Differential Item Functioning (DIF) among Spanish-Speaking English Language Learners (ELLs) in State Science Tests

Maria O. Ilich

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:
Min Li, Chair
Catherine S. Taylor
Guillermo Solano-Flores

Program Authorized to Offer Degree:
College of Education

University of Washington

Abstract

Differential Item Functioning (DIF) among Spanish-Speaking English Language Learners (ELLs) in State Science Tests

Maria O. Ilich

Chair of the Supervisory Committee:
Associate Professor Min Li, Ph.D.
College of Education

Psychometricians and test developers evaluate standardized tests for potential bias against groups of test-takers by using differential item functioning (DIF). English language learners (ELLs) are a diverse group of students whose native language is not English. While they are still learning the English language, they must take their standardized tests for their school subjects, including science, in English. In this study, linguistic complexity was examined as a possible source of DIF that may result in test scores that confound science knowledge with a lack of English proficiency among ELLs. Two years of fifth-grade state science tests were analyzed for evidence of DIF using two DIF methods, Simultaneous Item Bias Test (SIBTest) and logistic regression. The tests presented a unique challenge in that the test items were grouped together into testlets—groups of items referring to a scientific scenario to measure knowledge of different science content or skills. Very large samples of 10, 256 students in 2006 and 13,571 students in 2007 were examined. Half of each sample was composed of Spanish-speaking ELLs; the balance was comprised of native English speakers. The two DIF methods were in agreement about the items that favored non-ELLs and the items that favored ELLs. Logistic regression effect sizes were all negligible, while SIBTest flagged items with low to high DIF. A decrease in socioeconomic status and Spanish-speaking ELL diversity may have led to inconsistent SIBTest effect sizes for items used in both testing years. The DIF results for the testlets suggested that ELLs lacked sufficient opportunity to learn science content. The DIF results further suggest that those constructed response test items requiring the student to draw a conclusion about a scientific investigation or to plan a new investigation tended to favor ELLs.

Table of Contents

List of Figures

List of Tables

Dedication

To my parents.

Acknowledgements

There are many people I need to thank for their support while writing this dissertation and during this degree. First, and foremost, I thank my tireless committee, who were always ready with constructive criticism, guidance, and expertise. Dr. Cathy Taylor and Dr. Willy Solano-Flores are inspirational and diligent in making sure nothing was missed. I would like to express my sincere gratitude to Dr. Buz Hunt for his deep insights, his original and enlightening ideas and for sticking with me to the end. I'm very grateful to Dr. Min Li, my strong and supportive advisor, who was always available and confident, from the early morning to late into the night. My many thanks go out to Dr. Liz Sanders for donating her time and her statistical skill.

Ming-Chi "Aarron" Lan, Dr. Shin-Ping Tsai, Nicole Casillas, Dr. Tina Wang and Dr. Wendy Smith, my doctoral classmates, I thank you for the pep talks and the technical discussions, and for helping me navigate through the program.

I appreciate all the encouragement and support I received from my brother, Bobby Ilich, and my friends, Christina Rachal, Melanie Gersten and Evelyn Flores who provided a sounding board for my almost daily thoughts and questions, many of which probably made very little sense.

When there was very little time to spare, I received the help I needed from the raters, and from Anne Seitz who sprang to the rescue, as she has a tendency to do.

I couldn't have made it through the doctoral process without Miyako Kodama and James Ott. Their delicious cooking nourished me, and they pushed me to take fun breaks and exercise with them, often when I had passed the point of productivity, so that I could come back with a clearer mind and more energy.

Two very capable women who never fail to awe me, Erin Riesland and Marisabel Almer kept me on track, particularly towards the end of this journey. They spent their precious time proofreading and editing the dissertation and always believed in me. I cannot find the words to fully express my gratitude for their encouragement and help.

Finally, I'm very appreciative of the generous financial support to The Boeing Company and for the patience and rallying of my managers Billy Gibbs, Andy Thomsen, and my co-workers Sharon Cassidy, and the rest of the South Park office.

Chapter I: Introduction

According to the U.S. Census Bureau report in 2010[1], there are more than 55 million people in the United States who do not speak English at home. English Language Learners (ELLs) are children whose primary language is not English. ELLs are a growing population of students in the United States and, as such, are gaining the attention of researchers of education. Research on ELLs in recent years covers a broad range of areas, from examining the effectiveness of test accommodations (Wolf, Kim, Kao, & Rivera, 2009) to state policies meant to increase English acquisition (Lawton, 2012). In school classrooms, many children from these homes are learning English and their school subjects simultaneously. They will hear school lectures and read textbooks in English even though they haven't reached a level of English proficiency adequate enough to fully grasp the lessons in their entirety. Previous academic mastery in a foreign language does not prevent ELLs from struggling to understand the same level of material when it is presented in English. When tested on science or math knowledge, ELL students are expected to also exhibit their command of English in order to read and respond to the test questions, even though English is not the subject being tested.

It has been reported that acquiring English language proficiency for grade school children can take approximately five to seven years (Abedi & Gándara, 2006). However, reaching a proficiency level of academic English reading and writing that would raise an ELL's academic achievement to the grade-level of their peers can sometimes take as long as eight to 10 years (Collier, 1987, 1995). Differences in speed of acquiring English

[1] http://www.census.gov/prod/2010pubs/acs-12.pdf.

proficiency may depend on a number of factors such as the age at which English is first taught, and how many years of formal schooling students received in their native language (Collier, 1995). This indicates that when ELLs are ready to be tested in English can vary. More importantly, it can take many years for them to perform as well as native English speakers on a test written in English.

Standardized tests are administered by each state in the nation every year to determine the effectiveness of schools to provide the requisite education to all students. The use of national, state, and classroom exams designed for native English speakers (non-ELLs) have been shown to be problematic for ELLs because they don't understand the test questions (Martiniello, 2008). As a result, researchers express concerns about the accuracy of ELLs' individual and group academic rankings for standardized math and science subject tests (Abedi, 2008; Abella, Urrutia, & Shneyderman, 2005; Solano-Flores & Trumbull, 2003).

When test developers are formulating the standardized tests, to prevent potential bias, test items are piloted by comparing the performance of demographically varied groups. To ensure that the test items do not function differently for equally skilled students, testing programs obtain item statistics that match the performance of a minority group to those of the majority group to investigate potential sources of bias so they can be prevented in the future. Calculating students' true abilities is not a simple task; instead, statistical methods are used to estimate true ability scores. In the context of assessment and psychometrics, bias specifically refers to the influence of any factor resulting in a misjudgment of the skill level of a demographically related group (Scheuneman &

Slaughter, 1991). It is important to note that conclusive evidence is required to indicate that an item has bias. The analysis must go further than identifying any disparity in performance between two groups (Camilli & Shepard, 1994; Hambleton & Jones, 1994).

During the item analysis stage of the test development process, comparisons for each test item are made between groups of various ethnic backgrounds and genders. However, technical reports that describe piloting of items and statistical comparisons made to examine bias prior to administrating the tests do not commonly use ELLs as one of the groups of interest. Recently, researchers have analyzed post-administration test results to determine how accurately ELLs' scores reflected their knowledge level and to attempt to identify which features of test items may be problematic for them (Abedi, 2009; Abedi & Lord, 2008; Luykx et al., 2007; Martiniello, 2008; Shaftel, Belton-Kocher, Glasnapp & Poggio, 2006; Wolf & Leon, 2009; Young et al., 2008; Young, Holtzman & Steinberg, 2011). In these studies test items have been identified that demonstrate the English proficiency level required to pass the item is higher than the English proficiency level of ELLs taking the test.

ELL students may find science tests complex due to the academic level vocabulary they contain; however, at this time, few studies have focused on ELLs' science test performance in comparison to non-ELLs' (Luykx et al., 2007; Wolf & Leon, 2009; Young et al., 2008). Studies of potential bias in science test items is essential because of the US government's goal to increase interest and encourage pursuit of higher education in the fields of science, technology, engineering and math (STEM). Educators and policy makers maintain that science education is central to both preparation to be

socially responsible adults and academically successful in higher education. Right now, there is a concern that the U.S. has a shortage of college graduate sufficient to meet the needs of the science and math professions. In order to maintain a globally competitive workforce and to fulfill the needs of future employers, all students must receive a rigorous education in science and their strengths in these areas must be nurtured and developed early to ensure continued interest in STEM fields. Today's ELLs are the future's workforce, so identifying opportunities to learn and accurately guiding future instruction will help the nation to prosper. For that reason, this dissertation concentrates on examining potential test bias for Spanish-speaking ELLs' on state science assessments.

Current Approaches to ELL Testing Issues

In the last few years, research using *differential item functioning* (DIF) analysis has become a popular for comparing ELLs to non-ELLs of the same ability level to learn whether students have an equal chance of passing each item on a test. DIF is the term used to describe a set of statistical methods that determine when an item on an exam functions differently for one group of people than for another, such that the mean score or the distribution of scores (or both) for those groups differ significantly. When undertaking DIF analyses, the two groups being compared are referred to as the reference group (generally considered to be dominant), and the focal group (the group of concern). From the sample based on demographic criteria (for example males and females), all the subjects are divided into one of two groups of test-takers, for example males and females. The theory, expectation, or incident this study is built upon dictates which groups are

deemed the focal group. For example, if an item is suspected to have gender bias against females, the males would be the reference group and the females the focal group.

Linguistic Complexity. Linguistic complexity, a possible culprit of problematic test questions, includes such issues as the use of idioms, colloquialisms, excessively long sentences or overly complicated language structures (Abedi & Lord, 2001; Kopriva, Emick, Hipolito-Delgado, & Cameron, 2007; Martiniello, 2008; Solano-Flores & Trumbull, 2003; Wolf & Leon, 2009). Improving how test items are constructed by specifically reducing their linguistic complexity has been strongly encouraged by researchers as a way to diminish cognitive load, thus increasing the validity of assessment scores (Abedi, 2004; Kopriva, 1999; Rivera & Stansfield, 2004).

In this dissertation, linguistic complexity is the lens used to examine the language on test items as a source of measurement error. The modern views of literacy hold that semiotics is an integral part of the study of linguistics. While maintaining awareness of the importance of non-textual features of linguistics, the linguistic complexity framework in this dissertation addresses mainly the textual features of test items. It is meant to guide and assist test developers in recognizing features of test items that may inadvertently confound subject area knowledge (e.g. science or math) with lack of English proficiency.

Linguistic complexity in the following chapters diverges from previous definitions by taking into consideration the heterogeneity of the ELL population that can vary greatly from sample to sample in their knowledge structures. Since each ELL group possesses different sets English academic skills, it is not possible to anticipate what will

work for each and all the students. What is and what is not linguistically complex may vary with each sub-group of ELLs.

Abedi (2008) and Kopriva (2000) describe linguistic modification and linguistic simplification as methods to increase ELLs' understanding of test items by using the most common words that would be encountered in plain conversations. These approaches serve as a departure point for the linguistic complexity conceptual framework in this dissertation. Linguistic simplification involves a set of rules for the use of grammar, discourse style, and vocabulary on standardized tests in order to make them easier for ELLs to decode and make sense of. While these categories of language have been created for researchers to study the language on the tests, they do not reflect how language is remembered or used in written text or spoken conversation by ELLs.

In addition, the framework acknowledges that each test item's set of linguistic demands is shaped by the content and format in which it is administered (Solano-Flores, 2012). For this reason, each item has different constraints for how it can be written. Because of the heterogeneity of ELLs, items cannot be expected to be more or less difficult for ELLs just because they have or do not have a particular linguistic feature. Instead it is more reasonable to focus on probablisitic patterns in which items' linguistic features that are more difficult for ELLs than non-ELLs.

Issues with ELLs as Test Takers

No Child Left Behind (NCLB) act describes ELLs as those either not born in the United States or whose native language is not English and who are unable to meet a state's English proficiency level when assessed. The law also indicates that ELLs could

be of Native American or Alaskan decent as long as another language is primarily used in the home instead of English. Clearly, this is a very broad definition, for which the original intent was school accountability for ELLs' education. However, this blanket category created by *No Child Left Behind* does not address the degree of diversity within this large group of students across the United States. Some researchers have looked deeper into the demographics of ELLs to include such variables as language spoken at home, highest parental educational level and socioeconomic status, refugee status, permanent residency, naturalization and the breakdown of country of origin of the parent or the ELL. The problem remains that categorical data are not useful in illuminating how much English ELLs have been exposed to, or for instance, what specific words they know. Such definitions and statistics therefore remain limited in their ability to inform test developers how best to create test items for use with ELLs. In the end, the difficulty of testing ELLs' content area knowledge without confounding it with their English proficiency remains.

ELLs cannot be expected to perform at the same level in reading and writing as their peers while they are still developing the language of instruction as a second language. ELLs' English reading and writing skills will not yet have reached their peers' level of English proficiency. Even for math and science, which are sometimes wrongly assumed not to pose linguistic challenges to students, ELLs might not be able demonstrate knowledge of required concepts due to cultural differences in the item format, presentation, or non-verbal structural components. It must be mentioned that no test is fully language or culture-free, and that reading and writing cannot be entirely

removed from any test in English, unless the test is of purely non-verbal mathematical calculation or spatial manipulation. However, variability in the test's level of English usage can affect the accurate measurement of the construct of interest, e.g., math or science. ELLs may struggle with responding to questions in English due to English writing skills. Test responses that require drawing diagrams rather than writing out explanations may assist ELLs in communicating their understanding of a science concept.

In examining performance differences between ELLs and non-ELLs little attention is paid to examining or questioning the characteristics of the items. Typically, test developers provide limited evidence in support of the argument that the items are appropriate for ELLs–and more importantly that ELL scores would accurately reflect their knowledge level of a given subject. Valid interpretations of test results depend on ELLs' full understanding of the test questions and the ability to communicate that knowledge in English. An analysis of test items through the perspective of ELLs alone could validate their viewpoints and needs while simultaneously concentrate research attention on the assumptions being made by the test.

Validity, Construct Irrelevant Variance and Impact

A crucial part in the test development and validation process is the need to validate the use and interpretation of test scores. Large-scale assessments, such as the National Assessment of Educational Progress (NAEP), are put through a rigorous process to ensure test results have high validity, and to support test developers' claim that the test is reliable and that the results provide an accurate measure what they purport to measure. Kane (2006) asserts that meaning derived from test scores must be supported by data in

order to justify the social ramifications of the interpretation. According to Kane, strong evidence is needed to support claims made from test score interpretations, and the application of those scores to a specific population.

Evidence type and validation strategy are dependent on the consequences of those claims, as well as the expected interpretation of the scores. Accordingly, the greater the consequences, the more rigorous the validation process will need to be (Kane, 2006). In addition, the claims must match the evidence. For example, if the test is meant to generalize to a clearly defined ability, the validation process should involve measures of the following: the dependability of scores, whether the content covers the competencies of the ability, related abilities, and whether scores exclude unrelated ability measurements. Studies should provide evidence that high scores indicate high ability outside of the test and vice versa. The type and process of validation required depends on social considerations, claims, and practical uses of the test. However, a minimal validation should involve evidence that the assessment results can accurately and reliably use to evaluate the population and the ability being assessed.

There are varying degrees of validity—validity is not an all or nothing concept—meaning that there is no such thing as "a perfect test." Although efforts are taken to reduce unintended sources of error during the construction process, there will always be some sources of error in any test. An in-depth discussion on all the considerations involved in item and test validation is beyond the scope of this paper; however the studies focusing on item characteristics will be viewed through the lens of such validity considerations.

Psychometricians focus on uncovering potential causes of bias. In the context of assessment and psychometrics, the term *bias* specifically refers to the influence of any factor which results in mis-measuring the skill level of a demographically related group by either overrating or underrating individuals' ability levels (Scheuneman & Slaughter, 1991). Still, it is not the test taking itself that results in bias, rather it is the applied consequences of the test's scores that will result in bias (Shepard, 1982). State and classroom test results are used to make educational decisions that, besides social consequences, can adversely affect the long-range achievement levels of these students. Thus, the best use of these tests scores is evaluation of what students have learned in school. For that use to be feasible for ELLs, researchers need to consider the population's various characteristics, including individual's English proficiency levels and any other factors that could impact their scores.

The *construct related evidence for validity* of an assessment refers to the degree of association between the test score and what trait or ability it is meant to describe or predict. For example, an algebra assessment with high construct validity would imply that the test score has accurately represented the algebraic knowledge level of each student who takes the test. Conversely, construct-irrelevant variance, according to Haladyna and Downing (2004), is any facet of the test items that inflates or deflates test scores of a group of participants such that the true scores are obscured. Thus, we are not apprised of the participants' true knowledge levels. There are threats to the validity of test score interpretations if factors other than the targeted construct can affect the scores, i.e., if the scores do not isolate what is being measured and are corrupted by the influence

of construct irrelevant factors or variables. A poorly written or overly complicated sentence on a test can potentially cause confusion for some students, and this confusion may lead to misunderstanding the question being asked. In this case, the long and complicated sentence could be replaced with a simplified sentence while maintaining the same meaning. Therefore, developers must explore systematic sources of error variance during the item piloting process and work to remove their sources of error from the final distributed test.

While all students are required to know the technical language relevant to the assessed content area in order to get a test item correct, additional academic language can appear in a test item as well, and is considered by researchers to contribute to construct-irrelevant variance (Abedi, 2006, 2010; Solano-Flores, 2006). Science terms taught at each grade level may be entirely new, or may build upon previous science lessons. For example, a common third grade science lesson introduces the following terms: eco-system, conifer, and precipitation. These scientific terms would constitute the construct-relevant vocabulary of the lesson. These words are directly related to the construct being assessed—third grade science—and should be included on the assessment. By contrast, construct-irrelevant vocabulary would constitute terminology unrelated to the lessons taught in third grade science. For example, the phrase "based on" is an academic phrase, regional colloquialisms like "pop" used to describe bubbly soft drinks, and the non-academic word "raffle" could all constitute construct-irrelevant language. Since measurement errors can occur when an aspect of the test item is unclear or misinterpreted due to unfamiliarity with phrases, words, and colloquialisms, researchers advise that

standardized test results will have higher validity for ELLs if test developers consider ELL students when creating test items and scenarios (Kopriva, 2000; Kopriva, & Sexton, 1999; Kopriva et al., 2007).

DIF refers to different groups with equal skill or knowledge that do not have an equal chance of getting an item correct. A test item is described as having DIF if it favors a specific demographic group over another. If differences in the probability of getting an item correct are dependent on variables related to group membership, rather than true differences in knowledge of the subject being tested, then construct-irrelevant variance could potentially be the cause. Since researchers have found evidence that linguistic complexity may be hindering ELLs from making sense of the items and displaying their knowledge, it is posited that DIF may be present in those items with linguistic complexity.

It is important to emphasize that these comparisons for DIF must match students on their ability or skill level in order to compare their scores on a particular item. The two groups' overall distribution of scores can differ; however, that circumstance is not classified as DIF. A significant gap between the performance of ELLs and non-ELLs on standardized tests has existed for many years now, such that a majority of ELLs are not reaching the class-level standard for reading or math (Fry, 2007). There are many possible reasons for these score differences. For example, the majority of ELLs come from lower socio-economic status groups, a demographic factor that has been linked to lower test scores (Abedi & Lord, 2001) simply because ELLs may not have had the same chance to learn the material as non-ELLs. In other words, if ELLs have been taught in

separate classrooms from non-ELLs, learning experiences will differ. When true differences in group performance exist, it is referred to as *impact*–where unequal opportunities to learn result in true test score differences.

This dissertation addresses the concerns of psychometricians, test developers, educators and ELL researchers regarding the need for empirical research on ELL testing and the validity of their academic achievement measures. Linguistic features that significantly and adversely affect ELLs can vary from different years of testing and in different regions and states. Comparisons between the results from studies with different ELL samples are important in highlighting the interaction between linguistic features and the characteristics of this diverse group. Thus, this dissertation focuses on two samples of fifth grade non-ELLs and Spanish-speaking ELLs and their performances on two state science assessments from administrations in 2006 and 2007. I analyze the science tests' items DIF using two methods to determine whether any items have DIF that favors non-ELLs because of linguistic complexity, while controlling for the effect of variables that could account for some of the variance in ability scores. The control variables included are intended to have a clarifying effect on the relationship between those items exhibiting DIF and linguistic complexity. My dissertation also examines possible ameliorating non-linguistic aspects that could assist these students in making better sense of the items.

The following chapter provides a comprehensive overview of research regarding sources of DIF against ELLs and will describe the methods for analyzing DIF and their respective limitations.

Chapter II: Literature Review

For many decades, researchers have been concerned about the possibility that ELLs' test scores do not accurately reflect their school subject knowledge (Aguirre-Muñoz & Baker, 1998). These researchers have hypothesized both broad and specific features of test items that might contribute to underestimations of ELLs' knowledge on state exams. This section will compare findings from these studies and the theories surrounding item features that may be causing comprehension problems for ELLs on tests. Recently, ELLs have become a focal group of interest in the DIF analyses traditionally centered on ethnic minorities and females. This literature review will concentrate on relevant results of DIF analyses conducted with ELLs and the potential sources of measurement error discovered by item analyses.

