	-1.7933***	-0.0372	0.0204	0.1208***	0.0958	-0.0449***
8.1.38	0.1508	0.0012	0.0004	0.0016	-1.9793***	-0.0138	-0.0223	0.0796***	0.3399***	-0.0179*
8.1.39	0.3523	0.0002	0.0006	0.0008	-4.4912***	-0.0985	0.0283	0.1649***	-0.1628	0.0399
8.1.40	0.3198	0.0003	0.0007	0.0010	-3.8577***	-0.1378	0.0495	0.1351***	0.2001	-0.0332**
8.1.41	0.3912	0.0000	0.0013	0.0013	-4.5618***	-0.0267	0.0496	0.1624***	0.0623	-0.0535***
8.1.42	0.2982	0.0003	0.0001	0.0004	-3.0213***	0.3609***	0.212*	0.1391***	-0.177	-0.0107
8.1.43	0.3488	0.0000	0.0026	0.0026	-5.4713***	0.0154	-0.1241	0.1876***	0.0703	-0.1025***
8.1.44	0.1863	0.0002	0.0004	0.0006	-2.8715***	0.0486	0.0782	0.0891***	0.1456	-0.0203*

[1] $z = \tau_0 + \tau_1 SES1 + \tau_2 SES2 + \tau_3 \theta + \tau_4 g + \tau_5(\theta g)$; where SES1 = “free lunch,” and SES2 = “no lunch support”
*p < .05. **p < .01. ***p < .001.

Grade 8 Logistic Regression Differential Item Functioning (DIF) Results

	Nagelkerke ΔR^2				Logistic Regression DIF Coefficients[1]					
Item	Lunch & score	Group	Score X group	Total DIF	τ_0	τ_1	τ_2	τ_3	τ_4	τ_5
8.1.45	0.2794	0.0024	0.0002	0.0026	-2.9216***	-0.0011	-0.0993	0.1201***	-0.5708***	-0.0179
8.1.46	0.3120	0.0008	0.0004	0.0012	-3.073***	0.0938	-0.1091	0.1322***	-0.3294**	-0.0249*
8.1.47	0.3115	0.0007	0.0000	0.0007	-3.9815***	-0.1172	0.1472	0.1414***	0.2894*	0.0004
8.1.48	0.1382	0.0001	0.0001	0.0002	-3.2079***	0.0777	0.1485*	0.0747***	0.1426	-0.0109
8.1.49	0.1663	0.0004	0.0003	0.0007	-2.4847***	0.1637*	0.1345*	0.0830***	0.2121*	-0.018*
8.1.50	0.2881	0.0015	0.0000	0.0015	-3.2666***	0.0556	-0.1394	0.1250***	-0.4511***	0.0005
8.1.51	0.1443	0.0001	0.0003	0.0004	-1.3433***	-0.0604	-0.0927	0.0807***	0.0688	-0.0172
8.1.52	0.2003	0.0002	0.0004	0.0006	-1.9569***	0.024	0.0167	0.0976***	-0.1527	-0.0212*
8.1.53	0.2573	0.0001	0.0002	0.0003	-4.1547***	0.1088	0.1354*	0.1120***	0.0898	-0.0173
8.1.54	0.0761	0.0000	0.0010	0.001	-0.6479*	-0.1177	-0.0136	0.0536***	-0.0149	-0.0279***
8.1.55	0.2562	0.0000	0.0002	0.0002	-3.5776***	0.0549	0.1105	0.1115***	-0.0065	-0.0167
8.1.56	0.1985	0.0001	0.0004	0.0005	-2.0852***	-0.0335	-0.1651	0.1137***	0.0936	0.0246
8.1.57	0.3014	0.0001	0.0022	0.0023	-3.2818***	0.0488	0.213**	0.1302***	0.1058	-0.0562***
8.1.58	0.2346	0.0001	0.0006	0.0007	-1.9812***	0.0998	0.0484	0.1129***	-0.1079	-0.0267**
8.1.59	0.3309	0.0010	0.0004	0.0014	-3.0227***	0.092	-0.1305	0.1427***	-0.3904**	-0.0229
8.1.60	0.3741	0.0000	0.0002	0.0002	-5.0796***	-0.0273	0.1638*	0.1548***	0.0873	-0.02
8.1.61	0.1095	0.0008	0.0004	0.0012	-1.9282***	0.0875	0.0231	0.0654***	-0.2785**	-0.0192*
8.2.1	0.1694	0.0000	0.0001	0.0001	-2.936***	0.0657	0.0778	0.0901***	0.0898	-0.0097
8.2.2	0.3158	0.0005	0.0015	0.0020	-4.3793***	0.0029	0.0543	0.1499***	0.2999**	-0.063***
8.2.3	0.2932	0.0001	0.0001	0.0002	-2.9578***	-0.4103**	-0.3071*	0.1588***	0.1228	0.0075
8.2.4	0.1043	0.0000	0.0002	0.0002	-1.6351***	-0.0041	0.0188	0.0743***	0.0847	0.0127
8.2.5	0.3256	0.0002	0.0007	0.0009	-4.3409***	0.0037	0.1369*	0.1493***	-0.1505	-0.04***

[1] $z = \tau_0 + \tau_1 SES1 + \tau_2 SES2 + \tau_3 \theta + \tau_4 g + \tau_5(\theta g)$; where SES1 = “free lunch,” and SES2 = “no lunch support”
*p < .05. **p < .01. ***p < .001.