Linguistic Complexity

Linguistic complexity is a broad term that can include, but is not limited to the following: a high number of words used in the item's question and multiple-choice response options, a high number of academic-style phrases (e.g., "in order to"), use of words that are not frequently used in conversations (e.g., "the following set of"), use of relative pronouns (e.g., "which", "whom"), use of passive voice (e.g., "was given"), and conditional clauses (e.g., "if", "could", "would"), and use of complex verb phrases (e.g., "have been going") (Abedi & Lord, 2001; Abedi, Hofstettler, Baker, & Lord, 2001; Brown, 1999). Studies that remove linguistic complexity in exam items have shown moderate increases in ELL scores compared to the original (Shaftel et al., 2006; Siegel, 2007). In some cases, non-ELLs, who are struggling readers, showed improvement as

well (Abedi & Lord, 2001). Rivera and Stansfield (2004) identified linguistic complexity in fourth and sixth grade science exams items then modified those items with the help of science teachers. Their research team and discovered that simplifying language had no effect on the item's difficulty level for native English speakers. These results imply that reducing the linguistic complexity of items may improve ELLs' test scores without compromising the validity of the results for other students. Rivera and Stansfield compared reading linguistically simplified exams to wearing needed correction eyeglass lenses and explained that, for those with inadequate vision, wearing prescription eyeglasses does not improve reading skill – it only improves basic comprehension. To avoid results confounding the knowledge area with English language reading proficiency levels, Riviera and Stansfield recommended that linguistic complexity be removed from test items to improve the resultant validity level for ELLs.

Due to being situated in mainstream U.S. culture, test items reflect the writing style, word choices and perceptions of the test writers. However, ELLs are not considered part of this homogeneously imagined group of test-takers. Even though items are assumed to contain measurement error (no test is perfect), for the most part, results are accepted as if the test contains scant measurement error. The connections these students make to words and phrases will depend on their familiarity, or exposure with those phrases and words – thus measurement error may be present for a sub-group of ELLs based on certain characteristics. Finding an academic word, like "however", that is difficult for some ELLs to understand, will not necessarily generalize to a larger group of ELLs. Another word in the same sentence may affect another proportion of the test-

takers. Natural language is seldom precise, thus interpretations are often contextual (Hunt & Agnoli, 1991). The subject area being written about often constraints the linguistic features such that modifying the language may come at the cost of sacrificing meaning. So, while it is important to measure and reduce linguistic complexity, it is important to keep in mind the probabilistic nature of language (Solano-Flores, 2008). Removing complex linguistic features from items alone may not necessarily account for a significant amount of measurement error for ELLs (Ockey, 2007) and may not be the best solution for improving comprehension. Predicting which parts of speech cause difficulties in comprehension for some ELLs and then removing those parts of speech from test items may not make a significant difference for other ELLs. Each ELL has his or her own unique combination of English skills and challenges. Reducing a compound sentence into two shorter sentences may increase test item comprehension for the ELL who has not yet become adept with compound sentences, but may not aid the ELL with stronger reading skills and better understanding of the science concept being assessed. For the latter, the original compound verb phrase may better describe the relationship between the actions consistent with how the lesson was taught in class.

These linguistic simplifications do not account for different socioeconomic groups having differing experiences and interpretations of the same sentence. For ELLs that are not part of the mainstream culture there are many sub-populations within the greater ELL group. An obvious example is how the experiences of children who are political refugees or migrant workers can be expected to differ from those of middle-class

non-ELLs. For this reason it is also important to define a specific sample and attempt to account for some of the sample's variability in order to generalize to that specific group.

While results vary from study to study and grade level, test item features showing improvement are word choice (for example idioms or multi-meaning words) and item length (linked to lower scores for ELLs in multiple studies) (Abedi, Lord, & Plummer, 1997; Abedi & Lord, 2001; Martiniello, 2008, 2009; Shaftel et al., 2006; Wolf & Leon, 2009). The main focus of the next section is the likelihood that removing these types of linguistic complexity successfully will positively affect ELLs' performance.

Aspects of Linguistic Complexity

Recent studies have used composite linguistic complexity scores to explain the DIF against ELLs in math and science items; however, a breakdown of item features contribute most to DIF by using academic or unfamiliar word choices (Martiniello, 2008; Martiniello, 2009; Shaftel, 2006; Wolf & Leon, 2009).

Several ELL researchers recommend the use of high frequency words in test items (Abedi, 2006; August, Carlo, Dressler, & Snow, 2005; Kopriva, 2000). *High frequency words* are words more commonly encountered in spoken conversations, as opposed to words often used exclusively in formal written pieces. Conversely, *low frequency words* can be overly academic, and may distract or confuse students from the test item's main objective. Also, *polysemous words*—ambiguous words that can have multiple meanings—such as "left" (which could be used as an adjective indicating a direction or the past tense of the verb "to leave") can also confuse ELLs when used on assessments. Specifically, using regression analysis with linguistic complexity as a predictor of math

item difficulty, Shaftel et al. (2006), found that polysemous words are problematic for ELLs in fourth grade.

Wolf and Leon (2009) measured the linguistic complexity of science and math tests for several grade levels, and found that the tests' easier items are more difficult for ELLs to answer correctly than non-ELL. These easier items are found to contain a significantly greater amount of academic vocabulary versus technical math or science words. Examples of problematic academic vocabulary are the terms "based on" and "substantial." Conversely, the more content difficult math and science items contained a much more technical vocabulary which they strongly believe was actively taught to ELLs, and thus more familiar. Familiarity provided them greater opportunity to make sense of the questions asked. The researchers believe that in the referred to "easier items," the use of academic language may have been employed to compensate for a perceived lack of depth and technical difficulty which resulted in making those items more difficult for ELLs than non-ELLs.

Each exam item must carefully include appropriate words and examples to avoid ELL miscomprehension. For example, Martinello (2008) interviewed Spanish speaking students with varying levels of English proficiency about their understanding of the items on the Massachusetts state math assessment. She found that most ELLs are unfamiliar with certain words considered common relating to the home- such as "chores," "rake," "weed" and "dust," while they are familiar with other common words related to school, like "pencil" and "notebook". They also had trouble with the words "owe," "identical," "likely," "unlikely," "certain," and "even." It is not easy to make assumptions regarding

ELLs' experiential knowledge, which is why test development will require item analyses from ELLs' viewpoints.

Martiniello (2009) found that the presence of linguistic complexity, specifically items that contain non-math related vocabulary, predicted DIF in favor of non-ELLs. When asked what they believed an item was asking, ELLs could not determine which noun certain adjectives referenced. Complex sentence structures combined with unknown words confused ELLs about what the item was asking (Martiniello, 2008). Also, ELLs, when seeing the word "one," always assumed it was for the amount or number '*1'* even when it was used as a pronoun. Students' feedback about linguistically modified items and their ability to understand those items better has been positive (Abedi & Lord, 2001; Siegel, 2007).

To aid the large majority of ELLs who are native Spanish speakers in making sense of test items, the use of Spanish-English cognates is encouraged (August et al., 2005; Carlo, et al., 2004). Spanish-English cognates are words that sound and look similar in structure and have the same meaning in both languages. For example, "combination" in English is a cognate to the word with the same meaning in Spanish, "combinación." This word was used on two items of the same fourth grade math test, and was understood by almost all of the ELLs (Martiniello, 2008). Approximately sixty-eight percent of the technical language or scientific words used in textbooks have Latin roots and could potentially make sense to Spanish speakers (Carlo, et al., 2004). In addition, teachers of Spanish-speaking ELLs have long known of the benefit of creating Spanish-English cognate awareness (August et al., 2005; Krippner, 1966; Nagy et al.,

1993; Williams, 2005). However, teaching ELLs to use cognates is only feasible for students whose native language has the same roots as English; e.g. French, German, Spanish, Italian and Spanish, but not for Cyrillic or east Asian languages.

While comparing Chinese and Spanish speaking ELLs in Canada, researchers found that Spanish cognates are easier for Spanish speaking ELLs (Chen et al., 2011). When confronted with Spanish cognates, Spanish speaking ELLs out performed Chinese speaking ELLs. In contrast, non-cognate vocabulary was equally difficult for both groups, and was associated with number of years living in Canada, which only explained a small amount of variance in score differences. Since instruction on how to recognize cognates is provided for ELLs during class lessons, it is possible that items containing cognates could counteract the effect of some linguistic complexity present in the item.

Still, English words that have the same root and meaning as Spanish words will not always be recognized as a cognate by some ELL students. Martiniello (2008) found that at least half of the twenty-four fourth-graders that she interviewed didn't know the meaning of the word "exactly," even though it is similar to "exactamente," a word with the same meaning in Spanish. Although most ELLs are able to understand the mathematical use of the word "combination" (a cognate of "combinación" in Spanish), one of the two math items involved combinations that the students couldn't figure out what was to be counted and combined when an item contained several unfamiliar words. This result could be due to illiteracy among many Spanish-speaking ELLs (Martiniello, 2008).

There are also false cognates—words that seem to mean the same thing in both languages because they sound the same. These could confuse ELLs, such as the word "firm", which looks like the Spanish word "(una) firma," meaning "a signature." Luykx et al. (2007) provide an example of confusion that resulted over the word "gaseous," a word that resembles "gaseosa," meaning soft drink in some Spanish speaking countries. Young et al. (2008) believe that use of Latin-based terms in wrong answer choice options may have caused ELLs to pick answers not based on whether they are right or wrong, but because they looked familiar. ELLs are also more likely to choose the multiple choice response containing unfamiliar academic looking words, possibly because they see this as a test ploy and believe the most likely answer choice contains the least comprehensible word (Abedi, 2010). Therefore, it is not surprising that studies have not always found academic or ambiguous words to be correlated with an ELLs' difficulty with an item (Abedi et al., 2005). Results can be inconsistent across tests and grade levels; for example, ELLs' difficulty interpreting math test items due to multi-meaning linguistic complexity in fourth grade was not reflected in third or eighth grade math tests (Shaftel et al., 2006).

Although low frequency words may not always be the main influence in reading comprehension, it has been found to be a significant contributing factor in difficulty passing test items for many decades, now (Graves et al., 1980). When constructing test items, use of high frequency words over less frequently encountered words should lessen the confusion and cognitive reading load, and thus decrease the chance of misinterpreting the item, while increasing the speed of reading so that the student can focus their

attention on the task and demonstrate their skill in the subject area (Kopriva, 2000). From a longer set of linguistic simplification suggestions, Kopriva (2000) also provides one example of a math item involving pizza and hamburger toppings where she advises test developers to choose more familiar words like "pepperoni" rather than "anchovies," and "tomato" rather than "pickles."

Some studies have concentrated on glossaries as a possible solution to the inclusion of unfamiliar words. Providing a glossary of terms that are unrelated to the construct being measured is the most successful accommodation for ELLs (Abedi, 2009; Abedi, Courtney, & Leon, 2003). However, to use this accommodation requires ELLs to be trained to use glossaries and, therefore, will require extra time for searching for these words during the testing process (Wolf et al., 2009). The problem for developers and those working on this accommodation is they cannot assume which words any ELLs would or would not know, and therefore, as with development of the test item, developers would need to ask ELLs directly. Subtle linguistic variation in items with the same content could potentially lead to very different interpretations and score results. This can be seen from generalizability studies that have focused on dialectic differences. In particular, studies have shown how, even with the students who speak a certain dialect, translating an item into a second dialect can result in score improvements for some students and score declines for other students (Solano-Flores & Li, 2006; Solano-Flores, Lara, Sexton, & Navarrete, 2001). Since the dialects of ELLs vary even within the same language, awareness of which words to include in a glossary does not solve the problem. Multiple glossary translations would be needed for each dialect of each language.

This solution is somewhat complex since glossary definitions must also protect against revealing the answer to an item, or inadvertently giving clues about required technical language. Mainly, it has been suggested that the best solution is to avoid all high level academic English terminology on tests so that English reading proficiency levels and math and science assessment scores will not be confounded (Abedi & Gándara, 2006; Kopriva, 2000; Mahon, 2006). Still, researchers question what specific words are considered high level academic words that ELLs would not be familiar with when attempting to generalize to a highly heterogeneous group (Solano-Flores, 2012).

Item length is another aspect of linguistic complexity that has been linked with a lower proportion of correct responses on an item for ELLs**.** Ways of measuring item length depends on which study is being examined, from number of lines in a question (Abedi et al., 1997) to number of words in the item or the response options (Wolf & Leon, 2009). The number of words or sentences required to be read in order to answer an item have been identified as problematic for ELLs, such that longer test items are more difficult than shorter items on a state math assessment at three different grade levels for ELLs (Shaftel et al., 2006). In another study of eighth grade math items, non-ELL students did significantly better than ELLs on items with longer explanations (Abedi & Lord, 2001). Non-ELLs outperformed ELLs on NAEP 8th grade multiple choice math items containing three sentences or more, and item response choices a sentence length or longer (Abedi et al., 1997).

Not every lengthy test item has been shown to adversely affect ELLs, however, studies using a wide variety of data have found similar results. On a fourth grade math

test with twenty-nine multiple choice items and ten constructed response items, Martiniello's (2008) content analysis revealed that two multiple-choice items with twelve to sixteen word sentences exhibited high DIF against ELLs. Wolf and Leon (2009) undertook one of the few ELL studies to use data from both science and math assessments and over multiple grades, from fourth to eighth grade. Over the eleven different assessments they found a pattern, such that sentence length correlated with DIF against ELLs for both math and science items that are considered easy in their subject content area and level.

However, Ockey's (2007) study was unable to find more than one item with DIF disfavoring ELLs. Ockey reanalyzed the same data set and was unable to attain the same results as Abedi and Lord (2001). This evidence suggests that simplifying the language may not result in a marked improvement in ELL scores. Making changes to the linguistic complexity, therefore, may sometimes lead to statistically reliable differences, but not practically significant ones.

It is possible that sampling and score matching issues may explain why DIF was not identified in the items that contained the most linguistic complexity in Ockey's study. Of 1174 students, only 372 mostly low socioeconomic status ELLs were represented, meaning that the sample consisted of a majority of non-ELLs, thus perhaps an insufficient number of ELLs' scores to match. Also, some of the non-ELLs had previously been reclassified from ELL status, but were not examined separately, which could have contaminated the data. The author did not detail whether he examined the internal consistency of the test separately for each of the groups to determine whether

multidimensionality varied by English proficiency. Methodological changes like the ones mentioned above could lead to discovering different numbers of items disfavoring ELLs and multidimensionality in the items.

At present, no known study has yet experimentally manipulated item length. Therefore, the actual correlation of item length with ELLs' item scores in the literature is uncertain. However, as reviewed above, this aspect of linguistic complexity has been identified as adversely affecting ELLs' item comprehension. Item length should continue to be investigated in future studies, and perhaps in more studies that include constructed response items.

Language Scaffolds

While measuring the linguistically complex features of an item, some studies have also tried to determine how much an item relies on aspects other than language as a way to gauge the potential influence of the non-verbal features of the items (Abedi et al., 1997; Martiniello, 2009; Wolf & Leon, 2009). For example, the use of pictures, diagrams, graphs or tables are thought to relieve some of the cognitive load attributed to linguistic complexity.

Shaw (1997) studied five high school science classes where he examined the ELL assessment process from several different perspectives. The results of these students' scores revealed that the items that relied most on language are positively correlated with English proficiency levels, while English proficiency levels are not correlated with items that are purely calculation items or those containing tables or graphs. Similarly, on the math assessment of the Iowa Test of Basic Skills (ITBS) for grades 6 and 8, test items

that relied more on language—referred to as high language load or high language demand—are more difficult for ELLs than non-ELLs (Abedi et al., 2005). Finally, on both science and math assessments an item's reliance on language (as rated by expert judges) was the strongest predictor of DIF against ELLs, such that those items without visual aids are more likely to favor non-ELLs over ELLs of equal ability (Wolf & Leon, 2009). Thus, inclusion of visual aids on assessment items could potentially increase the scores of ELLs to reflect their true knowledge without compromising the validity of the assessment

In a science classroom assessment study by Siegel (2007), ELLs commented that they appreciated the inclusion of pictorial representations of the actions being described within the items. Martiniello (2009) takes the issue of visual scaffolding a significant step further by describing a certain type of visual aids items need. These visual aids are classified as being "schematic" rather than pictorial, meaning that the illustrations non-verbally tell a story or describe a situation, rather than, for example, representing an object without background characteristics or motion. In her study, linguistic complexity found on fourth grade math items with DIF in favor of non-ELLs was greater when there are no figures or equations in the item to help make a connection to the terms and question being asked (Martiniello, 2009).

Research is currently being undertaken to create carefully designed illustrations to accompany science items that will show as closely as possible what the text in scenarios are meant to communicate (Solano-Flores & Wang, 2011). The conceptual framework for creating these illustrations provides categories of details of the development and

design, such as how text is included in the illustration and if time, space or motion is represented. The conceptual framework for the design and use of illustrations has already shown itself to be sensitive to cultural differences. There is a great deal of potential for them to support ELLs in their test item understanding. The prescriptive procedure for creating illustrations that could help ELLs to make sense of an item is a complicated process still being researched (Solano-Flores & Wang, 2011), but results have already shown differences across cultures, such that students prefer illustrations from their own culture (Wang & Solano-Flores, 2011). Therefore, the inclusion of visual aids with test items becomes is only one consideration. Visual aids share with the non-textual elements of linguistic complexity in that they impact the comprehension of the linguistic features of the items. The current study will attend to the potential influence of visual aids to increase the coherence of the item.

ELL Characteristics

There are several other factors that can shape the effectiveness with which linguistic complexity can be identified. First, as has been mentioned previously, mean test scores for ELLs are almost always well below those of non-ELLs. Stoneberg (2004) points out that when comparing groups using DIF analyses, if the focal group's ability scores are a standard deviation or more lower than the reference group's, as in this case where the ELLs' expected performance is already quite low, it can be harder to find differences between expected and actual scores. Limited information is available for extremely low or extremely high scores in empirical analyses, as extreme scores do not provide enough variability to specify the knowledge of the students in these groups.

Since the ELLs scores are quite low, there may not be enough information to make final pronouncements on their performance and thus their true science skill proficiency. Examining scores of all students for whom English is not their native language instead of just those who are labeled as being in ELL/bilingual programs may increase the score variability. Another solution proposed to increase the variability in scores is to include more items at a variety of difficulty levels. Having more items on a test on the same content area but asking the questions in a variety of ways may better pinpoint knowledge levels.

Some of the DIF results that show items favor non-ELLs may actually be due to *impact*—differences in knowledge levels which can arise due to contextual factors such as lack of opportunity to learn or divergent educational experiences (Abedi & Herman, 2010). Lack of opportunity can result from insufficient practice speaking and reading the scientific terminology to gain understanding of the concepts or not being taught the scientific terminology at all. A pronounced example of divergent educational experiences is comparing the science classrooms of different socio-economic groups. Schools with the most resources are more likely to have access to expensive technological equipment like microscopes and greater opportunities for access to in-situ learning and field trips. When considering the likely reason for DIF, a pattern may emerge such that items that cover a specific content may favor non-ELLs. Lack of opportunity to learn some content areas in math, especially math data analysis and probability, may be the cause of DIF. Martiniello (2008) identified high DIF against ELLs for data analysis and probability content items. While fewer DIF studies with

ELLs have focused on science content, when Ercikan, Gierl, Mcreith, Puhan, & Koh (2004) compared science tests translated into French they reported that almost all items for the content strand of "science technology" exhibited DIF favoring native English speakers (Ercikan et al., 2004). These results point to the fact that ELLs, many of whom are from low socio-economic backgrounds, have not had equal an opportunity to learn about scientific technology in school or are differentially exposed to technology outside of school.

Cultural and Experiential Issues

The cultural causes of measurement errors contained in test items are more difficult to uncover but are an important area of recent research in ELL testing (Wu & Ercikan, 2006; Zenisky, Hambleton & Robin, 2004). ELLs bring with them their own culture's experiences and knowledge while many aspects of mainstream U.S. culture remain unfamiliar to them, including some that they will discover for the first time in school, for example a new style of tests. Students are encoding information in their own environment, based on their age, school and home experiences, their cultural background, their linguistic background, their subject area knowledge, their previous experience with tests, and their feelings about the test question and the test as a whole. For this reason, concentrating on a particular sub-set of ELLs can improve the generalizability of a study's results.

In their classroom assessment of science, Luykx et al. (2007) note that science itself has cultural components that cannot be studied separately from the linguistic components of items, and that test developers can clarify and remove any cultural biases

they are able to identify, but not all cultural components can be removed from the test because the test itself is a cultural device. Test styles can be quite sophisticated, and can take many years to learn because of how a test is constructed or the type of response that is expected is never explicitly explained even after students go through the experience (Scarcella, 2003).

Test developers make assumptions about what test-takers would be familiar with in terms of test format, interrogative style, and types of situations used in item examples. This is in part due to the fact that test language and style are imbedded in culture. Therefore, a test, with its particular form and style is an artifact of the culture, one that generally reflects the cultural norms and experiences of the middle to higher echelons of the society, and is likely to be unfamiliar to ELLs, who tend to be from lower socioeconomic groups. Duran (2008) and Solano-Flores (2008) explain how the forms of questions asked on tests evolve from a culture that students learn early on in school. Moreover, an understanding of the testing process is built upon over the years or during a student's academic career. For students who arrive in the school system later on in their education, the style of question asked on a test is entirely new. Since the discursive style of exam items and the item question and response formats are often introduced during ELLs' first assessment experiences, it will likely take some time for them to become accustomed to each particular aspect of the U.S. testing culture (Solano-Flores, 2006).

Likewise, issues regarding gender and race are also affected by the cultural facets of tests. Due to the paucity of empirical research on ELLs' difficulties due to linguistic complexity on mathematics and science items, not much attention has been given to

certain item attributes, such as item format and structure. DIF studies comparing Caucasians to minority ethnic groups have found that open-ended—also called constructed response—test items favor minorities while multiple choice items favor Caucasians (Taylor & Lee, 2011). In gender comparison DIF studies, constructed response items are differentially easier for females while multiple choice items are differentially easier males (Taylor & Lee, 2012; Zenisky, Hambleton, & Robin, 2004). There is growing evidence to suggest that multiple choice items favor non-ELLs and constructed response items favor ELLs (Ilich, 2011; Smith, 2009). Multiple choice items may not provide much information about students' breadth of knowledge in any topic area, therefore, they could be underestimating ELLs' knowledge. For example, the results from the multiple choice questions were lower for ELLs, particularly for the beginner ELLs, than for non-ELLs (Kopriva et al., 2007).

The majority of the ELL studies mentioned in this paper use mainly, or in some cases exclusively, multiple choice test items. A greater number of constructed response items will need to be used on the tests for DIF researchers to determine how they function for ELLs. For all items there are still unknown aspects of the formatting that could affect ELLs' comprehension. In classroom studies, while scoring ELLs' constructed responses and during interviews, researchers uncovered confusion due to item formatting inconsistencies in the way the questions are presented, such as putting several questions or excessive information into the stem of the item prior to asking the question or using bold or italics fonts (Lukyx et al., 2007; Siegel, 2007). Increasing the number of constructed response items on tests may seem counter-intuitive, since these items require

more writing from a group of students who are not entirely proficient in English writing. Nevertheless, when given the chance to provide explanations of what they know on constructed response items the students' knowledge levels might be more precisely measured.

Socioeconomic status. The types of experiences that a child is exposed to may depend to a large degree on family income level. The community and lifestyle of all children, not only ELLs, can vary greatly and affects the age at which a child would be exposed to certain types of information about the world outside their homes, for example the types of activities they participate in ranging from hiking, fishing and camping to skiing, to the variety of foods they have encountered.

Many studies suffer from lack of control over the size of demographic groups to allow comparisons within or between groups, others simply do not have access to demographic information that would shed light on their results. Thus, matching or weighting scores in a DIF analysis for socioeconomic status (SES) could be an important factor in comparing the results of linguistic modifications and uncovering other factors that could cause score variance, especially for the studies that conclude certain learning strands or subject content are the cause of DIF. However SES is rarely used as a matching control variable or covariate, or mentioned in DIF studies. Abedi and Lord (2001) and Krashen and Brown (2005) found that SES was correlated with math scores such that the low SES ELLs did not do as well as the high SES ELLs. Additionally, English proficiency level was strongly correlated with SES in studies where eighty-two percent of ELLs are low SES (Martiniello, 2008, 2009).

Since the majority of ELLs in the nation are low-SES, as measured by their eligibility for a reduced or free lunch subsidy, it may be more appropriate to take SES into account or to compare low-SES non-ELLs with low-SES ELLs, as comparing unmatched groups could affect the magnitude of the DIF found or could inflate the number of items with DIF.

One way to examine the influence of SES on item performance is to identify items that assume certain background knowledge. Solano-Flores and Li (2009a) asked low SES students to explain their understanding of a NAEP math item, asking how much lunch money a child needs. They asked students to talk through the logic of how they answered the question. Interpretation of what was being stated in the item varied with SES. For the low SES students, having experienced scarcity of adequate funds detracted from their ability to answer the word problem correctly causing them to misread and misunderstand the information in the question.

Abedi and Lord (2001) note that having a high SES can compensate for ELL status, such that high SES ELLs can do as well on standardized tests as low SES non-ELLs. For over a decade, ELLs' opportunity to learn has been observed and reviewed and found wanting (Abedi et al., 2006; Abedi & Herman, 2010; Aguirre-Muñoz, Parks, & Benner, 2006; Minicucci & Olsen, 1992). Research that compares high SES students to low SES students has revealed poorer educational outcomes can occur due to: lack parental involvement, lower parental education level, less stimulation provided as an infant, less school resources, lack of the availability of advanced placement courses in high school and overall differences in content covered in class lessons (Griffin, Allen, &

Kimura-Walsh, 2007; Kuhl, 2011; Schmidt, Cogan, McKnight, 2011). Although having low SES does not guarantee that these negative circumstances will definitely occur, they are considered a powerful trend that affects students' performance, so that the lack of opportunity to learn for low SES students has been strongly linked to adverse educational outcomes. Therefore, such influences can be difficult to pinpoint without extensive knowledge about the backgrounds of the sample it is still important to consider SES and other background characteristics that could aid in identifying aspects of items that would be problematic for these students (Betts et al., 2009; Kopriva, Wiley, & Emick, 2007; Mahon, 2006; Solano-Flores, 2008).

Reading comprehension. Every language has different types of words and conventions which come about from the culture. How often a word is used, or its frequency, is also dependent on the culture. For example, in English the word "inure" is not frequently used, however this is not the case in some other languages, for example in Serbo-Croatian-Bosnian. The word for "inure" has three synonyms to conveying its meaning in different contexts that are frequently used in both everyday speech and writing in the Serbo-Croatian-Bosnian language. The importance of such word use is that it sheds light on a culture, in that it can explain why a cultural outsider would have a different interpretation of the same situation, emotion or behavior from a given text. It can take time for students new to English to become familiar with the frequency and grasp the flexibility of English word usage.

Likewise, word arrangement in sentences also has an impact on comprehension. The ways in which a sentence is written and a question is communicated will change the

meaning of what is read depending on who is reading the sentence. In addition, the placement of words in a sentence, and the inclusion or exclusion of words can make a huge difference in how meaning is constructed. Thus, the conventions of writing and speaking are also important to examine when analyzing test items during test construction, since exams generally use more formal English while English speech is informal. This has led to test items with format, syntax and cultural references that skew toward the unfamiliar for a great majority of ELLs who also have not yet become accustomed to the testing process.

Determining why ELLs may have difficulty understanding or identifying which words are not being understood could also be due to an ELL lacking experience with a situation presented in an item and thus a potential source of construct irrelevant variance. Cultural experience often interacts with other factors, such as the previously mentioned SES. Specifically, SES can determine which words are most frequently encountered in daily life. While test items which feature rarely-experienced scenarios arguably are appropriate on certain types of exams like the Graduate Record Exam (GRE), with grade school aged ELLs a wide range of what seems like everyday scenarios to non-ELLs may be totally unfamiliar and unknown to some ELLs.

Speaking English primarily at home has also been linked to higher test scores, especially with respect to reading comprehension skills (Klein & Jimerson, 2005). This may be why the number of years spent in the U.S. may not always strongly correlate with higher English proficiency scores. SES in conjunction with unmeasured socio-cultural and socio-linguistic factors could have an effect on speed of attaining English proficiency

(Angoff, 1989). The prevalence of community support for the ELLs' cultural or language background or lack of these resources can also affect ELLs' speed in acquiring English reading skills. In parts of the U.S. with large Latin American communities, Spanish speaking families may frequent Spanish-only supermarkets and restaurants, and exclusively use Spanish television stations and radio channels. Therefore, contrary to the findings of Wang, Park and Lee (2006) showing gains in reading comprehension due to language similarities, Betts et al. (2009) found that when comparing Somali students and Spanish-speaking students who had been in the United States the same length of time, Somali students increased their English fluency at a faster rate than the Spanish-speakers, most likely due to the lack of community resources available in the Somali language. Caution should be taken, however, with these results as they do not address the benefits of bilingualism and multiculturalism (Wong & Fillmore, 1991). In one noted study, children of Chinese and Korean immigrants who maintained their familial culture and language scholastically outperformed their peers who became immersed in American culture (Lee, 2002).

One such relevant example that precisely displays the complexity of such cultural influences was found in Martiniello's study (2008) where an item used an unfamiliar word "coupon" incorporated in the colloquial phrase "coupon for $1.00 off". ELLs could not determine what the word "off" meant in the context of the item because the idiomatic phrasing terminology ("$1.00 off") can only to be encountered after actually acquiring and using coupons, and Spanish-speaking ELLs could not recognize the context in English. Researchers claim that many ELLs' English proficiency levels are actually still

miscategorized (Abedi, 2008; Kopriva et al., 2007). Using English proficiency level data, such that ELLs are said to be at beginning, intermediate or advanced levels of English proficiency, may not be helpful as this designation does not provide enough information about ELLs' specific knowledge. In part, this type of categorization is difficult to assess with the amount of variability present in cases of different usage: casual English usage, formal English writing, English listening, and simply the variance in vocabulary, such as "does a student know this specific term" or "do they know the multiple uses for this term" arising from the variance in experiences (Scarcella, 2003; Solano-Flores, 2008). When and how English is used in school communities will ultimately lead to differences in reading test scores (Solano-Flores, 2008). These scores do not necessarily predict familiarity with, for example use of conditional tense or relative pronouns. Solano-Flores (2006) refers to this issue as linguistic alignment, notably that the type of English language complexity that ELLs are familiar with and the language used on the test should be matched in order for the test to be appropriate.

Martiniello (2008) points out that high frequency words found in the math items she analyzed were terms associated with the school environment and, with one exception (the word "vacuum"), the words that referenced the home, such as "dust" and "weed," were unknown to most of the ELLs in her sample, even though many of these words are considered to be high-frequency words by other researchers. Although many of the ELLs were familiar with the word "vacuum," which provided a context for one question, the concepts of "inside chores" and "outside chores" could reference experiences which may not apply to all cultures or fourth grade children's experiences, especially if in their

cultural milieu responsibilities for keeping up the home are not given to children, or they live in an apartment building where there are no outside chores or gardening opportunities.

The evidence provided in this section details why researchers advocate for the consideration of the sociocultural backgrounds of ELLs during test development processes, particularly in the piloting of items, preferably with interviews to confirm the test item has the meaning intended and to demonstrate validity evidence that the item evokes the cognitive processes that the construct being measured specifies (Sireci, Han, & Wells, 2008; Solano-Flores & Nelson-Barber, 2001; Solano-Flores & Trumbull, 2003). Data regarding the influence of experiential factors presented in the items needs to be compiled to potentially explain why ELLs who demonstrate high science skill levels do not pass test items at the same rate as equally skilled non-ELLs.

Errors in measurement of ELLs' science knowledge may stem from ELL background characteristics such as culture and socioeconomic status that complicate the identification of linguistic complexity. Ercikan (2002) concluded that about half of the DIF in science items on the Trends in International Mathematics and Science Study (TIMSS) that disfavored ELLs was due to translation adaptation problems and opportunity to learn certain content strands. It is likely that many of the causes of DIF disfavoring ELLs could be tied to the huge amount of diversity and complexity in the background characteristics of ELLs that are incalculable or unavailable to researchers. Figuring out and removing language that causes difficulty in comprehension for some ELLs is unlikely to resolve the problem for all ELL students (Solano-Flores, 2008).

The long-range consequences of not addressing test score validity can affect student interest and performance in the school subject and, therefore, limit career goals. Testing, a situation imbedded in culture, is often perceived as stressful because it makes a judgment about the competency of the test-takers. Even parents' attitudes and knowledge that is passed down affect how students will view and feel about the test. Stereotype threat is a good example of how low test scores may influence negative perceptions, and then continue to affect student progress and outcomes, particularly for young students whose self-images are being formed (Steele & Aronson, 1995). Low expectations and stereotyping can negatively affect student performance, and the societal ramifications of these behaviors could potentially lead to most of these students underperforming continuously over the years at school which can lead to lower career goals and presumably missed opportunities for higher paid jobs (Friend & Degen, 2007; Griffin, Allen, & Kimura-Walsh, 2007).

Given all of these language and cultural issues, it is likely that ELL test scores may not adequately reflect ELLs' science knowledge. For the reasons provided above, it is important to avoid the confounding of language skills, culture or socioeconomic status with science proficiency. Studies regarding testing of ELLs must measure the contribution of unique sources of item variance to uncover the degree of measurement error due to those variables.

Summation

The research studies mentioned in the previous section have demonstrated that ELLs may be losing points on exams due to the inability to make sense of linguistically

complex items that include variables outside the subject content that is being assessed. Kopriva and Sexton (1999) call "item overload" the situation that occurs when ELLs expend most of their time and mental energy decoding the language and making sense of the exam item, rather than responding to the question being asked (see also, Beal, Adams, & Cohen, 2010). However, it is difficult to specifically isolate those linguistically complex factors in every instance that create the most problems for ELLs in decoding an item's meaning because several types of linguistic complexity coexist on any given item. The scarcity of definitive conclusions for prescriptive recommendations for this multi-cultural and distinctive population of students has paved the way for promising areas for further research.

It is evident that gaps remain after reviewing the existing research on ELLs, some of which this dissertation will attempt to address. First, few DIF studies for ELLs have been carried out using the presence of linguistic complexity as a variable. Several correlational studies have inconsistently or inconclusively determined linguistic complexity to be the underlying cause of ELLs scoring lower on test items than non-ELLs. Stronger statistical evidence is required of a repeated pattern relating linguistic complexity to ELLs' test score deflation. Second, many studies about ELLs have focused on math assessments. Science assessments have been shown to involve more academic language (Wolf & Leon, 2009), so more research that focuses on science assessments could shed light on what types of linguistically complexity are present in science items and whether the problems in science items are the same as on math assessments. While there may be a greater linguistic burden and more lesson-specific

technical language in science tests, much of the technical terminology contains Latin roots, which are Spanish cognates and could assist Spanish-speaking ELLs. More research is needed that approaches linguistic complexity in relation to other factors present in the items. This dissertation constrains the ELL sample to Spanish-speaking ELLs because they are the largest non-English speaking group in the nation. It also takes into consideration ELL characteristics, their SES and reading level scores, and item characteristics, the presences of cognates and visual aids, that could clarify the strength of the relationship between DIF and linguistic complexity.

Differential Item Functioning

After performing the DIF analysis between students of equal ability, item level results indicate whether a particular exam item is disproportionately difficult for the focal group at each ability level, such that when DIF is present, the two equivalent groups do not have an equal chance of correctly responding to an item on the test. When it's been shown that there is DIF in suspected item (or items), it is said that the item "favors" one demographic group over the other.

DIF assumes that the differential performance of the focal group on the suspect item is not solely related to the construct being assessed, rather an unaccounted for construct is systematically affecting one of the group's scores. The resulting differences in between-group scores for an item exhibiting DIF, therefore, are not entirely due to different content-related ability levels since the unaccounted for construct is affecting one group's scores or differentially affecting both of the groups' scores. Only when the groups of students are matched on ability levels can DIF be examined. Observed scores

for the two groups are compared to their expected scores at each ability level. In DIF analyses, DIF is examined at each ability level by comparing focal and reference group members, where the probability of getting the item correct should be the same for both groups. If there is DIF found due to bias, these items can lead to lower overall test scores, such that one group is not performing as expected on the biased item based on their ability estimates. Therefore, the item difficulty measures, essentially a natural log transformation of the probability of passing the item at each ability level for the two groups, should be the same for each item under conditions of no DIF. Most DIF procedures use the total test score as the measure of ability, however if a high proportion of the items' measures are confounded by another dimension or factor the calculated ability estimates may not be reliable (Douglas, Roussos & Stout, 1995).

A central feature of DIF analyses is comparison of individuals with the same ability levels. The number of ability levels being matched for the two groups varies from method to method and can generally be divided into two categories, "thick" and "thin" matching. Thick matching means that there are fewer ability levels for comparison than for thin matching. For thin matching, all score levels an attempt is made to include from 0 to the total test score as the ability levels. Each DIF analysis method has its own measures of effect size, to avoid statistically significant results that might have little practical value.

It is important to note that finding an item has significant DIF against one of the groups is not sufficient to make the statement that the item is biased against that group. When determining a test is biased, one must examine whether the DIF is balanced with

some items in favor of one group and some in favor of another. One or a few items with DIF will not necessarily invalidate the test or test score. Sometimes, an item displays DIF due to actual differences in the groups' knowledge, e.g., if one group was not taught the material so that group scores lower on the item, which is referred to as impact. Only where the difference between the probabilities of each group passing the item is caused by construct-irrelevant factors can an item with DIF be viewed as potentially biased. One limitation of DIF studies is that results are not always able to reveal a clear cause of DIF, interaction effects between groups and items. The ELL population is a particularly difficult group to make predictions about because there are so many different ways they can vary, due to the presence of multiple cultures, SES differences, widely varying levels of knowledge in speaking or reading English, and differences in experience. Therefore, although some experimental control can be achieved in choosing only one language group or one country of origin, experiential factors will still vary due to the reality of there being several sub-groups of ELLs within any language sample.

In what follows, four commonly used methods for detecting DIF are reviewed: the Mantel-Haenszel, Rasch model, Simultaneous Item Bias Analysis, and logistic regression. Each of these methods has a unique procedure for calculating DIF and, therefore, may identify different items as having DIF. Thus, using more than one method is advised (Ackerman, 1992; Camilli & Shepard, 1994). After explaining the differences between the methods' approaches and assumptions, the reasons that SIBTEST and logistic regression are more suitable for this dissertation are provided.

Mantel-Haenszel. First conceived of in 1959 to compare the probability of contracting a disease and adopted for use with DIF as described by Holland and Thayer in 1988, Mantel-Haenszel (MH) is a non-iterative contingency tables method that uses a chi-square statistic ($\chi 2$) to determine significance of DIF (Holland & Thayer, 1988; Linacre & Wright, 1989). The MH alpha (α_{MH}) is calculated by comparing the item pass and fail rate for both focal group and reference group members at each test-score level then taking an average of the absolute value of the differences. For a dichotomous item, at each score level there is a 2x2 contingency table showing passing and failing on that item for each of the groups; the Generalized Mantel-Haenszel (Roussos, Schnipke, & Pashley, 1999) is used for polytomous items and the contingency table is 2 x *n*, where *n* will equal the highest number of points possible for that item.

Often five ability levels are established for Mantel-Haenszel comparisons but there can be as many levels (*k*) as needed (Donoghue & Allen, 1993). Generally the ability levels are taken from the total score on the test. The reference and focal groups are matched within score bands, such that each ability level has its own contingency table (see below MH Contingency Table 1). A similar contingency table is created for all ability levels.

Table 1

MH Contingency Table for One Ability Level

Group	Right (1)	Wrong (0)	Score level Totals
Reference (r)	**r1**	**r0**	Nr
Focal (f)	**f1**	**f0**	Nf
Group Totals	c1	c0	T

The odds of getting the item correct for the reference group (r) and the focal group (f) is calculated at each ability level to create the α_{MH} statistic using the components of the contingency table where c0 is the combined odds of an incorrect response and c1 is the combined odds of a correct response, and T is above.

For this equation, the null hypothesis should be equal to 1[2] , meaning that the odds of passing the item for the two groups at each score level are the same. This parameter α_{MH} is referred to as the common odds ratio. The null hypothesis can be written:

$$\frac{R_f}{W_f} = \frac{R_r}{W_r} \tag{1}$$

R is odds of getting the item correct, W is the odds of getting the item incorrect. *f* refers to the focal group and *r* to the reference group for both R and W.

The α_{MH} test statistic measures of the difference in performance between the two groups summed across all score bands for each item.

---

[2] Technically, alpha can reach infinity.

$$\alpha_{MH} = \frac{\sum \frac{r1 \cdot f0}{T}}{\sum \frac{f1 \cdot r0}{T}} \qquad (2)$$

` The α_{MH} is then transformed and put on a scale from negative to positive, so that the scale goes from negative numbers to positive numbers with a mid-point of 0, called the Δ_{MH} :

Δ_{MH} = -2.35 ln (α_{MH}) or Δ_{MH} = - (4 / 1.7) ln (α_{MH}) (3)

The statistical significance of the difference between the groups is interpreted as a chi-square statistic[3].

Holland and Thayer (1988) provided measures of effect size for (Δ_{MH}) to indicate the severity of the difference. The positive or negative sign indicates which group the item favors and then the Δ_{MH} indicates whether the DIF is negligible (if less than 1.0), moderate (if greater than 1.0 and less than 1.5), or large (if equal to or greater than 1.5).

A successful process has been found for non-uniform DIF detection using a MH procedure that requires the sample to be split into high and low ability and the MH method used on the two sub-sets of data separately (Mazor, Clauser, & Hambleton 1993).