Grade 8 Logistic Regression Differential Item Functioning (DIF) Results

	Nagelkerke ΔR^2				Logistic Regression DIF Coefficients[1]					
Item	Lunch & score	Group	Score X group	Total DIF	τ_0	τ_1	τ_2	τ_3	τ_4	τ_5
8.2.6	0.2110	0.0002	0.0001	0.0003	-2.5352***	-0.0259	0.032	0.1175***	-0.1639	0.009
8.2.7	0.1875	0.0008	0.0001	0.0009	-2.2576***	-0.0341	-0.0683	0.0988***	-0.2875**	-0.0128
8.2.8	0.2495	0.0002	0.0013	0.0015	-2.7522***	-0.0439	0.1755**	0.1217***	-0.1604	-0.0447***
8.2.9	0.2391	0.0001	0.0015	0.0016	-2.3142***	0.0288	-0.0991	0.1160***	-0.0972	-0.0436***
8.2.10	0.2735	0.0003	0.0009	0.0012	-4.2763***	-0.0266	-0.0464	0.1358***	-0.2194	0.0427**
8.2.11	0.3059	0.0013	0.0001	0.0014	-4.5053***	0.0742	0.075	0.1399***	-0.4396***	-0.0204
8.2.12	0.3366	0.0001	0.0000	0.0001	-4.1061***	-0.0256	-0.2429**	0.1550***	-0.1459	-0.0018
8.2.13	0.2174	0.0004	0.0009	0.0013	0.2981	-0.2607	-0.1706	0.1562***	-0.2414	-0.0409
8.2.14	0.1963	0.0002	0.0019	0.0021	-2.3411***	0.0232	0.0662	0.1031***	-0.1701	-0.0498***
8.2.15	0.3383	0.0000	0.0001	0.0001	-4.0789***	0.1291	-0.0367	0.1579***	0.0592	-0.0135
8.2.16	0.2711	0.0018	0.0013	0.0031	-2.8475***	-0.1823*	-0.0336	0.1327***	0.4777***	-0.0426***
8.2.17	0.2454	0.0005	0.0005	0.0010	-2.5875***	-0.1324	-0.0548	0.1254***	0.2351*	-0.0257*
8.2.18	0.1721	0.0001	0.0001	0.0002	-2.2817***	0.0079	-0.0049	0.0932***	-0.0754	-0.0133
8.2.19	0.2967	0.0003	0.0017	0.0020	-2.6948***	-0.0624	0.2106*	0.1490***	0.175	-0.0502***
8.2.20	0.2831	0.0003	0.0000	0.0003	-2.7705***	-0.1684	-0.1161	0.1504***	-0.221	0.0005
8.2.21	0.1981	0.0000	0.0001	0.0001	-2.5993***	0.0046	-0.0765	0.1060***	-0.0562	0.0022
8.2.22	0.3232	0.0001	0.0011	0.0012	-3.8968***	0.0854	0.0807	0.1467***	-0.109	-0.0453***
8.2.23	0.2032	0.0003	0.0009	0.0012	-4.8836***	-0.044	0.012	0.1261***	0.2975*	-0.0523***
8.2.24	0.1994	0.0003	0.0007	0.0010	-2.73***	-0.0284	-0.0055	0.1013***	0.1815	-0.0309**
8.2.25	0.1633	0.0014	0.0000	0.0014	-1.8115***	0.0266	-0.0678	0.0928***	-0.3982***	-0.0059
8.2.26	0.3421	0.0000	0.0000	0.0000	-4.3284***	0.066	0.0724	0.1634***	0.0051	-0.0025
8.2.27	0.2753	0.0000	0.0000	0.0000	-4.026***	0.1051	-0.0132	0.1305***	0.065	0.0066

[1] $z = \tau_0 + \tau_1 SES1 + \tau_2 SES2 + \tau_3\theta + \tau_4 g + \tau_5(\theta g)$; where SES1 = “free lunch,” and SES2 = “no lunch support”
$*p < .05$. $**p < .01$. $***p < .001$.