---

[3] x^2_{MH} = (∑ (r1) –∑ (nr·c1 / T(r1)))2 / ∑(nr·nf·c1·c0 / T·T (T-1)) (r1))

Rasch model. The Rasch model is an IRT model developed by George Rasch in 1960 is the least complicated IRT model with one item parameter, difficulty. IRT models separately estimate the item parameters for each of the groups being examined. One main difference between MH and the Rasch model is that expected scores in IRT models are not dependent on the sample and IRT models are iterative—the tentative estimates of item difficulty and person ability are compared to a model of fit, and then the model is compared to the data until there is convergence. The model uses a scale to estimates people's abilities separately from the difficulty measures each item. Therefore, the person ability estimates are not dependent on the specific sample of items used, and likewise the item difficulty estimates are not dependent on the scores of the people who took the test. Fit statistics, "infit" and "outfit" provide information at the mean and the tails of the distribution regarding how well the results of the sample of people and items match the predicted Rasch model.

Similar to MH, the Rasch model calculates and compares the two groups' odds of a correct response on the suspect item in establishing that DIF is present. The Rasch approach compares the scores for reference and focal groups on the suspect item at each ability level. The probability of getting the item correct increases as the ability of the student increases, and the Rasch model can match the ability of the student to the likelihood of passing that item based on the item's difficulty parameter.

If the groups of interest do not have ability distributions that are close in proximity, for example one group has few if any high scorers while the other group has few if any low scorers on the test, then there could be problems with the ability estimates

and with the matching of scores for the tail ends of the distribution since there is not enough data for either group at some levels of ability. For IRT models, the item or test characteristics can be separated from the subjects' characteristics, so subjects' ability can be generalized to other tests of the same construct where the test parameter is determinable. The test's target ability level makes a difference in how a student scores on the test, therefore, knowing the difficulty level of the test and the student's ability level in that domain provides a score level for that test. IRT addresses this phenomenon and generates theta estimates for examinees that are independent of the difficult of the test.

Thus, for the Rasch model (Linacre & Wright, 1989) there are two initial measures to be concerned with—the ability (β) of the test-takers and the difficulty (δ) of each item. In this model, the ability estimates of the test-takers are grouped by score levels. Then, the reference group (R) estimate is compared:

$$\beta - \delta = \ln\left(\frac{p_{R1}}{p_{R0}}\right) \tag{4}$$

to the focal group (F) estimate:

$$\beta - \delta = \ln\left(\frac{p_{F1}}{p_{F0}}\right) \tag{5}$$

The formula takes the natural log (ln) of the proportion of those getting the item correct (P1) divided by the proportion of those getting the item incorrect (P0).

Then, the Rasch difficulty for the reference group members who gave the correct answer is compared to the Rasch difficulty for the focal group:

$$\delta_F - \delta_R = \ln\left(\frac{P_{R1}}{P_{R0}}\right) - \ln\left(\frac{P_{F1}}{P_{F0}}\right) = \ln(a_k) \quad (6)$$

The resulting α taken by performing this calculation at all (α_k) levels of performance provides the "estimate of the difference in performance in the two groups" (Linacre & Wright, 1989).

To determine whether the difference between the two groups is significant the following formula is used:

$$\tau = \frac{(\delta_F - \delta_R)}{SE(\delta_F - \delta_R)} \quad (7)$$

Here the τ is calculated by taking the difference between the item difficulty estimates of the groups, with the standard error (SE) of the difference in the denominator.

According to Linacre and Wright (1989) the Rasch model and the Mantel-Haenszel statistic have similar general assumption regarding the probability of passing an item at each ability level for the two groups, therefore can be used interchangeably. However, they believe the Rasch model is the superior statistical model due to it being an IRT model that separates ability estimates from item difficulty estimates and therefore calculates the standard error more effectively.

For the estimates of effect size, DIF contrast statistics of 0.42 or less are indicative of no DIF, beta statistics of 0.43 to 0.63 are indicative of slight to moderate DIF and DIF contrast statistics of 0.64 are indicative of moderate to high DIF (Linacre & Wright, 1989).

Logistic regression. **S**everal researchers' recommended statistical approach to take in DIF studies is logistic regression (LR), to calculate DIF (Clauser & Mazor, 1998; Swaminathan & Rogers, 1990; Zumbo, 1999). This method is a probabilistic odds ratio approach that uses the natural log scaling technique to test whether group membership will vary the probability of a correct answer. The total score, the demographically defined groups, and the interaction between these two are the variables in a regression analysis for predicting the probability of a correct response for each item. A significant main effect for group membership and the interaction between group membership and ability level in the regression indicates that ability level alone does not predict successfully passing the item. A significant interaction means that the DIF is non-uniform and that the slopes differ for the two groups such that their regression lines may cross and the item favors one group either at the higher or lower end of the ability spectrum.

LR has been found to be as powerful as the Mantel-Haenszel method in detecting uniform DIF (Swaminathan & Rogers, 1990; Rogers & Swaminathan, 1993).

The logistic regression equation looks very much like a general regression equation, but with the natural log-odds of the proportion of examinees who correctly answer the test item being predicted to:

$$\ln\left[\frac{p_i}{1-p_i}\right] = \beta 0 + \beta 1(totalscore) + \beta 2(group) + \beta 3(totalscore \times group) \qquad (8)$$

Alternately, this formula can be calculated separately for each group:

$$\Pr(u_{ij}) = \left[\frac{1}{\theta_{ij}}\right] = \frac{e^{(\beta 0 j + \beta 1 j \theta_{ij})}}{1 + e^{(\beta 0 j + \beta 1 j \theta_{ij})}} \quad (9)$$

with each person (*i*) in each group (*j*) with the correct response (*u*), with β0 as the intercept parameter and β1*j* as the slope parameter for that group, while *θij* is the ability of the person in the group being calculated, all converted to the natural log called base "*e*."

To determine whether an item has DIF, the logistic regression curves for the two groups should be different such that the *β*0 (intercept parameters) for each group are different indicating uniform DIF, or the *β*1 (slope parameters) for each group are different indicating non-uniform, or sometimes called, "crossing DIF"[4] (Swaminathan & Rogers, 1990).

The test of the DIF significance can be calculated by taking the chi-square for the total score and deducting it from the chi-square of the interaction and using the chi-square table to compare the results with two degrees of freedom. The calculation of effect size can be calculated similarly by taken the regression co-efficient (R-squared) for the interaction and subtracting the regression co-efficient for the total score (Zumbo, 1999).

For polytomous items, that is when the item responses are ordinal with partial credit scoring, such as for constructed response or short-answer test items, LR uses the same equation but requires the logistic regression analyses to be run multiple times after

[4] For logistic regression which uses the total test score as an estimate of true ability, non-uniform DIF could be indicative of adverse impact. See Dorans, 1989 and DeMars, 2010.

re-coding responses. Items that award two or three points for correct responses are converted into dichotomous data, meaning into correct and incorrect responses. Depending on the number of ordinal categories, the number of logistic regressions to be performed and binary coded files required will vary. For example, with four ordinal response categories, where an item can have scores of 0, 1, 2, or 3 points, three logistic regressions will need to be performed.

There are three model options for coding ordinal level data into dichotomous data with one predictor variable: cumulative odds ratio, continuation ratio, adjacent categories (Agresti, 1990). Using the four scoring levels example, in the adjacent categories model, binary coding is used to examine the probability difference in the 'steps'—the step is the increase from an incorrect answer to receiving one point, or from receiving one point to receiving two points, etc. For the same example with the cumulative odds ratio binary coding would compare the probability on each item of scoring one or more points to scoring two or more points to scoring three points, etc. Then, for the continuation ratio option, on any item the probability of getting the item incorrect is compared to receiving one or more points, the probability of scoring one point is compared to the probability of scoring two or more points, and the probability of scoring two points is compared to the probability of scoring three points, etc.

Reporting the results of DIF using LR for polytomous items can be quite cumbersome. Zumbo (1999) used a process for the constrained cumulative logits model, called the proportional odds model, with polytomous items that results in an omnibus DIF test. This model seems ideal for interpreting the DIF results for polytomous items, but

assumes equal slopes for each ordered response category. If polytomous items' slopes vary one of the unconstrained models should be used. Kristjansson, Aylesworth, Mcdowell, and Zumbo (2005), using the unconstrained cumulative logits model, reported significant DIF if any of the three regressions for a polytomous item was significant at the .017 level.

Zumbo (1999) first established effect sizes for LR by subtracting the full model chi-square from the chi-square of the total score, with two degrees of freedom and checked for significance. If the chi-square is significant then the magnitude or effect size of the DIF is calculated by taking the difference between the R^2 at the total score alone from the step where the group is added for uniform DIF, and the last step where the interaction is added for non-uniform DIF. Moderate DIF will yield an R^2 between 0.13 and 0.26, while for large DIF the R^2 should exceed 0.26.

Jodoin and Gierl (2001) were concerned that Type I errors might increase as sample size increased. In their study the data was taken from large scale achievement tests with sample sizes of 4,400 subjects. They proposed new guidelines for LR effect sizes to parallel those of SIBTest. According to Jodoin and Gierl (2001) R^2 of less than 0.035 indicates negligible level of DIF, R^2 between 0.036 and 0.070 indicates medium DIF and R^2 greater than 0.071 indicates large DIF. They conducted a simulation study to confirm that these effect sizes were adequate for larger sample sizes and varying ability distributions.

SIBTest. Simultaneous Item Bias Test (SIBTest, Shealy, & Stout, 1993) software program employs a multidimensional model for the analysis of DIF. The PolySIBTest

procedure was used for these DIF analyses, because it is an extension of the SIBTest procedure that can analyze items constructed response items which award a different number of points depending on the degree to which the correct answer was given. SIBTest begins by matching students from the reference and focal groups on ability based on estimates of their true score taken from their total test score (Shealy & Stout, 1993). If a significant portion of the focal group does not perform as expected on an item, or a significant portion of the reference group does not perform as expected on an item, then the item is exhibiting DIF. The group that is performing better than expected has the DIF in their favor. In the cases where the score distributions of the two groups being compared do not align for side by side statistical comparisons due to different group means and variances, SIBTest uses a statistical correction process (Shealy & Stout, 1993).

The Simultaneous Item Bias (SIB) method uses a multidimensionality or latent variable non-parametric approach to confirm or disconfirm the presence of DIF for an item or set of items (Shealy & Stout, 1993). The method compares the performance of the reference group to the focal group on the ability being measured and identifies DIF due variance from a secondary dimension, possibly a nuisance factor, not intended to be measured by the test.

First, validated items that measure the ability or trait of interest must be identified so that total scores on this valid test can be used to generate ability levels for use in comparing item performance of the groups. In some cases, the "valid test" may be a sub-set of the original test and include items that are suspected to have DIF (Bolt, 2000). The

suspect items are the focus for the DIF analyses. For Bundle DIF, the suspect items are bundled together to create a suspect test.

The valid subtest provides the estimates of true ability (θ) level for each subject, the latent variable, so that the members of the two groups with the same valid subtest scores will be compared to each other on the item or set of items that are suspected of DIF.

Probability of a correct response (P) for both the reference (r) and focal (f) groups is based on true scores estimates (T)—which stand in for latent ability estimates:

$$B(T) = P_r(T) - P_f(T) \tag{10}$$

Mean scores of the reference group and the focal group on the item being examined for DIF are estimated separately at each ability (i.e., valid true score) level (later represented as k):

$$\beta_{uni} = \int B(\theta) f_F(\theta) d(\theta) \tag{11}$$

Here the fF(θ) represents the probability density function for the focal group(F), yielding a βuni (unidirectional) index of the focal group's amount of DIF for that item across all ability levels, and $d(\theta)$ is the differential of theta (Bolt & Stout, 1996). Recall from the previous equation that $B(T)$, now $B(\theta)$, signifies the difference between the probability of the reference group getting the suspect item correct and the probability of the focal group getting the suspect item correct, when they are matched on their ability level for that latent trait.

Shealy and Stout (1993) took into consideration the likelihood of the focal and reference groups having a different distribution of scores so they added a regression

correction. Without this correction bias would be over-identified due to for example, the focal group having more scorers in the lower end of the scoring range than the reference group.

At each ability level (k) the true score mean estimates (V) will be re-calculated beginning with the following formula for each of the items and each of the groups, with $\overline{X}$ as groups' average valid subtest score:

The $\hat{V}$ is calculated

$$\hat{V}_{gk} = \frac{1}{m}(\overline{X}_g + b_g(k - \overline{X}_g)) \tag{12}$$

and

$$\hat{M}_{gk} = \left(\frac{\overline{Y}_{g,k+1} - \overline{Y}_{g,k-1}}{\hat{V}_g(k+1) - \hat{V}_g\ (k-1)} \right) \tag{13}$$

Here, $\overline{Y}_g$ represents the group's mean score on the item being examined at level k. Then the adjusted mean scores of the suspect subtest for each group at matching subtest score k are given by the regression correction

$$Y^*_{gk} = \overline{Y}_{gk} + \hat{M}_{gk}(\hat{V}_k - \hat{V}_g(k)) \tag{14}$$

$\overline{Y}_{gk}$ is the proportion of group g members at score level k who got this suspect item correct, $\hat{M}_{gk}$ is "a first-order Taylor series approximation of the rate at which the proportion is changing as true score changes," (from Shealy & Stout, 1993) and $\hat{V}$ is the average of true ability scores estimates of each group at level k.

After the regression correction, the DIF is calculated:

$$\hat{\beta}_u = \sum_{k=1}^{n} \hat{p}_k(\overline{Y}_{Rk} - \overline{Y}_{Fk}) \tag{15}$$

where $\hat{p}_k$ is the proportion of those who obtaining a score k on the suspect item and n is the number of items on the test. The DIF test statistic (β) is a calculated by taking the average across all ability levels. To determine significance of (β), the following formula is used by SIBTest, with the sample variance ($\hat{\sigma}$):

$$B = \frac{\hat{\beta}_u}{\hat{\sigma}(\hat{\beta}_u)} \tag{16}$$

Naturally, this $|B|$ should be greater than 1.96 significant at the .05 level for that item to contain DIF.

SIBTest has been expanded for use with polytomous data (for graded response, partial credit and generalized partial credit models) by Chang, Mazzeo, and Roussos (1996). They made two changes, first replacing the KR-20 during the regression correction, since it applies to dichotomous items, with Cronbach's alpha, and second instead of using the total number of items on the test (n) they matched the groups on ability based on their test score with nh the total number of possible points, to allow for polytomous items where the highest score can be more than one point:

$$\hat{\beta}_u = \sum_{k=1}^{nh} \hat{p}_k (\overline{Y}_{Rk} - \overline{Y}_{Fk}) \tag{17}$$

Since SIBTest does not allow any subject with "not-reached" missing items to be included in the data file, the missing are re-coded as non-passing scores or removed. For SIBTest, DIF estimates of effect size are based on significant beta statistics, such that, beta statistics of less than 0.05 or are indicative of practically insignificant DIF, beta statistics of 0.06 to 0.09 are indicative of low DIF, and beta statistics of 0.10 to 0.12 are

indicative of moderate DIF, and beta statistics greater than .12 are indicative of high DIF (for examples see Abbott, 2007; Stoneberg, 2004).

Grouping a sub-set of items together to be compared to the remaining items and treating that sub-set as one item suspected of having DIF is called Bundling or Differential Bundle Functioning (DBF). This option provided by SIBTest is used to further clarify any uniting hypothesis of item classifications that might underlie DIF or for groups of items that are linked together by referencing a chart or reading passage, often referred to as "*testlets*." DBF can be examined when an overall pattern in the DIF has been confirmed. For a small bundle of items, the DBF effect sizes are: Beta less than 0.10 should be disregarded as insignificant and Beta greater than 0.10 is moderate to large bundle DIF. Related items that could together result in DIF can be examined in such a way, but could also disconfirm a hypothesis about causes of DIF. For instance, grouping all items with total length greater than 30 words could be one way to test a hypothesis about item length. To check for DIF amplification, a situation that results from small amounts of DIF in one item of a bundle making that bundle significant, the bundle should also be purified by removing an item and checking the significance of the bundle.

Error Rates and Impact

Research suggests analyzing a set of data with more than one DIF method to strengthen one's case for the presence of DIF, accompanied by purification methods for final DIF confirmation (Ackerman, 1992; Camilli & Shepard, 1994). Purification methods vary but generally involve similar steps; after finding DIF the differentially

functioning item or items are removed from the test and then the total score matching procedure is run without that item or items and this new total score will be used when assessing the DIF. Each of the original items with DIF can be examined one at a time, iteratively. Purification has been shown to provide improved DIF results (Clauser, Mazor, & Hambleton, 1993) and controlling Type I error (Su & Wang, 2005), but may not change the results if a valid subtest has been used initially (French & Maller, 2007; Soares, Goncalves, & Gamerman, 2009). French and Maller (2007) further note that their results show that when DIF rates are in actuality high—above 25%—it is best not to perform the purification technique for DIF using LR because it will likely increase Type I error, and to instead depend on measures of effect size to best control Type I error rates. In other words, purification should not be performed under conditions of pervasive DIF.

Simulations to test error rates generally vary the percentage of DIF, sample sizes, effect size of DIF, test length and degree of impact, the latter being of the most concern and providing the most resistance to purification. For example, SIBTest performs well even under situations of low sample sizes of 200 subjects without significant increases in Type I (Roussos & Stout, 1996), but Roussos and Stout (2005) warn that if ability distributions for the groups are too far apart, if there is severe impact, the Type I error will increase as the sample size or test length decreases. While some research has shown SIBTest to be robust under conditions of varying group performance distributions (Chang, Mazzeo, & Roussos, 1996; Pei & Li, 2010), Klockars and Lee (2008) state that as the reliability of the test decreases, the SIBTest method overestimates items favoring the focal group.

When non-uniform DIF is likely to occur, for example under situations where there is an interaction between the primary test measure and secondary variables that are unintentionally being measured, some research has shown that LR is mostly likely the strongest DIF method to employ (Kristjansson et al., 2005), but caution should be taken with LR when using a small sample size as the method may over-identify items with significant DIF (Gómez-Benito, Hidalgo, & Padilla 2009). The SIBTest method is also capable of identifying non-uniform DIF for dichotomous items while maintaining power and controlling for Type I error (Finch & French, 2007; Li & Stout, 1995). In a simulation study addressing the issue of maintaining statistical power and reducing Type I error comparing SIBTest, logistic regression and two other DIF methods, Finch & French (2007) found that SIBTest was the best method for non-uniform DIF. However, all the items on the test must be dichotomous to use crossing DIF in SIBTest to measure non-uniform DIF. So far, results have yielded uniform DIF and there has been no evidence to suggest the presence of non-uniform DIF in ELL comparisons (Martiniello, 2008; Wolf & Leon, 2009).

Pei and Li (2010) compared LR to SIBTest with simulated samples whose means are unequal and showed impact, SIBTest performed the best, while LR performed the worst, having the most inflation of Type I error. They also note that for any of these methods, as the distributions of the two groups of interest get farther apart there is also an increase in the likelihood of Type I error. Recently it has been suggested that, under situations where the groups' normality of distribution assumption is violated, nonparametric methods such as PolySIBTest should be used (Woods, 2011).

Research has shown that the M-H method may not respond to purification procedures, maintaining Type I error, in the presences of impact (Su & Wang, 2005). DeMars (2010), explains that when the distributions of the two groups differ due to impact, especially with large sample sizes but not necessarily large valid sub-tests, LR will yield Type I errors for DIF favoring the reference group, an effect that may be more pronounced for unreliable or shorter tests. DeMars (2010) provides a detailed explanation beginning with the understanding that researchers examining tests that mostly include multiple-choice items tend to use three parameter models such as LR to account for varying slopes and the lower asymptote for the guessing parameter, but do not take into account the curvilinear shape of the data. If impact is not present there will not be a problem since the log odds slope of both groups will be calculated by the LR procedure as slightly lower than is true. However, when the data reveals that impact is present, the focal group item difficulty is miscalculated by the LR method which by definition of its equation chooses the best *linear* relationship. For the more difficult multiple-choice items, the guessing parameter is the lower asymptote for the item so the prediction line starts higher and is affected by each of the group means but the predictions of scores will not curve where needed.

Even though the presence of impact in the sample comparison is a main concern, it is believed that specifically accounting for the influence of reading test scores will reduce the effect of the gap between the focal and reference group distributions. In a classroom study that compared beginner, intermediate or advanced ELLs to non-ELLs, math scores are best predicted by English reading proficiency levels, with beginner ELLs

receiving the lowest math scores (Beal, Adams, & Cohen, 2010). One limitation that has been previously mentioned is that the reading test has also been developed for non-ELLs to measure grade-level reading, and as such cannot represent an accurate measure of English as a second language. Therefore, while the reading scores may explain some variability in science test scores, they cannot be treated as an ideal measure of English reading and writing fluency. This variable alone will not be able to fully explain ELLs' interpretation of an item, but it is hoped that it may shed some light on the DIF that is uncovered. Additionally, to avoid the problem with a more accurate prediction line described by DeMars (2010), the LR approach is used with the addition of a quadratic term.

Performing DIF analyses with language learning samples have been uniquely difficult for psychometricians (Hauger & Sireci, 2008). The sample's characteristics and how to best address the research questions are considered in determining which DIF methods should be used. In the interest of reducing variability in the sample by taking into consideration another group characteristic logistic regression was chosen and, in order to yield the most power and the lowest error, it was determined that the most robust DIF method to use with this much complexity was SIBTest.

Next, the SIBTest method was chosen to analyze DIF and to account for the presence of *testlets*, also referred to as local item dependence, meaning that a common stimulus (e.g. a reading passage) is referenced for sets of items that together make up a testlet. Beretvas & Walker (2012) have used both multi-level models and SIBTest successfully to examine testlet level DIF, which, when found, indicates that some aspect

of the scenarios is causing DIF. After examining DIF for each item, DIF is examined for each scenario and for the group of independent items by using differential bundling functioning (DBF). Bundling together items that reference the same scenario is performed to determine the effect of testlets. For the most part, SIBTest is able to control for Type I error when used to examine DIF in testlests (Lee, Cohen & Toro, 2009; Zwick & Thayer, 2003).

Chapter III: Methods

This chapter describes the goals of the current study, the sample, the tests and the way linguistic complexity was examined. The dissertation focuses on three general aims, namely 1) examining replication of items used in multiple years of testing to determine whether patterns of DIF in the first sample are reflected with a second sample of students, 2) comparing the results of two methods of DIF with regard to items that favor either group, 3) identifying the properties of items that show DIF for ELLs, and 4) determining whether items that are part of a testlet exhibit DIF as a group, indicating that possibly something about the scenario rather than the item is correlated to the DIF.

Exploratory studies, which have found DIF and related it to linguistic complexity, have not found consistent results over different grade levels. Comparing samples in the same regional area, at the same grade level, and repeating items may be more helpful in uncovering and making sense of patterns of DIF. A little over one-third of the items are present on both tests used in this study, so when DIF is found in science items in separate samples it can be seen as confirmation of that DIF operating across ELLs with different unmeasured background variables such as different countries of origin.

One predictor that is relevant to linguistic complexity is reading level, therefore it is included in the analysis in order to determine whether it plays a significant role in DIF, beyond ELL group membership. Logistic regression has been chosen as the first method of analysis mainly for its flexibility. In addition to allowing for multiple predictors to be included in the same DIF analysis, logistic regression maintains low levels of Type I and II error while examining non-uniform DIF. Reading scores are added as a predictor

variable in order to attain a better sense of the construct being measured by accounting for the influence that reading level may have both for non-ELLs and ELLs. Since previous research shows a correlation of test performance with SES, it can also provide clues to DIF that may stem from SES differences.

SIBTest, the second method selected for this DIF comparison, is unique in taking a multidimensional and IRT approach, and able to perform bundling analyses. The method's relevant feature is the ability to bundle items because of its focus on examining any differential functioning at the testlet level. Because SIBTest is a multidimensional approach, it can recognize linguistic complexity as a second unintended dimension in tests.

Research Questions

My research seeks to answer the following questions based on previous findings and recommendations:

1) Does the pattern of DIF in the 2006 test repeat itself in the 2007 test for those fifth grade science items that are used in both years?

2) Do two DIF methods identify the same items with DIF favoring non-ELLs and items favoring ELLs?

3) Do any sets of items that refer to a scientific reading scenario, called a testlet, exhibit DIF when bundled together?

4) Do items with DIF favoring non-ELLs correspond to high linguistic complexity coding and those that favor ELLs incorporate schematic visual aids?

Sample

In the U.S., Spanish is spoken by over half of the population who do not speak English at home (The U.S. Census, 2010). Thus, the chosen sample takes into consideration the high number of students who speak Spanish at home. Focusing on one particular language group of ELLs limits the high degree of variability within the ELL population due to language. ELLs are defined in this study as students whose data records indicate their primary language is Spanish. Comparing all students for whom English is not the native language instead of just those who are labeled as being in ELL/bilingual programs will help increase the score variability. ELLs selected for this sample include students with special needs, students in gifted programs, foreign-born children, second or third generation Americans, children born in the U.S. to immigrant parents, and migrant students whose families work in the areas of agriculture and fisheries.

Only Spanish-speaking ELLs who did not complete the test and those who received a presentation accommodation were excluded from the sample. It is important to note that, while this sample consists of Spanish-speaking ELLs, different dialects of Spanish are represented within the sample but were not identified for the purposes of these analysis. The English reading proficiency scale scores of the sample were used as a covariate to control some of the high degree of variability within this group.

There were 6,254 Spanish-speaking ELLs and 54,962 non-ELLs for a total of 61,216 students in 2006; there were 8715 Spanish-speaking ELLs and 61,835 non-ELL students' for a total of 70,550 in 2007. Approximately forty-five percent of the 2006

sample and twenty-eight percent of the 2007 sample were in a specialized ELL or bilingual program.

For both tests, those students who did not answer any items on the science test and were coded by their teachers as "not tested" and have been removed. Those students who did not answer any items on the reading test were also removed. Finally, those students who received presentation accommodations on either test were removed since it cannot be determined whether this accommodation included reading assistance. The final sample consists of a random sample of non-ELLs that approximately equaled the total sample of ELLs to the nearest 50 for the 2006 and 2007 tests.

Table 2

Data Reduction of 2006 and 2007 Test Samples

Variable Removed	Removed 2006	Remaining 2006	Removed 2007	Remaining 2007
Science test not attempted	2379	58,837	3,524	67,026
Reading test not attempted	733	58,104	572	66,454
Science presentation accommodation	2073	56,031	2814	63,640
Reading presentation accommodation	89	55,942	112	63,528
Primary language-Spanish		5,156		6,771
Primary language-English		50,786		56,757
Sampled English speakers		5,200		6,800

Instrument

This study analyzed all the items from Washington State's standardized science test for fifth-grade from the spring of 2006 and the spring of 2007. The tests present a

unique challenge in that groups of items refer to scenarios, previously described as testlets. The testlets are composed of items that all refer to the same scientific scenario, generally an observation or an investigation. The items that do not refer to any scenario are independent items.

From hundreds of possible items that are created, items are reviewed for alignment to content standards, undergo a review for sensitivity and bias by a specially chosen diversity committee, and piloted prior to administration. Only items with acceptable item analysis data are included in the operational tests. Regarding the item response types in the final test forms, there are 21 multiple choice items, 10 short answer and 2 extended response items for each test. Extended response items award 0 to 4 possible points; short answer items award from 0 to 2 points, and multiple choice items award 1 point each. Cronbach's alpha for the both the 2006 and 2007 science tests are reported as 0.85. Of the 33 items per test, 14 items appear in both the 2006 and 2007 tests. Table 3 compares the item features of the two tests:

Table 3

Science Test Features by Year

Item Features	**2006**	**2007**
Multiple choice	21	21
Extended response	2	2
Short answer	10	10
Independent items	4	2
Scenario-based	29	31

The majority of items administered on the test are scenario-based items; these items that refer to one of the eight scenarios are analyzed on their own in one analysis and grouped together as a testlet for a second analysis. Independent items are those that can be answered on their own without referencing a scenario. Table 4 shows the different scenarios compares the items within each testlet.

Table 4

Science Scenarios

Scenario	Knowledge Type	Constructed Response (CR)	Multiple Choice (MC)	Total Items
Puddle	Investigation	1	5	6
The birds	System	3	2	5
Boiling	Investigation	3	4	7
Weather	System	1	4	5
Lettuce	System	2	4	6
Heat	Investigation	1	5	6
State tree	System	2	5	7
Compost	System	2	3	5

The scenarios that accompany each set of questions concern knowledge of either a scientific system or a scientific investigation. Each scenario is accompanied by a visual representation. The systems scenarios present a living system and focus on certain details within that system. The questions ask students to explain the way the system functions, to provide solutions, and to design their own experiments. The investigations begin with a research question or hypothesis and provide lists of materials used in the investigation,

the steps taken, and the results. The questions ask students to analyze the situation, draw conclusions from the results, and design a new investigation.

Coding Scheme

To guide interpretation of the DIF pattern of results, a coding scheme was devised. For each item, linguistically complex characteristics, tables, diagrams and schematic pictures were coded (see Appendix A and B). Coding visual aids was important for examining the effect compared to non-schematic visual aids, since the presence of schematic visual aids in an item has been shown to relieve some of the linguistic burden for ELLs (Martiniello, 2009; Wolf & Leon, 2009).

The conceptual framework diverges from the categorizing of linguistic features to a more probabilistic and pattern-based approach. On a given test item, combinations of different types of linguistic complexity occur together. Items that contain only one type of complexity are difficult to find or create, meaning that researchers must identify, for example, both the use of passive voice and conditional clauses within a given item; consequently, each item is examined for several features of linguistic complexity (see Appendix A). Some item features, such as English-Spanish cognates and item response format, require more in-depth analysis if a pattern is found that indicates DIF is related to one of these features; therefore, to identify patterns the items flagged for DIF and the items not flagged for DIF are compared with regard to their item features.

A manageable set of linguistic features were explored and coded for each test item. As explained in the previous chapter, empirical research has identified several

features as causing confusion for ELLs on tests. Additionally, the framework takes into account the findings of recent similar studies that address the possible scaffolding effect of the non-textual elements found on tests on ELLs' understanding (Martiniello, 2008; Wolf & Leon, 2009).

It is important to note that this coding scheme of linguistically complex features is in no way all inclusive or fixed**.** First, it would not be efficient or fruitful to examine every linguistic feature, either because that feature is not present in any items on the given test being examined or if the feature hasn't been shown to be problematic for ELLs. Second, due to the nature of a specific item, some linguistic features may be problematic when encountered on one item but would not be problematic when encountered on a different item. Solano-Flores (2012) refers to this quandary as the probabilistic nature of language.

Ratings of Linguistic Complexity

After coding the combinations of linguistic features for each test item, the linguistic complexity is examined from the perspective of educators whose native language is Spanish and who teaches ELL students. This part of the conceptual framework relates to the importance of including the judgments of Spanish-speakers with experience in the field of education. Although raters cannot be expected to predict precisely which academic vocabulary or grammar has been taught to ELLs, they are expected to provide a reasonable evaluation of linguistic complexity posed based on their knowledge and experiences of their own ELL students.

For the items that were released by the state's department of education, three native Spanish speakers who work within academic fields ranked the set of items by increasing linguistic complexity—from the easiest to the most difficult to read and comprehend. These raters were instructed not to consider crossed out scientific words as part of the task. These rankings were compared to the DIF results to determine whether those items ranked as having the highest linguistic complexity are flagged for DIF favoring non-ELLs and those items ranked as having the lowest linguistic complexity were not flagged for DIF at all or were flagged for DIF favoring ELLs.

DIF Analyses

For this dissertation, datasets from two contiguous years of a fifth-grade state science tests were analyzed using two DIF techniques with a focal group of Spanish-speaking ELLs and a reference group of non-ELLs. These DIF methods were SIBTest and logistic regression. The DIF results are presented with the effect sizes from low to high signifying the magnitude of DIF. The coefficient displays either a positive sign indicating the reference group (non-ELL) is favored or a negative sign to indicating the focal group (ELL) is favored. Results from the two methods were then compared for agreement as to which items and testlets were flagged for DIF favoring either group.

Logistic regression. For each item, DIF was calculated using a hierarchical logistic regression approach, with variables entered in the order shown in the equations. For each item, the logistic regression included reading scores as a predictor and a quadratic term to improve the prediction line (DeMars, 2010). The natural log-odds of the proportion of examinees who correctly answer the test item was found by adding together the products

of the coefficients (β) and each term, with the total z-transformed score and total score squared (θ^2):

$$\ln\left[\frac{p_i}{1-p_i}\right] = \beta 0 + \beta 1(totalZscore) + \beta 2(\theta^2) + \beta 3(group) + \beta 4(totalZscore \times group) + \beta 5(readingscore) + \beta 6(readingscore \times group) + \beta 7(totalZscore \times group \times readingscore) \quad (18)$$

Jodoin and Gierl (2001) state that z-score transformation helps decrease possible collinearity between predictor variables.

The reading scale scores acquired from the state reading test served as a predictor variable in place of cut-off points that create categorical reading levels as suggested by other researchers (Solano-Flores, 2008). Using reading test scores increases the variance and provides a more accurate measurement of reading comprehension to explain a portion of the variance in science item passing rates. The cumulative odds ratio coding scheme for the constructed response items does not code any items as missing and therefore every score is included in each analysis. This process of combining categories is an analysis of differential step functioning, it compares the differential functioning at each scoring level from zero to the highest score. The cumulative odds ratio coding strategy was used for LR constructed response items (Agresti, 1990). French and Miller (1996) found that using the continuation ratio or the cumulative odds ratio resulted in higher power than the adjacent categories strategy. A polytomous item is reported to have significant DIF if any of the logits was significant at the .000 level, and the coefficient reported is for the significant logit.

For the testlet analyses, as with polytomous items, re-coding is required at each score level. The total testlet score is used to examine differential functioning at the testlet level. The number of regressions analyses needed is one less than the highest number of points possible for each testlet with cumulative odds ratio coding. The differential bundling equation looks the same as the logistic regression when applied to polytomous items, except that the dependent variable becomes the testlet score, replacing the item score. The hierarchical logistic regression equation for the bundling of the items from each testlet:

$$\ln\left[\frac{p_i}{1-p_i}\right] = \beta 0 + \beta 1(totalZscor\ e) + \beta 2(\theta^2) + \beta 3(group) + \beta 3(totalZscor\ e \times group) \quad (19)$$

SIBTest. Each item was examined for DIF individually using the SIBTest procedure, which employs a multidimensional Item Response Theory model for the analysis of DIF. The PolySIBTest procedure was used for the polytomous items, those constructed response items which are scored from zero to four. Since the SIBTest method does not allow any subject with missing values to be included in the data file, there were nine to twelve percent fewer cases in each group analyzed by SIBTest than by logistic regression. Testlets were examined for differential bundle functioning using SIBTest's DBF, such that each testlet was bundled and treated as an item and compared to the remaining items.

Two years of Washington State's fifth grade science assessments (WASL-science) were analyzed to examine replication of patterns of DIF, to compare the results

of two DIF methods, and to identify the properties of items that show DIF for ELLs. Coding and bilingual educators' ratings were used to describe the level of linguistic complexity in the replicated items to gauge whether language issues were related to the DIF patterns found. In the next section, each of the statistical results for the comparison of the two methods, the item replication, the individual DIF and the testlet level DIF using both methods are described. Item analyses and testlet analyses with regard to linguistic complexity for items with significant DIF are further explained.

Chapter IV: Results

This chapter begins with the general demographic information and then each of the research questions is addressed. The majority of Spanish-speaking ELLs in both years of testing were ethnically Hispanic Latino and there was a balance of males and females (see Table 4). The low SES groups scored lower than the high SES groups on both the 2006 and 2007 science tests (see Tables 5 and 6). This corresponds with results found in previous studies regarding the correlation between low SES and lower test scores (Abedi & Lord, 2001; Krashen & Brown, 2005).

Table 5

Demographic Frequencies

	2006 Test			2007 Test		
	ELL (*N*=5156)	Non-ELL (*N*=5200)	Total (*N*=10,356)	ELL (*N*=6771)	Non-ELL (*N*=6800)	Total (*N*=13,571)
Native American/Alaskan	7	151	158	1	1	2
Asian	8	306	314	38	427	465
Black African-American	5	298	303	87	435	522
Hispanic Latino	5095	351	5446	4997	486	5483
White	34	4020	4054	1471	5084	6555
Hawaiian/Pacific Islander	0	16	16	6	27	33
Multi-Racial	5	45	50	87	106	193
Females	2591	2596	5187	3387	3461	6848
Males	2565	2604	5169	3384	3339	6723

Note: Ethnicity data was unavailable for some students.

In general ELLs' mean performance was lower than non-ELLs' mean performance on the total science test and on the reading test, while minimum and maximum scores differed by two to four points. The 2006 data set contained an extremely low number of high SES ELLs in comparison to the 2007 data set; therefore, even though each test had the same number of items the difference in ELL scores was less pronounced on the 2007 test (see Tables 6 and 7). The low SES ELLs, particularly on the 2006 test, had the lowest mean of science scores. Conversely, the high SES ELLs performed almost as well as non-ELLs on the 2007 test, such that the high SES ELLs had higher mean scores than the low SES non-ELLs.

Table 6

2006 Total Science Score by Group

	M	SD	Median	Min	Max
Low SES (*N=6463*)	20.90	7.42	20	4	44
High SES (*N=3891*)	28.76	8.57	29	6	48
ELL					
Low SES (*N*=4768)	19.69	7.42	19	4	42
High SES (*N*=388)	22.82	7.37	22	7	44
Total ELL (*N*=5156)	19.92	7.05	19	4	44
Non-ELL					
Low SES (*N*=1697)	24.29	7.61	24	5	44
High SES (*N*=3503)	29.42	7.89	30	6	48
Total Non-ELL (*N*=5200)	27.74	8.17	28	5	48

Table 7

2007 Total Science Score by Group

	M	SD	Median	Min	Max
Low SES (*N=7471*)	20.79	7.77	20	1	47
High SES (*N=6100*)	28.61	8.57	29	5	48
ELL					
Low SES (*N*=5299)	19.87	7.50	19	1	47
High SES (*N*=1472)	26.62	9.19	27	5	47
Total ELL (*N*=6771)	21.87	8.36	21	3	47
Non-ELL					
Low SES (*N*=2172)	23.02	7.98	22	3	47
High SES (*N*=4628)	29.25	8.50	29	5	48
Total Non-ELL (*N*=6800)	27.98	8.67	28	5	48

The lower scores of ELLs on the 2006 test can be observed as a slightly wider and slightly less negatively skewed distribution of ELL total scores than in the larger sample of ELLs in the 2007 test where there is an increase in the number of ELLs performing one standard deviation above the mean. The difference in the ELL score distributions between the tests can be seen in Figures 1 and 2.

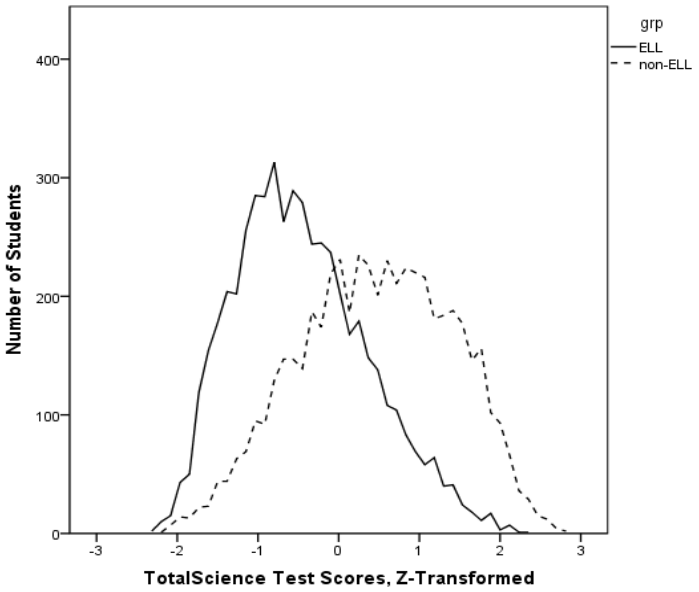


Figure 1. 2006 total science test score distribution by group

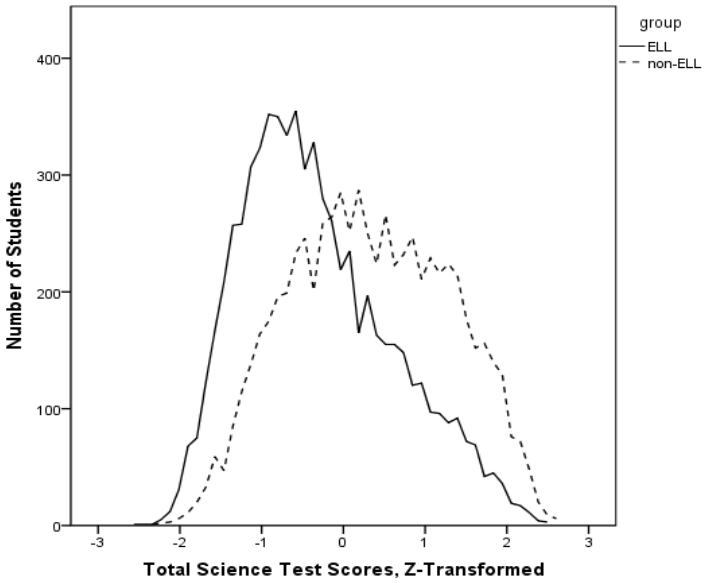


Figure 2. 2007 total science test score distribution by group

Logistic Regression Item Analyses

In what follows, items will be identified in terms of the title of the science scenario (e.g. Weather) and the item number within the scenario. For example, Weather1 is the first item in the Weather scenario. For the polytomous items, in previous research the determination of significance employed a Bonferroni correction (Kristjansson, et al., 2005). In this paper, a polytomous item with 3 step analyses is reported to contain significant DIF if the hierarchical logistic regression analysis for any of the differential step analyses was significant at the $p \leq .000$ level. When more than 3 step analyses are required, more than one step must be significant at the $p \leq .000$ for that item to be flagged for DIF.