Grade 8 Logistic Regression Differential Item Functioning (DIF) Results

	Nagelkerke ΔR^2				Logistic Regression DIF Coefficients[1]					
Item	Lunch & score	Group	Score X group	Total DIF	τ_0	τ_1	τ_2	τ_3	τ_4	τ_5
8.2.28	0.2965	0.0005	0.0001	0.0006	-4.0036***	0.1199	-0.034	0.1340***	-0.2554*	-0.0155
8.2.29	0.1136	0.0002	0.0000	0.0002	-2.2993***	0.0843	0.0314	0.0703***	-0.1159	-0.0054
8.2.30	0.1723	0.0000	0.0001	0.0001	-3.2335***	-0.0198	-0.0489	0.0913***	0.0567	-0.0123
8.2.31	0.2063	0.0001	0.0011	0.0012	-3.8539***	0.0082	-0.0174	0.1078***	-0.1058	0.0438**
8.2.32	0.1260	0.0000	0.0038	0.0038	-1.5193***	0.1932**	0.1261*	0.0807***	-0.0477	-0.064***
8.2.33	0.3881	0.0000	0.0001	0.0001	-4.8087***	0.1663	0.0452	0.1822***	0.0283	-0.0098
8.2.34	0.3199	0.0001	0.0012	0.0013	-4.4898***	0.1318	0.0286	0.1485***	0.1015	-0.0528***
8.2.35	0.2224	0.0002	0.0000	0.0002	-3.4596***	-0.0571	-0.056	0.1100***	0.131	0.0023
8.2.36	0.3150	0.0000	0.0018	0.0018	-3.6362***	0.0259	0.0541	0.1446***	0.0747	-0.0572***
8.2.37	0.2872	0.0000	0.0000	0.0000	-4.4123***	-0.1331	-0.1011	0.1320***	0.0054	0.0056
8.2.38	0.2972	0.0003	0.0000	0.0003	-2.8938***	0.3261**	0.3029*	0.1606***	-0.1955	-0.0124
8.2.39	0.1700	0.0000	0.0004	0.0004	-2.3494***	-0.0212	0.0222	0.0915***	0.0058	-0.0205*
8.2.40	0.1544	0.0002	0.0003	0.0005	-2.3933***	0.1867*	0.1366	0.0947***	-0.1306	0.0211
8.2.41	0.2796	0.0001	0.0001	0.0002	-3.1151***	-0.0659	-0.0996	0.1398***	0.1074	-0.0121
8.2.42	0.2640	0.0010	0.0016	0.0026	-2.6118***	-0.0622	0.0631	0.1308***	0.356***	-0.0458***
8.2.43	0.1023	0.0012	0.0000	0.0012	-2.0219***	0.0711	-0.0013	0.0664***	-0.3548***	0.002
8.2.44	0.2241	0.0004	0.0018	0.0022	-1.3891***	0.1701	0.1286	0.1215***	-0.2296*	-0.0479***
8.2.45	0.1231	0.0000	0.0013	0.0013	-1.53**	-0.0604	-0.1411	0.0961***	-0.0017	0.047*
8.2.46	0.1545	0.0000	0.0003	0.0003	-2.6611***	0.0614	0.0197	0.0852***	0.0344	-0.0188*
8.2.47	0.2087	0.0003	0.0001	0.0004	-2.2719***	-0.1621	0.056	0.1129***	-0.1967	-0.0098
8.2.48	0.1513	0.0002	0.0047	0.0049	-1.6391***	0.0047	0.0365	0.0946***	-0.1438	-0.0777***
8.2.49	0.2178	0.0000	0.0000	0.0000	-1.7735***	0.0058	-0.173	0.1302***	-0.016	-0.0051

[1] $z = \tau_0 + \tau_1 SES1 + \tau_2 SES2 + \tau_3\theta + \tau_4 g + \tau_5(\theta g)$; where SES1 = “free lunch,” and SES2 = “no lunch support”
*p < .05. **p < .01. ***p < .001.

Grade 8 Logistic Regression Differential Item Functioning (DIF) Results

	Nagelkerke ΔR^2				Logistic Regression DIF Coefficients[1]					
Item	Lunch & score	Group	Score X group	Total DIF	τ_0	τ_1	τ_2	τ_3	τ_4	τ_5
8.2.50	0.1714	0.0000	0.0001	0.0001	-3.7713***	0.1899**	-0.0433	0.0894***	-0.0264	0.0069
8.2.51	0.2139	0.0000	0.0000	0.0000	-2.4169***	-0.214*	0.0111	0.1198***	0.0086	0.0008
8.2.52	0.2327	0.0001	0.0005	0.0006	-3.8465***	-0.0165	-0.1	0.1156***	-0.0957	0.029*
8.2.53	0.1445	0.0017	0.0000	0.0017	-0.9105*	-0.0618	-0.1057	0.1064***	0.4253**	-0.0076
8.2.54	0.1984	0.0003	0.0000	0.0003	-3.5358***	-0.0691	0.1906**	0.1022***	0.1746	0.0021
8.2.55	0.3401	0.0002	0.0002	0.0004	-4.3512***	0.0813	-0.0601	0.1513***	-0.1871	-0.019
8.2.56	0.1689	0.0035	0.0005	0.0040	-0.8652*	-0.0681	-0.0813	0.0991***	-0.6711***	-0.026*
8.2.57	0.1871	0.0004	0.0015	0.0019	-2.7149***	-0.0023	0.1365*	0.1025***	-0.2063*	-0.047***
8.2.58	0.2971	0.0003	0.0021	0.0024	-3.1159***	-0.0147	0.1554*	0.1397***	0.2223*	-0.0583***

[1] $z = \tau_0 + \tau_1 SES1 + \tau_2 SES2 + \tau_3 \theta + \tau_4 g + \tau_5(\theta g)$; where SES1 = “free lunch,” and SES2 = “no lunch support”
*p < .05. **p < .01. ***p < .001.