For the 2006 test, a total of 16 items were flagged for significant DIF, 10 items favoring non-ELLs and 6 items favoring ELLs. Initially 2 of these 10 items were significant for uniform DIF favoring non-ELLs, but also contained a significant interaction, indicating there was non-uniform DIF. Of the 16 items with significant DIF, 8 were multiple choice items and 8 were constructed response items (see Table 8).

The effect size was calculated by examining the difference or change in the Nagelkerke R^2 ($R^2\Delta$) between the variance accounted for by the z-transformed total score on the science test and the variance accounted for by adding the group variable to the regression. None of the items flagged for significant DIF had an effect size above the 0.035 required to be considered more than negligible DIF, thus the DIF is said to be of little practical significance.

In general, the pattern of significant DIF ($p \leq .000$) was such that multiple choice items tended to favor non-ELLs and most of the constructed response items favored ELLs. In 2006, all the items flagged for significant DIF favoring ELLs were constructed response items, and all but two items flagged for significant DIF favoring non-ELLs were multiple choice items (see Table 8).

Three items contained non-uniform DIF. Initially, for each of these items the DIF favored non-ELLs, but at the next step in the process when the interaction was examined non-uniform DIF was found to be significant. First, Weather5 with non-uniform DIF on the 2006 test contained both a significant reading x group interaction and a more complicated three way interaction of reading x group x totalZscore (see Figure 3). The next item flagged for non-uniform DIF, Boiling3, contained a significant three-way interaction—reading x group x total score z-transformed (see Figure 4). Boiling7, a four-point constructed response item contained a significant interaction—reading x item. The DIF for Boiling7 looks negligible when viewing the proportion of ELLs passing the item compared to the proportion of non-ELLs passing the item at each total test score (see Figure 5). The item has a high demand for writing skills, so a higher proportion of high scoring ELLs have a better chance of being awarded the full four points on the item, and low scoring ELLs have a lower chance of receiving one point on this item.

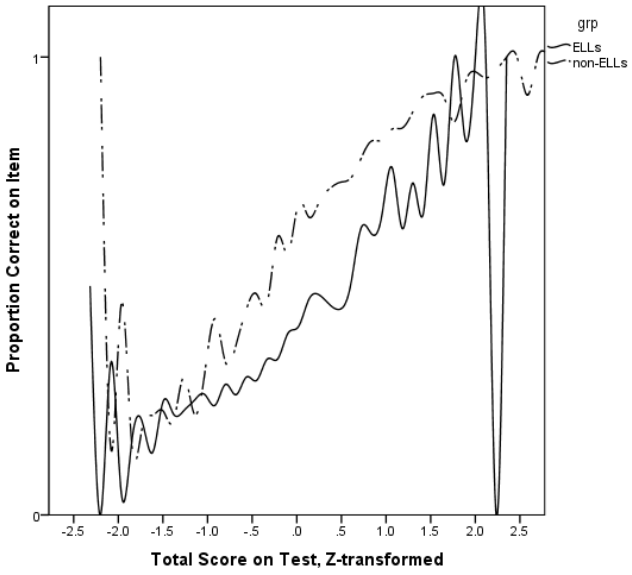


Figure 3. Weather5 with non-uniform DIF in 2006

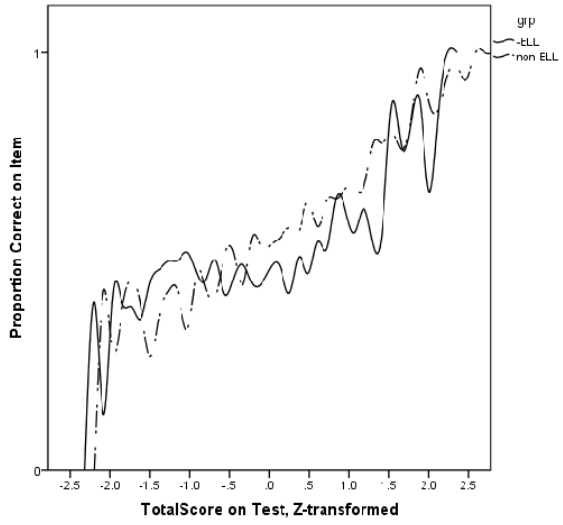


Figure 4. Boiling3 with non-uniform DIF in 2006

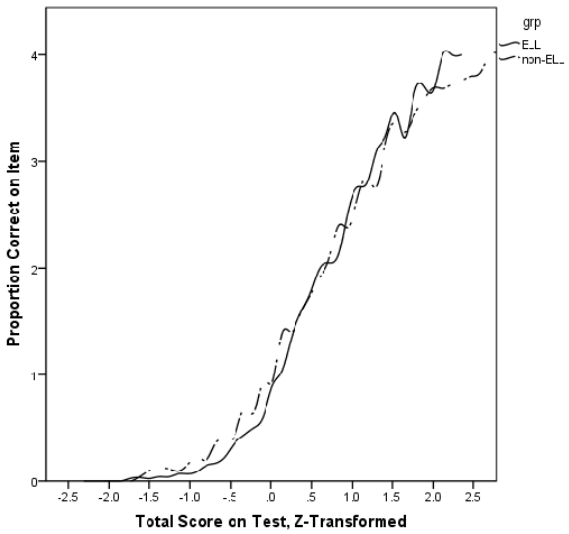


Figure 5. Boiling7 with non-uniform DIF in 2006

Table 8

Logistic Regression 2006 Test Item Analyses

Item	Response Type	Scenario ID	Logistic Regression Coefficient	$R^2\Delta$	Favors	Reading
1	mc	Puddle1				
2	mc	Puddle2				
3	mc	Puddle3	0.106	0.002	Non-ELL	
4	cr	Puddle4	-0.265	0.008	ELL	
5	mc	Puddle5	0.144	0.004	Non-ELL	
6	mc	Puddle6				
7	mc	The birds1				
8	cr	The birds2				
9	mc	The birds3	0.109	0.002	Non-ELL	Y
10	cr	The birds4	-0.031	0.003	ELL	Y
11	cr	The birds5				Y
12	mc	Independent#1				Y
13	mc	Boiling#1				
14	mc	Boiling#2	0.095	0.001	Non-ELL	
15	mc	Boiling#3	0.173[a]	0.003	Non-ELL	Y
16	cr	Boiling#4	-0.220	0.003	ELL	
17	mc	Boiling#5				Y
18	cr	Boiling#6	-0.305	0.010	ELL	Y
19	cr	Boiling#7	0.211[a]	0.004	Non-ELL	
20	mc	Weather#1	0.107	0.002	Non-ELL	Y
21	mc	Weather#2				Y
22	cr	Weather#3				Y
23	mc	Weather#4				Y
24	mc	Weather#5	0.369[a]	0.023	Non-ELL	
25	cr	Independent3	0.117	0.002	Non-ELL	
26	mc	Lettuce1				Y
27	mc	Lettuce2	0.219	0.008	Non-ELL	
28	cr	Lettuce3				
29	mc	Lettuce4				Y
30	mc	Lettuce5				Y
31	cr	Lettuce6	-0.274	0.010	ELL	Y
32	cr	Independent#2	-0.260	0.007	ELL	
33	mc	Independent4				

Note. Polytomous item coefficient and $R^2\Delta$ reported pertain to highest significant step; #items used in 2006 and 2007; cr = constructed response; mc = multiple choice; [a] = interaction; Y=significant at the $p \leq .000$; $p \leq .000$.

For the 2007 test, a total of 14 items were flagged for significant DIF, 9 favoring non-ELLs and 5 favoring ELLs. Of these 14 items with DIF, 9 were multiple choice items and 5 were constructed response items (see Table 9). Among the 5 constructed response items flagged for significant DIF, 4 constructed response items favored ELLs and 1 constructed response item favored non-ELLs. Among the 9 multiple choice items flagged for significant DIF, 8 multiple choice items favored non-ELLs and only 1 multiple choice item favored ELLs. The 3 items that contained non-uniform DIF on the 2006 test did not contain non-uniform DIF on the 2007 test.

Again, none of the items with significant DIF has an effect size above the 0.035 required to be considered more than negligible DIF.

Table 9

Logistic Regression 2007 Test Item Analyses

Item	Response Type	Scenario ID	Logistic Regression Coefficient	$R^2\Delta$	Favors	Reading
1	mc	Heat1	-0.151	0.001	ELL	Y
2	mc	Heat2	0.213	0.002	Non-ELL	
3	cr	Heat3	-0.299	0.002	ELL	
4	mc	Heat4	0.160	0.001	Non-ELL	
5	mc	Heat5				
6	cr	Heat6				
7	mc	Weather#1	0.130	0.001	Non-ELL	
8	mc	Weather#2				Y
9	cr	Weather#3				Y
10	mc	Weather#4				Y
11	mc	Weather#5	0.499	0.011	Non-ELL	
12	cr	Independent5				
13	mc	Boiling#1				
14	mc	Boiling#2				
15	mc	Boiling#3				Y
16	cr	Boiling#4	-0.157	0.001	ELL	
17	mc	Boiling#5				Y
18	cr	Boiling#6	-0.279	0.003	ELL	Y
19	cr	Boiling#7				Y
20	mc	StateTree1				
21	mc	StateTree2				Y
22	cr	StateTree3				Y
23	mc	StateTree4				
24	mc	StateTree5				
25	mc	StateTree6	0.604	0.008	Non-ELL	Y
26	cr	StateTree7				Y
27	mc	Compost1	0.286	0.001	Non-ELL	
28	cr	Compost2	0.217	0.002	Non-ELL	Y
29	mc	Compost3	0.320	0.007	Non-ELL	
30	mc	Compost4	0.312	0.005	Non-ELL	
31	cr	Compost5				Y
32	cr	Independent#2	-0.473	0.001	ELL	
33	mc	Independent#1				Y

Note. Polytomous item coefficient and $R^2\Delta$ reported pertain to highest significant step; #items used in 2006 and 2007; cr = constructed response; mc = multiple choice; Y=significant to the $p \leq .0000$; $p \leq .000$.

Reading score as covariate. The inclusion of the reading variable does not change the DIF results but is helpful in explaining some of the significant DIF. It must be noted that having adequate reading skills is important on almost any test. Reading scores are a significant predictor of item scores in both items with and without DIF and were the strongest predictors in scoring well for half the items on the test. Reading scores were a significant predictor for 15 items on the 2006 test and 15 items on the 2007 test.

Weather5 on the 2006 test did not contain a significant main effect for reading, but there was a significant reading x group interaction, and a more complicated significant three way interaction of reading x group x totalZscore. On the 2007 test, there was a significant reading x group x totalZscore interaction for Weather2, however Weather2 is not flagged for DIF. These significant interactions illuminate the potential complexity of explaining the DIF against ELLs.

SIBTest Item Analyses

Using SIBTest for the 2006 test, a total of 11 items were flagged for practically significant DIF, 7 items favoring non-ELLs and 4 items favoring ELLs. Four items favoring non-ELLs had a negligible effect (see Table 10). The 4 items with DIF favoring ELLs with moderate to large effect sizes were constructed response items, while the items favoring non-ELLs were both multiple-choice and constructed response items. The effect sizes varied; however, items favoring non-ELLs included a few items with low and negligible DIF, while the items favoring ELLs mainly result in moderate to high effect sizes. Two items with the highest effect sizes favoring ELLs required students to write a summary or conclusion about experiments. Overall, the tests were not differentially problematic for ELLs since few items had significant DIF with moderate or high effect sizes favoring non-ELLs.

For the 2007 test the DIF was balanced such that a total of 6 items were flagged for practically significant DIF, 3 items were identified with significant DIF favoring non-ELLs, and 3 items were identified with significant DIF favoring ELLs. Four items favoring non-ELLs and 2 items favoring ELLs had a negligible effect size (see Table 11). For the three items favoring either group, 2 items had a low effect size, and 1 item a moderate effect size. The items with DIF favoring ELLs were constructed response items, while the items with DIF favoring non-ELLs were mainly multiple choice and one constructed response item. This pattern is relatively similar to the 2006 test.

Table 10

SIBTest 2006 Test Item Analyses

Item	Response Type	Scenario ID	SIBTest Beta	Effect Size	Favors
1	mc	Puddle1			
2	mc	Puddle2			
3	mc	Puddle3			
4	cr	Puddle4	-0.099	Moderate	ELL
5	mc	Puddle5	0.040		Non-ELL
6	mc	Puddle6			
7	mc	The birds1			
8	cr	The birds2	0.088	Low	Non-ELL
9	mc	The birds3	0.073	Low	Non-ELL
10	cr	The birds4			
11	cr	The birds5			
12	mc	Independent#1			
13	mc	Boiling#1			
14	mc	Boiling#2	0.031		Non-ELL
15	mc	Boiling#3	0.055	Low	Non-ELL
16	cr	Boiling#4	-0.095	Moderate	ELL
17	mc	Boiling#5			
18	cr	Boiling#6	-0.150	High	ELL
19	cr	Boiling#7			
20	mc	Weather#1	0.050		Non-ELL
21	mc	Weather#2	0.049		Non-ELL
22	cr	Weather#3			
23	mc	Weather#4			
24	mc	Weather#5	0.147	High	Non-ELL
25	cr	Independent3	0.055	Low	Non-ELL
26	mc	Lettuce1			
27	mc	Lettuce2	0.069	Low	Non-ELL
28	cr	Lettuce3			
29	mc	Lettuce4			
30	mc	Lettuce5			
31	cr	Lettuce6	-0.163	High	Non-ELL
32	cr	Independent#2	-0.124	High	ELL
33	mc	Independent4			

Note. cr = constructed response; mc = multiple choice; #items used in both tests. $p \leq .000$.

Table 11

SIBTest 2007 Test Items DIF Results

Item	Response Type	Scenario ID	SIBTest Beta	Effect Size	Favors
1	mc	Heat1	-0.034		ELL
2	mc	Heat2			
3	cr	Heat3	-0.077	Low	ELL
4	mc	Heat4	0.028		Non-ELL
5	mc	Heat5			
6	cr	Heat6			
7	mc	Weather#1	0.035		Non-ELL
8	mc	Weather#2			
9	cr	Weather#3			
10	mc	Weather#4			
11	mc	Weather#5	0.099	Moderate	Non-ELL
12	cr	Independent5			
13	mc	Boiling#1			
14	mc	Boiling#2			
15	mc	Boiling#3			
16	cr	Boiling#4	-0.042		ELL
17	mc	Boiling#5			
18	cr	Boiling#6	-0.084	Low	ELL
19	cr	Boiling#7			
20	mc	StateTree1			
21	mc	StateTree2			
22	cr	StateTree3			
23	mc	StateTree4			
24	mc	StateTree5			
25	mc	StateTree6	0.022		Non-ELL
26	cr	StateTree7			
27	mc	Compost1	0.034		Non-ELL
28	cr	Compost2	0.067	Low	Non-ELL
29	mc	Compost3	0.079	Low	Non-ELL
30	mc	Compost4	0.048		Non-ELL
31	cr	Compost5			
32	cr	Independent#2	-0.109	Moderate	ELL
33	mc	Independent#1			

Notes. cr = constructed response; mc = multiple choice; #=items used in both tests. $p \leq .000$.

While there are no high effect sizes in the 2007 test,Weather5 favoring non-ELLs in the 2006 with a high effect size has a moderate effect size in 2007, and Independent2 favoring ELLs with a high effect size in 2006 has a moderate effect size in 2007. Finally, Boiling6 favoring ELLs with a high effect size in 2006 has a low effect size in the 2007.

Replication

To address the first research question asking whether the 14 items used in both the 2006 and the 2007 tests display the same pattern of DIF administrations, the results suggest that replication is successful for the most part. With SIBTest and with logistic regression, 11 of the 14 common items were consistent as to whether they did or did not contain significant DIF in both years. The comparison of significant items with DIF for both tests can be seen in Table 12. Two exceptions were found with each method. Boiling2 and Boiling3 display significant low DIF in 2006 but not in 2007 with either method. With logistic regression, there were no interactions on the 2007 test and Boiling7 was no longer significant. SIBTest identified Weather2 for negligible significant DIF in 2006 but it was not significant in 2007. Overall, these three items were identified with DIF favoring non-ELLs in the 2006 test but were not identified as having DIF in the 2007 test.

The correlation between the coefficients for 2006 and 2007 was 0.90 with logistic regression and 0.94 with SIBTest. This is a high correlation between the two testing years, considering that the demographic characteristics of ELLs each year

probably vary greatly in both their known (measured) and unknown (unmeasured) characteristics.

Table 12

Replication of Items

Scenario ID	SIBTest Beta Coefficients		Logistic Regression Coefficients		Significant Uniform DIF Favors
	2006	2007	2006	2007	
Boiling#1	0.001	0.005	0.041	0.032	
Boiling#2	0.031*	-0.057	0.095*	-0.011	Non-ELL
Boiling#3	0.055*	0.013	0.068*[a]	-0.005	Non-ELL
Boiling#4	-0.095*	-0.042*	-0.220*	-0.081*	ELL
Boiling#5	0.008	-0.010	-0.033	-0.030	
Boiling#6	-0.150*	-0.084*	-0.305*	-0.177*	ELL
Boiling#7	0.070	0.015	0.211*[a]	0.055	Non-ELL
Weather#1	0.050*	0.035*	0.107*	0.130*	Non-ELL
Weather#2	0.049*	0.024	0.054*	0.040	Non-ELL
Weather#3	-0.001	-0.024	0.004	-0.080	
Weather#4	0.019	0.054	-0.001	0.054	
Weather#5	0.147*	0.099*	0.369*[a]	0.499*	Non-ELL
Independent#1	-0.008	0.009	-0.070	-0.023	
Independent#2	-0.124*	-0.109*	-0.260*	-0.473*	ELL

Note. Items with inconsistent DIF significance were shaded; a=interaction; *= significant at $p \leq .000$.

DIF effect sizes for SIBTest were greater in 2006 than in 2007, such that the high DIF in 2006 was moderate DIF in 2007, the moderate DIF in 2006 was low DIF in 2007, and the DIF with low effect size in 2006 was not flagged for DIF in 2007. While none of the effect sizes were significant for logistic regression, there was also a decrease in the R^2 of the DIF in the 2007 data compared to the 2006 data. The replicated items contain very little significant DIF against ELLs.

DIF Item Comparison by Method

The second research question asks whether the two methods would identify the same items with significant DIF favoring non-ELLs and significant DIF favoring ELLs.

The comparison was made based on the results of item analysis from both methods (see Table 13) as well as the comparison of the testlet analyses (see Table 16).

There was a high agreement between the two methods with regard to which items were flagged and the group favored. The correlation between the method's coefficients on the total 2006 tests was 0.95, and on the total 2007 test was 0.86. When the methods identified the same items for significant DIF they favored the same group. Logistic regression identified non-uniform DIF for constructed response items, for example for Boiling7 on the 2006 test. SIBTest can identify non-uniform DIF for multiple choice items when there are no constructed response items include in the analyses. The items with uniform DIF identified by logistic regression were in agreement with SIBTest regarding which group they favored. It is important to note that because SIBTest cannot analyze any subject with missing data, 10% few cases in 2006 and 21% fewer cases in 2007 were analyzed with the SIBTest method than with the logistic regression method.

SIBTest flagged only two fewer items with significant DIF than logistic regression in the 2007 test. Where both methods flagged the same item as having significant uniform DIF, the methods were in agreement as to which group is favored. Using either method for the both sets of comparisons, about twice as many items favored non-ELLs than favored ELLs. For SIBTest, the 2007 test contained fewer items with significant DIF than the 2006 test, while logistic regression identified about the same number of items with significant DIF for the 2007 test as for the 2006 test. Since logistic regression uses the total test score in place of IRT ability estimates the

sample's characteristics may have influenced the DIF results. In 2007, the sample was larger and there were higher proportions of high SES ELLs than in 2006 (2006 = 7.5%, 2007 = 21.7%). Lower effect sizes were also found with SIBTest, which uses an IRT approach, intended to separate the samples' characteristics from the items' characteristics. The error rate for SIBTest when there are differences in the distributions of ability has been said to be lower than for other DIF methods (Finch & French, 2007; Woods, 2011), but not when the probability of passing is different due to adverse impact (Klockars & Lee, 2008). Most of the research regarding Type I error rates comes from simulation studies rather than real data, so it cannot be said with certainty that irregularities in the sample distributions and other sample characteristics could vary effect size estimates.

Table 13

DIF Significance by Method

2006				2007			
Scenario ID	SIBTest	LR	Favors	Scenario ID	SIBTest	LR	Favors
Puddle1	-0.011	-0.094		Heat1	-0.034*	-0.151*	ELL
Puddle2	0.005	0.071		Heat2	0.013	0.213*	Non-ELL
Puddle3	0.014	0.106*	Non-ELL	Heat3	-0.077*	-0.299*	ELL
Puddle4	-0.099*	-0.265*	ELL	Heat4	0.028*	0.160*	Non-ELL
Puddle5	0.040*	0.144*	Non-ELL	Heat5	-0.011	-0.134	
Puddle6	0.008	0.058		Heat6	-0.037	-0.081	
The birds1	0.007	-0.014		Weather#1	0.035*	0.130*	Non-ELL
The birds2	0.088*	0.094	Non-ELL	Weather#2	0.024	0.040	
The birds3	0.073*	0.109*	Non-ELL	Weather#3	-0.024	-0.080	
The birds4	-0.029	-0.031*	ELL	Weather#4	0.024	0.054	
The birds5	-0.030	-0.140		Weather#5	0.099*	0.499*	Non-ELL
Independent#1	-0.008	-0.070		Independent5	-0.003	-0.041	
Boiling#1	0.001	0.041		Boiling#1	0.005	0.064	
Boiling#2	0.031*	0.095*	Non-ELL	Boiling#2	-0.011	-0.115	
Boiling#3	0.055*	0.068*	Non-ELL	Boiling#3	0.013	-0.010	
Boiling#4	-0.095*	-0.220*	ELL	Boiling#4	-0.042*	-0.157*	ELL
Boiling#5	0.008	0.033		Boiling#5	-0.010	-0.061	
Boiling#6	-0.150*	-0.305*	ELL	Boiling#6	-0.084*	-0.279*	ELL
Boiling#7	0.070	0.211*	Non-ELL	Boiling#7	0.015	0.251	
Weather#1	0.050*	0.107*	Non-ELL	StateTree1	-0.006	-0.107	
Weather#2	0.049*	0.054	Non-ELL	StateTree2	0.002	-0.020	
Weather#3	-0.001	0.004		StateTree3	0.002	0.020	
Weather#4	0.019	-0.001		StateTree4	0.013	0.063	
Weather#5	0.147*	0.369*	Non-ELL	StateTree5	0.022	0.060	
Independent3	0.055*	0.117*	Non-ELL	StateTree6	0.022*	0.604*	Non-ELL
Lettuce1	-0.012	0.008		StateTree7	-0.001	-0.042	
Lettuce2	0.069*	0.219*	Non-ELL	Compost1	0.034*	0.286*	Non-ELL
Lettuce3	0.005	0.017		Compost2	0.067*	0.217*	Non-ELL
Lettuce4	-0.009	-0.047		Compost3	0.079*	0.320*	Non-ELL
Lettuce5	0.029	0.061		Compost4	0.048*	0.312*	Non-ELL
Lettuce6	-0.163*	-0.274*	ELL	Compost5	0.013	0.003	
Independent#2	-0.124*	-0.260*	ELL	Independent#2	-0.109*	-0.473*	ELL
Independent4	0.005	0.087		Independent#1	0.009	-0.023	

Note. #=used in both datasets; *p≤.000.

Testlet Analyses

The testlet analyses address the third research question as to whether any sets of items that refer to a scientific scenario, called a testlet, exhibit DIF when bundled together. Each testlet's scenario contains labeled pictures, some moderately to highly schematic with indications of movement and actions, and some that emphasize details within a larger picture. All but two of the scenarios include data tables, and some solitary items contained data tables, as well. The investigation scenarios entitled The Birds, Puddle, Heat and Boiling contain two page long descriptions. For investigation scenarios such as Heat, there are a couple sentences to introduce the scenario, then a research question, a prediction, a list of materials, a data table, the main visual aid, and an 8 part procedure. The systems scenarios entitled Lettuce, State Tree, Weather and Compost contain one page with three sentence descriptions. In contrast to the Heat scenario, State Tree's visual aid is accompanied by a three sentence description for the scenario.

Logistic regression testlet analyses. LR was used to analyze the differential testlet functioning in order to provide a clearer understanding of the significance of the regression and effect size of the testlets. The items from the testlets were bundled together and the testlet score was used in place of the item score. If any step of the testlet analysis contained a significant step, the testlet was considered significant[5]. The results reveal that the testlets entitled Boiling, Heat, and Independent items without a testlet

[5] The highest significant step's $R^2\Delta$ was reported for any testlet where more than one step was significant.

favored ELLs. The testlets entitled Weather and Compost favored non-ELLs. None of the effect sizes were above negligible; however, the Weather testlet had the highest effect size is in the 2006 test favoring non-ELLs.

Table 14

Logistic Regression Analyses of Testlets

Testlet	2006 $R^2\Delta$	2007 $R^2\Delta$	Favors
Boiling	0	0.001*	ELL
Weather	0.006*	0.002*	Non-ELL
Independents	0.003*	0.006*	ELL
Puddle	0		
The birds	0		
Lettuce	0		
Heat		0.001*	ELL
State tree		0	
Compost		0.006*	Non-ELL

Note. *p $\leq$.000

SIBTest testlet analyses (DBF). The initial DBF for 2006 identified testlets entitled Weather and The Birds as favoring non-ELLs, and identifies the Lettuce testlet and the Independent items as favoring ELLs. For the 2007 DBF, testlets entitled Weather and Compost favored non-ELLs and testlets entitled Heat and Boiling and the Independent items favored ELLs. Except for the independent item bundle, the testlets that favored ELLs are all experiments, while the testlets that favored non-ELLs were all scientific systems. These results may suggest that differences are due to opportunity to learn scientific content around systems.

The testlets with moderate to large effect sizes, with beta coefficients over 0.10 were: Boiling, Weather, Compost and Heat. Table 15 displays each of the SIBTest DBF coefficients. The Weather testlet was significant with a moderate to large effect size in both 2006 and 2007 favoring non-ELLs, while the Boiling testlet favored ELLs in the 2007 test, only.

Table 15

SIBTest Differential Bundling Functioning of Testlets

Testlet	**2006 Beta**	**2007 Beta**	**Favors**
Boiling	0	-0.128*	ELL
Weather	0.302*	0.186*	Non-ELL
Independents	-0.071	-0.078*	ELL
Puddle	0		
The birds	0.126*		Non-ELL
Lettuce	0		
Heat		-0.130*	ELL
State tree		0.083	Non-ELL
Compost		0.282*	Non-ELL

Note. *p $\leq$.001

The initial significance using SIBTest was re-tested using the purification technique by removing the item with the highest DIF to examine whether the bundle was still significant. Only two testlets favoring non-ELLs remained significant after purification—Compost from the 2007 test and Weather in both the 2006 and 2007 tests.

Neither the Weather testlet nor the Compost testlet describes experiments; instead both contain a short approximately three sentence description of a scientific system. The Weather testlet includes a data table and a labeled map, and two of the five items contain

their own visual aids. The Compost testlet only includes a labeled non-schematic drawing of a playground with sections of the playground labeled.

DBF testlet comparison by method. The two methods agreed on significant DBF flagged and the group that DBF favored on 10 of the 12 analyses. SIBTest identified for DBF favoring non-ELLs for two testlets that logistic regression did not flag; these are State Tree in 2006 and The Birds in 2007. Both methods flagged the Boiling testlet for DBF favoring ELLs in 2007 but not in 2006, and the Weather testlet favored non-ELLs in both years (see Table 16).

Table 16

Comparison of Testlet Analyses

	2006		2007		
Testlet	LR $R^2\Delta$	SIB Beta	LR $R^2\Delta$	SIB Beta	Favors
Boiling	0	0	0.001**	-0.128**	ELL
Weather	0.011**	0.302**	0.002**	0.186**	Non-ELL
Independents	0.003**	-0.071*	0.007**	-0.078**	ELL
Puddle	0	0			
The birds	0	0.126**			Non-ELL
Lettuce	0	0			
Heat			0.001**	-0.130**	ELL
State tree			0	0.083*	Non-ELL
Compost			0.010**	0.282**	Non-ELL

Note. *p ≤.01, **p≤ .001

Language-Related DIF

The final research question asked whether items with DIF favoring non-ELLs would correspond to high linguistic complexity coding and whether those favoring ELLs incorporated schematic visual aids. This section will describe the possible

language-related trend in the differentially functioning items for the few released items that contained linguistic complexity. Since logistic regression's effect sizes were all negligible, SIBTest effect sizes are referenced for the duration of the analyses.

Linguistic complexity. The statistical methodology framework guided the data interpretation. Identifying a pattern of coded linguistic features relating to DIF disfavoring ELLs does not mean that every coded feature of the same type (e.g. all academic language) is found to have significant DIF. It is not a simple process to predict which specific words an ELL or group of ELLs will know, due to the high degree of variability between individual ELLs and groups of ELLs. This is also why it is likely that there will be differences in the pattern between the two samples.

Another challenge in identifying linguistic complexity was that there is not one type of complexity per question; so, it was not always obvious why the item might be difficult to understand. Some of the items flagged for DIF against ELLs contain multiple types of linguistic complexity. The coding for linguistic complexity and the raters rankings may shed more light on the DIF in these items.

The most apparent linguistic feature that could be problematic for ELLs in understanding the test item was the use of low frequency words, such as can be seen in The Birds2[6]. The Birds2 requires a familiarity with two low frequency words, "*beak"* and "*pecking*".

---

[6] SIBTest identified The Birds2 as containing significant DIF favoring non-ELLs, however LR did not flag The Birds2 for significant DIF.

3 While observing birds in the neighborhood, Tim noticed that the birds pecking in the grassy areas had different beaks than the birds feeding at the bird feeder in a tree. Describe why birds have different beaks.

In your description, be sure to:

- Identify **two** different types of bird beaks.
- Describe **why** birds need these different types of beaks.

Use words, labeled pictures, and/or labeled diagrams in your answer.

Identify a type of beak: **Describe why a bird needs this type of beak:**
Identify another type of beak: **Describe why a bird needs this type of beak:**

Figure 6. The Birds2 with low DIF favoring non-ELLs

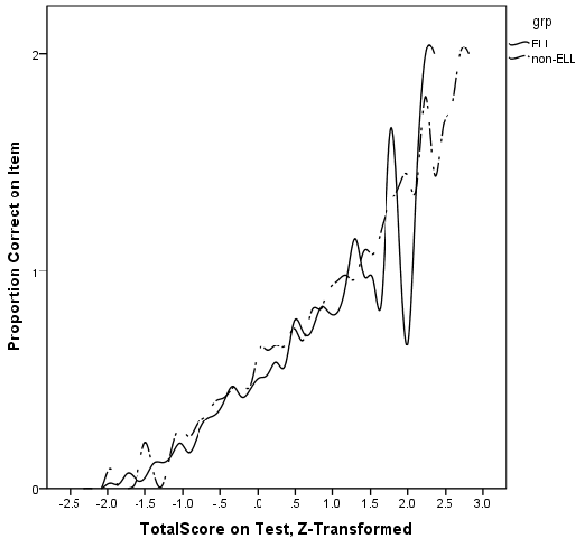


Figure 7. The Birds2 item characteristic curve

Figure 7 shows low DIF for Birds2, and that essentially the low frequency words do not seem to have much negative effect on ELLs' likelihood of receiving at least one point on the item.

Weather5, with the highest DIF against ELLs on both tests, contained a low frequency word "*source"* and a dependent clause in the question stem. The Spanish cognate for *source* is a word less frequently used in common every day speech but a scientific word —"*origin*"—that could be more difficult for non-ELL who are struggling readers. It is important to note that Weather5 does not depend on referencing either the visual aid or the scenario description to correctly answer the question.

4 The grass in each of the four cities is sometimes wet in the morning, even on days without rain. What is the source of the water on the grass?

- o **A.** Water evaporates from the grass.
- o **B.** Water in the air condenses on the grass.
- o **C.** Water on the grass is absorbed by the soil.

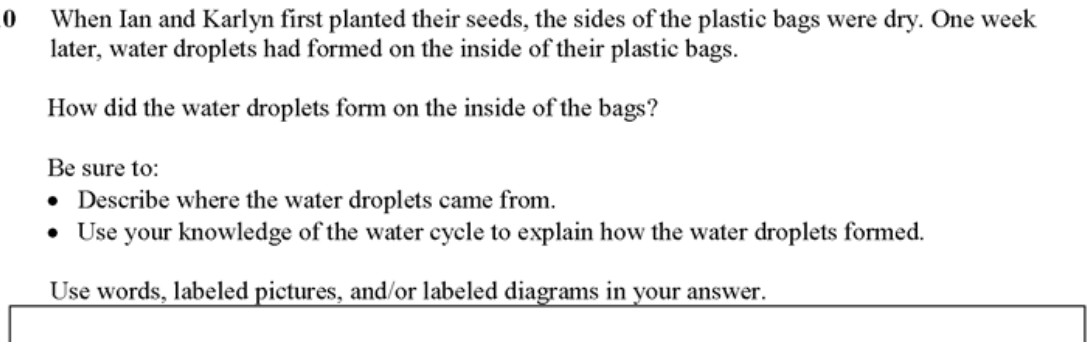
Figure 8. Weather5 with high DIF favoring non-ELLs

The word "*even*" is a multi-meaning word and is potentially confusing when used with the preposition "*on*". Besides the low frequency word use, Weather5 highlights the importance of prepositions in making sense of the answer choices on which the meaning of the sentence relies. The difference in the wording of the answer choices is only subtly different, so that placement of words can change the whole meaning of the answer choices. Logistic regression supports the assumption that this

item was especially difficult to read, revealing a significant three-way interaction between total score, group and reading at $p \leq .001$.

Adverse impact must be considered as an explanation for the DIF against ELLs on Weather5. Performing well on this item depends on having been taught the specific content in the classroom, and then specifically understanding the difference between the terms condensation and evaporation. The high effect size for the DIF favoring non-ELLs on Weather5 in the 2006 sample decreased to a moderate effect size in the 2007 sample. The change points to the effect of opportunity to learn and socioeconomic status on the magnitude of the DIF against ELLs (Boscardin et al., 2005).

It is important to compare Weather5 to another item that addresses the same knowledge. Lettuce3 requires the student to explain and differentiate between water condensation and evaporation in the form of a constructed response but is not flagged for DIF:



10 When Ian and Karlyn first planted their seeds, the sides of the plastic bags were dry. One week later, water droplets had formed on the inside of their plastic bags.

How did the water droplets form on the inside of the bags?

Be sure to:

- Describe where the water droplets came from.
- Use your knowledge of the water cycle to explain how the water droplets formed.

Use words, labeled pictures, and/or labeled diagrams in your answer.

Figure 9. Lettuce3 with no DIF

Lettuce3 is a constructed response item that clearly asks students to explain condensation, and includes several cognates, "*form,*" "*cycle,*" "*explain*," "*diagram,*" and

"*describe*". These last three cognates are present in most of the constructed response items that favored ELLs.

Some test related academic language can be found in items that favor ELLs and some that disfavor ELLs. The phrase, "*which of the following*" and the phrase "*based on*" can be found in multiple test items, including Boiling6 with high DIF favoring ELLs. Birds1 is not flagged for DIF but contains the academic phrase, "*which of the following,*" and the academic words "*gather*" and "*attract.*"

Complex verb phrases are found in many of the items, some with DIF favoring ELLs, such as in Lettuce6 where it states, "*will soon grow,*" and uses the potentially difficult academic word "*provide*."

The conditional tense, is present in items without DIF and one item with DIF against ELLs, Compost1. In Compost1, the conditional "*could*" is repeated in all three answer choices and in the stem question and contains the word "*benefit,*" a cognate. An example of an item without DIF is Independent4 that uses the conditional "*should*" and several cognates.

Accordingly, in some cases the presence of linguistic complexity in the item does not prevent the item from favoring ELLs. For example, Boiling6 and Lettuce6 contain the highest DIF favoring ELLs. Boiling6 contains some obvious linguistic complexity—an academic phrase, "based on," and the conditional tense—but nevertheless favors ELLs. Most likely due to that academic phrasing, conditional tense and length of reading required Boiling6 was ranked moderately to highly linguistically complex by raters. This indicates that, for the conditional tense, ELLs may have grown accustomed to this

linguistic feature. Boiling6 requires referencing the long page of information from the scenario and the schematic visual aid from that testlet (see Figure 10).

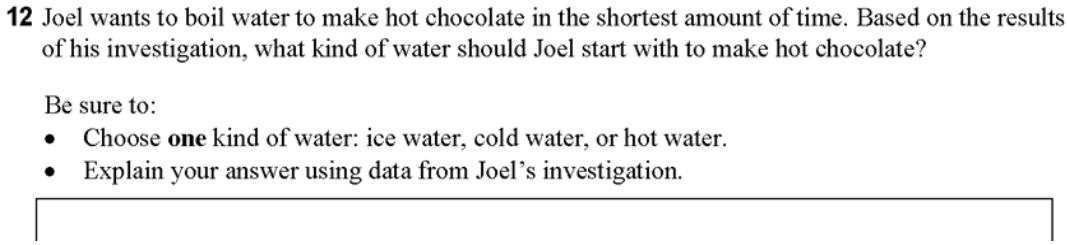


Figure 10. Boiling6 with DIF favoring ELLs

For Boiling6, the question that asks for a conclusion about the experiment and then application of such knowledge to a new situation particularly favors ELLs.

For Lettuce6, again students must make a conclusion about the experiment, apply the knowledge gained to a new problem, and reference the testlet scenario and the schematic visual aid (see Figure 11). This item is ranked as highly linguistically complex by the raters, but within a testlet that is rated low in overall linguistic complexity. This constructed response item contains a potentially difficult interrupted verb: "*will soon grow*" and the future conditional tense. The item characteristic curve for Lettuce6 shows that ELLs tend to have a higher probability of passing the item than non-ELLs (see Figure 12).

11 The lettuce plants will soon grow too big for the plastic bag. What should Ian and Karlyn do to provide their lettuce plants with everything the plants need to live and grow outside the bag?

Be sure to:

- Describe the new environment Ian and Karlyn should design for their growing plants.
- In your design, include **three** things the plants need to live and grow.

Use words, labeled pictures, and/or labeled diagrams in your answer.

Figure 11. Lettuce6 with high DIF favoring ELLs

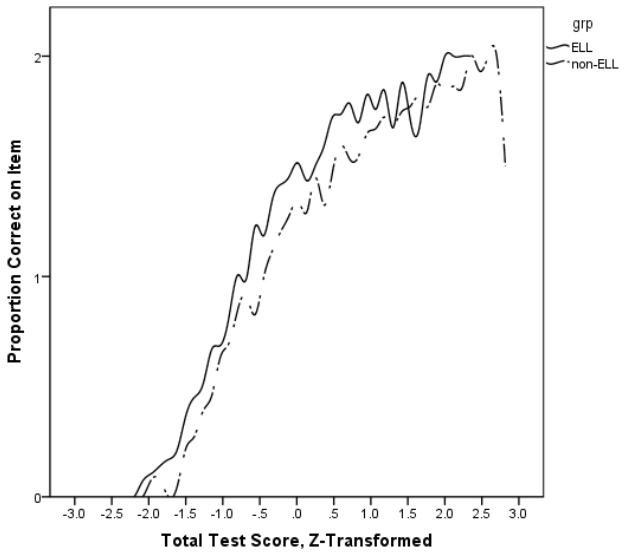


Figure 12. Lettuce6 item characteristic curve

Visual aids. The Birds testlet and the Compost testlet, which favored non-ELLs with the SIBTest method, include the least schematic visual aids in these scenarios. Visual aids were mainly pictorial and included labels for items with little depth or meaning. All but one item from the Compost testlet favors non-ELLs with low or negligible DIF. The Compost testlet assesses students' knowledge about that particular scientific system. The

content within these testlet items and the absence of a diagram or pictorial representation of composting indicate that students are expected to recall the information learned from the class lesson. Adverse impact is likely a strong predictor regarding why ELLs were disfavored by this testlet if ELLs were not thoroughly taught about composting at school.

Although adverse impact may explain the DIF for the Compost testlet, the influence of a potentially distracting non-schematic pictorial visual aid cannot be disregarded. ELLs may have believed that there was something in the visual aid meant to give them clues that would help them answer the items. The items refer to that visual aid, but the visual aid does not depict composting (See Figure 13).

Compost Pile

Directions: Use the following information to answer questions 1 through 5.

Simon's school has an area for a compost pile. A compost pile contains plant waste that can be decomposed. Compost is used in the garden. The diagram below shows the location of Simon's compost pile at his school.

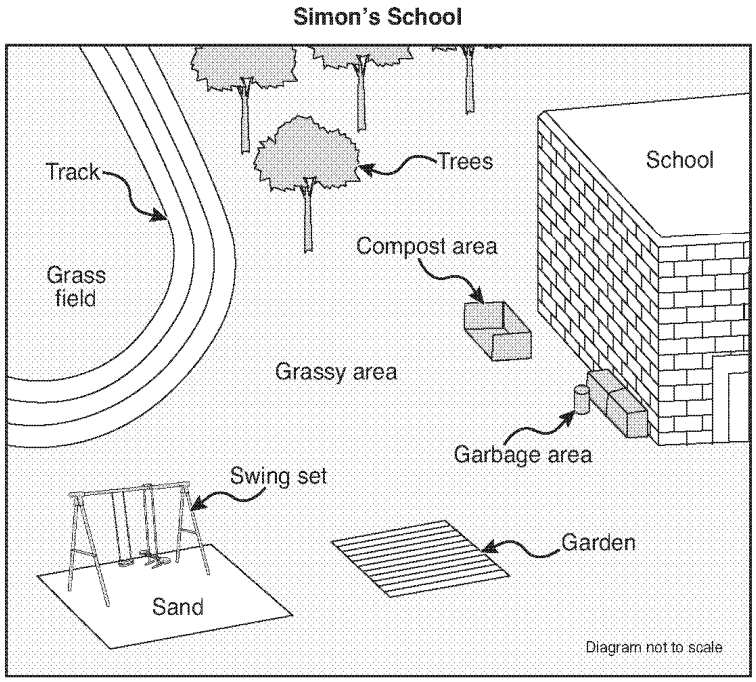


Figure 13. Compost testlet visual aid

Compost3 is rated as the lowest in linguistic complexity by two out of the three raters. It is a very short question with one word responses (Figure 14). There is one academic phrase, "which of the following" and incorporates cognates, "*decomposers*" and "*energy*". Generally, one might expect that pictures support students' comprehension of items. Students can look to the picture when they have trouble making sense of the question. In those cases where the picture shown in Figure 13 did

not scaffold the item the picture may serve as a detrimental distractor, but the low DIF that can be seen in Figure 15 does not support the raters' belief that a schematic visual aid is needed to aid ELLs on this item.

2 Which of the following is food energy for the decomposers in the compost pile?

- **A.** Heat
- **B.** Water
- **C.** Leaves

Figure 14. Compost3 with low DIF favoring non-ELLs

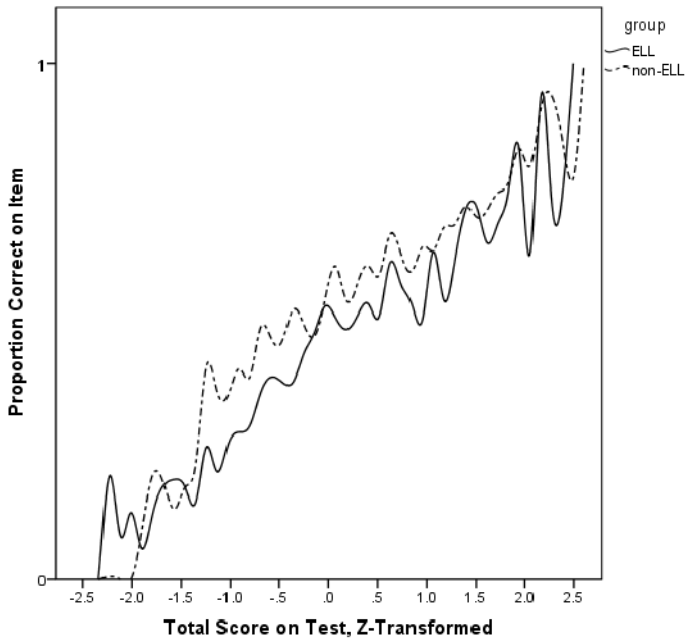
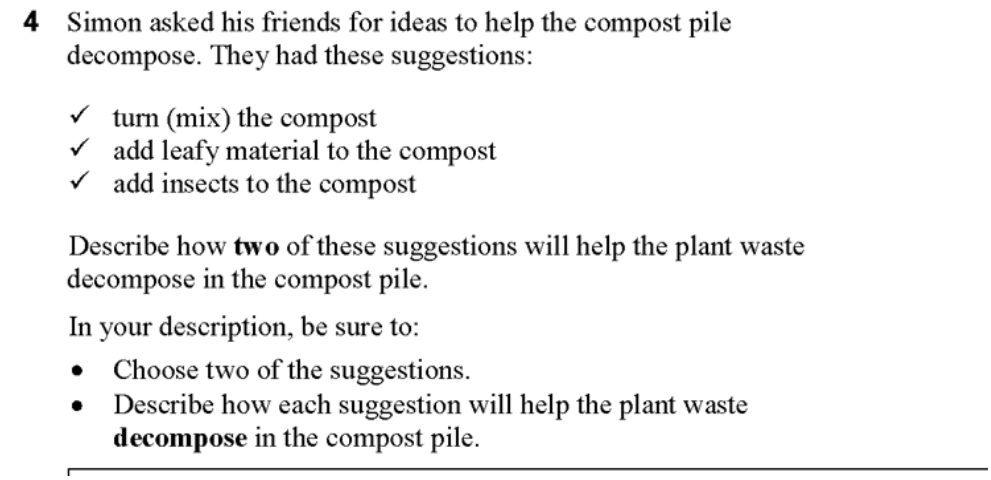


Figure 15. Compost3 item characteristic curve

Compost2 favoring non-ELLs is a constructed response item that initiates a research question about the natural observation. Compost2 requires the students to be creative in their responses and includes vague phrasing "*leafy mater*ial" and the pronoun reference "*these.*" However, "*material,*" "*suggestion,*" "*description*," "*decompose*," "*insects,*" "*ideas,*" and "*compost*" are all cognates. The raters believed it to be the most linguistically complex item from that testlet, possibly due to the length of the question and the lack of scaffolding in the visual aid.



4 Simon asked his friends for ideas to help the compost pile decompose. They had these suggestions:

- ✓ turn (mix) the compost
- ✓ add leafy material to the compost
- ✓ add insects to the compost

Describe how **two** of these suggestions will help the plant waste decompose in the compost pile.

In your description, be sure to:

- Choose two of the suggestions.
- Describe how each suggestion will help the plant waste **decompose** in the compost pile.

Figure 16. Compost2 with low DIF favoring non-ELLs

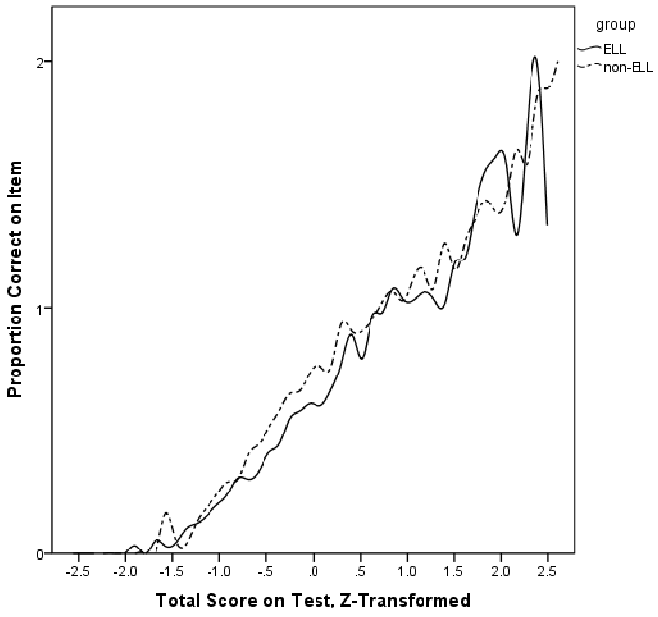


Figure 17. Compost2 item characteristic curve

Both The Birds2 and The Birds3 had low effect sizes favoring non-ELLs and refer to the non-schematic visual aid. The visual aid does not contain a picture of a bird or a labeled beak. Most non-ELLs may be familiar with the word *beak* or the word *pecking* and this is probably the reason that the visual aid does not provide clues about the word's meaning. Again, there is likely a high percentage of adverse impact, if ELLs were not familiar with the answer choices involving short scientific terms, which are cognates. The Birds testlet is a another natural observation of a system that requires students to recall class lessons that ELLs may not have had an equal opportunity to learn.

Still, it is possible that the non-schematic visual aid causes cognitive overload for The Birds3 because students expect the visual aid to help them make sense of the test item.

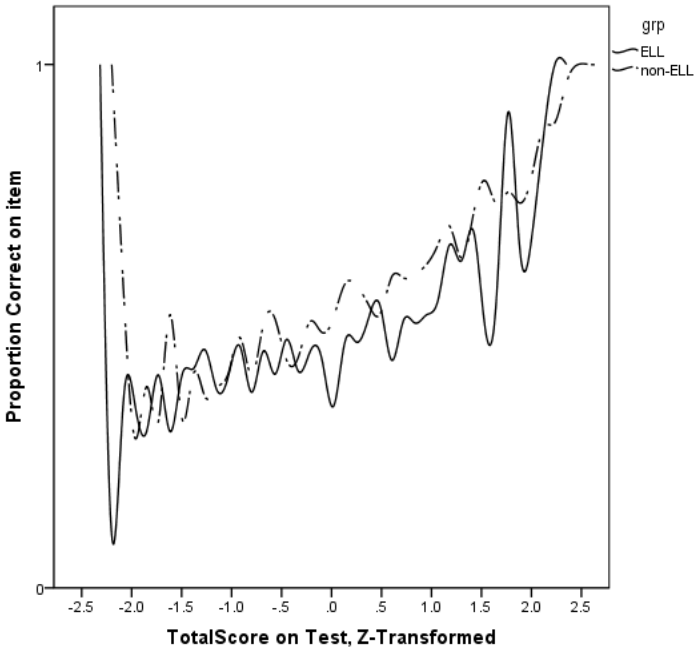


Figure 18. The Birds3 item characteristic curve

2 What does the grass in Tim's yard need to grow?

- o **A**. Soil, water, and insects
- o **B**. Water, sunlight, and air
- o **C**. Nutrients, birds, and roots

Figure 19. The Birds3, with low DIF favoring non-ELLs

Relationship between Raters' and DIF Results

Native Spanish speaking raters who work in the field of education served as experts in ranking the linguistic complexity for each item. Raters provided an overall rank of comprehensibility. The average intraclass reliability for the rank order of the items by linguistic complexity was 0.866, with a 95% confidence interval of 0.809 to 0.908. For the ranking of the testlets, the intraclass correlation between rater one and the other two raters was not significant, while between raters two and three there was a significant Spearman rho of 0.732. The inter-rater correlation between the first and third raters was the lowest, resulting in a correlation of 0.575. The raters were experienced in understanding the difficulties that ELLs experience with regard to academic English comprehension, but they sometimes found it difficult to separate or differentiate between linguistic complexity and academic complexity.

The item factors and linguistic aspects raters addressed tended to fall into three main categories—low frequency words, complex phrases, and scaffolding of understanding for the items. Each rater pointed out unfamiliar words that were not incorporated into visual aids. For example, one rater asked "What is this?" in reference to the words "*pop can*" from the Compost testlet. The raters brought attention to the fact that unfamiliar words can be distracting in the same way that non-schematic visual aids are. Often raters made statements about the items with the expectation that each question should address the visual aid and the scenario. The Birds2, with the low frequency words "*beak*" and "*pecking,*" is ranked as moderately linguistically complex. Reading scale scores were significant in the variance accounted for in responding correctly to the item.

This item was problematic in that ELL status was still a factor after variance accounted for by reading scores had been explained. Two raters commented about the item being confusing and scant on details and noted that this might be difficult for ELLs since neither birds, nor beaks are represented and labeled in the visual aid. Another rater asked why The Birds3, with low effect size favoring non-ELLs did not directly reference any information provided in the scenario. Overall, the raters did not find the visual aid or the scenario for The Birds testlet supportive in answering the items.

Much of the coding for linguistic complexity was identified by raters in their comments, particularly low-frequency words and complex phrasing. The difference in rankings by raters may have been due to differences in prioritizing features of linguistic complexity. One of the raters stated that sentences with conditional verb tense seemed more complex, such as "*what might be*" from State Tree4 or "*would most likely be*" from State Tree6. This is in agreement with what past research has suggested be removed to avoid linguistic complexity (Abedi & Lord, 2001; Abedi, Hofstettler, Baker, & Lord, 2001; Brown, 1999). However, the rankings for linguistic complexity tended to contradict the DIF results. For example, the raters did not agree on the linguistic complexity level of Weather5. The first rater ranked Weather5 and the testlet high in linguistic complexity. The second rater ranked Weather5 moderately linguistically complex, while the third rater ranked it moderately low in linguistic complexity. Both the second and third raters ranked the Weather testlet moderately low in linguistically complex. The Puddle and Lettuce testlets were rated among the least linguistically complex by raters, and the items that ranked high in linguistic complexity actually favored ELLs, such as Boiling6 with

high DIF favoring ELLs. While the average interrater reliability estimate of 0.866 was respectably high, there was little relationship between the ratings and the statistically significant DIF.

The majority of ELLs seemed to be able to comprehend the language used on the science tests. The testlet analyses showed that the items that were differentially difficult for ELLs mainly referred to systems scenarios whereas the items and testlets that favored ELLs tended to refer to the investigation scenarios. The two DIF methods were in agreement as to which items favored either group when identifying an item with significant DIF. I will summarize and discuss the implications of these findings in the next chapter.

Chapter V: Discussion and Conclusions

This study investigated the effect of linguistic complexity on DIF against Spanish speaking ELLs. There were two key findings with regard to this linguistic complexity. First, there was little correlation between the ratings on linguistic complexity and the patterns of DIF. Of further note is that ratings by the native Spanish speaking raters had little to no relationship with the items that were actually flagged for DIF. This is consistent with prior studies looking at whether the items flagged for DIF were also flagged for potential bias in bias and sensitivity reviews (Engelhard, Hansche, & Rutledge, 1990; Ercikan, 2002). Second, very few items stood out as containing linguistically complex features[7]. Many of the language issues found to be problematic in prior research had little bearing on which items had DIF. Overall, linguistic complexity was not a significant issue for this population of students.

The results did raise some important factors that may influence DIF. In this study, as has been found in previous studies examining gender and ethnicity, multiple choice items flagged for DIF tended to favor non-ELL students and constructed response items flagged for DIF tended to favor ELL students. Also, and probably most significant was that, although linguistic complexity did not emerge as a DIF issue for these data, a more serious finding is that DIF results suggest these Spanish speaking students lacked opportunity to learn science content. The majority of items that favored ELL students were tapping into scientific thinking, while the majority of items

[7] This is likely due to the rigorous series of piloting and evaluation processes that items used on Washington state tests must undergo before being administered.

that a favored non-ELL students measured science content knowledge. In what follows, I will elaborate on these findings.

Item Response Type

A significant pattern of DIF based on item type emerged on both tests; ELLs did not do as well on multiple choice items in comparison to non-ELLs with equal ability. All of the multiple choice items had only three response options, providing a fifty percent chance of successfully passing the item if just one of the incorrect options can be eliminated. DeMars and Wise (2010) point out that guessing can often be the reason for DIF, and they surmise that this could be a characteristic of the group, rather than the item. This effect in their simulation study led to finding a lower proportion of DIF in the case where the poorer guessing group also had a lower mean total score. Previous research has shown that constructed response items generally favor minority ethnic groups, while multiple choice items favor Caucasians (Henderson, 2001; Taylor & Lee, 2012; Zenisky, Hambleton, & Robin, 2004). The difference in guessing behavior can artificially inflate the disparity between the groups' actual knowledge of the science material. In addition, it is important to consider cultural differences in the acceptability of guessing and previous experience with multiple choice item tests that might explain why many multiple choice items tend to favor non-ELLs.

All but one of the items that favored ELLs were constructed response items. The constructed response items that required the student to draw a conclusion about a scientific investigation or to generate a new investigation tended to favor ELLs. These types of items assess the scientific thinking and critical thinking skills of students.

Constructed response items are often assumed to be difficult, especially because some degree of knowledge is necessary and there is no option to guess. In addition, English writing proficiency is necessary to receive the highest possible scores on constructed response items. Still, ELLs who are not yet at the same level as their peer in the English writing proficiency were able to successfully communicate their understanding and knowledge. The finding that ELLs received many of their points on constructed response items has important implications since many standardized tests have a high proportion of multiple choice items.

Differences in Content Knowledge (Impact)

A pattern emerged, such that testlets about scientific systems tended to favor non-ELLs and the testlets about investigations tended to favor ELLs. The finding that the systems scenarios tend to favor non-ELLs provides evidence for adverse impact. These natural observation scenarios mainly assess students' content knowledge about living systems. ELLs were likely not given the opportunity to learn that specific content in their classrooms that would give them an equal chance of passing those test items. Alternately, the investigation scenarios focus on science inquiry and scientific application learning requirements; they assess scientific interpretation, analysis, and problem solving skills. All of the test items require science knowledge, but investigation scenarios tend to assess different types of knowledge and different class lessons.

Individual items asking for a summary of an experiment all favored ELLs. ELLs generally seemed to be able to make sense of the active experiments. There are a few items from both tests that proposed new experiments, but most of those types of items

were not flagged for DIF against ELLs, so this is not a differentially difficult task for ELLs.

Linguistic Complexity Ratings and DIF

The coding by raters of each item for linguistic complexity did not yield much correspondence to the DIF favoring non-ELLs. Ranking of items by raters mainly focused on visual aids when they believed that items might be difficult for some ELLs. Raters predicted that lack of visual aids to scaffold understanding, particularly for the scientific systems items, would be problematic.

Raters believed that the most important feature in making sense of items was the visual aid. However, raters also pointed out that some of the visual aids did not provide any help in making sense of specific items. Raters found that what is excluded from the visual aid is more necessary for ELLs understanding than what is included, the reason being that the low frequency words may be more familiar to non-ELLs.

Raters perceptions regarding which items were linguistically complex for ELLs did not correspond to items with DIF disfavoring ELLs. Raters tended to find the shorter multiple choice sentences less linguistically complex than longer constructed response items overall, one rater stated that the multiple choice items with visual aids seemed more difficult and that meaning is conveyed mainly through the few science terms that make up the sentence. Identifying linguistic complexity when multiple choice distractors are linguistically subtle could potentially be difficult. Raters believed that close distractors were more likely to affect ELLs. Contrary to raters' beliefs, longer constructed response

items contained DIF favoring ELLs while the short multiple choice items disfavored ELLs.

The raters were not in full agreement regarding which of the testlets or items were the most linguistically complex. Even though they pointed out where simplified wording could improve the item's comprehensibility and words that might not be familiar to ELLs, the weight given to aspects of linguistic complexity depend on the rater. For example, the Weather testlet did not elicit many comments as to why it is rated high in linguistic complexity. Since raters were not able to identify differentially functioning items for ELLs, it may be necessary to provide raters with careful training to perform linguistic analyses

In particular the linguistic complexity level may have been more difficult for raters to determine on natural observations of science system items due to the demands and conventions of science language and science testing. The items that refer to scenarios about natural observations also happened to contain the least schematic visual aids and the most disciplinary language. These items were also chosen by raters to be more linguistically complex. However, the items in these testlets depend more heavily on previous knowledge gleaned from class lessons. Raters believed more schematic visual aids were necessary for the natural observation testlets to provide language scaffolds for ELLs. Specifically, raters felt the visual aids provided were inadequate in defining low frequency words found in six of the test items. Raters confided that an improved visual aid would be beneficial to ELLs familiar with the science lessons but not proficient in English to increase the likelihood of receiving at least partial credit on

constructed response items. Although adverse impact likely explains a high portion of the DIF favoring non-ELLs, those ELLs who had the opportunity to learn the material might still benefit from an improved visual aid.

The minimal amount of complex phrasing pointed out by raters is an encouraging finding with respect to past research showing that ELLs scores and some non-ELLs scores improved after simplifying the items' language (Abedi & Lord, 2001). This indicates that recommendations from researchers have influenced test developers into creating items with low complexity.

Of significant note is that none of the raters made specific mention of Spanish cognates that may aid ELLs in understanding the items. This is likely due to the emphasis placed on identifying linguistic roadblocks to understanding rather than focusing on why some items might be less linguistically complex. Cognates were incorporated in some of the shorter multiple choice items that raters deem to be less linguistically complex.

The linguistic complexity raters' main concern, beyond the importance of schematic visual aids, was that, in general, a scenario must ground the content for ELLs. For example, the item in the Compost testlet that proposed a new experiment did not require students to reference the visual aid or the text from the scenario. If ELLs are asked to build upon knowledge that references a scenario they must first have a full and clear understanding of that scenario to expound upon it.

Visual Aids

The test item raters believed that the scaffolding effect of visual aids, or lack thereof, would be a significant factor for some items in determining the linguistic complexity for ELLs. Each testlet contains visual aids in the form of a diagram, data table, and/or labeled pictures. Since two of the three least schematic visual aids pertained to items from systems scenarios that assessed science content knowledge, Weather and Compost, the results are inconclusive as to their level of linguistic scaffolding. For these two scenarios with the least schematic visual aids—those that did not include any abstraction, motion, or zooming in on details—their teslets tended to favor non-ELLs. For the most part the testlets with schematic visual aids either do not result in DIF or favored ELLs, but they were the testlets that pertained to an investigation. Although sometimes labeled drawings may help both non-ELLs and ELLs in place of using detailed photographic pictures of objects on the test, non-schematic visual aids has not been shown to help ELLs when those pictures are not schematic (Martiniello, 2009).

In this research, it seems as though non-schematic visual aids did not provide enough clues to the meaning of unfamiliar words for ELLs with lower English proficiency. The inclusion of these schematic visual aids may help ELLs overcome the difficulties associated with items containing low frequency words. Unfortunately, this research was unable to determine the importance of visual aids, and only leads to further questions. For example, are labels for objects not referenced in the items distracting? Alternately, does including a schematic visual aid portraying the objects referenced in the item help ELLs with low frequency words? First, future research must focus on

identifying exactly when visual aids are appropriate to include with any given test item. Then, schematic visual aids will need to be piloted with a diverse sample of ELLs to assure their effectiveness to clarify some confusion.

Martinello (2009) concentrates on schematic visual aids as abstract representations that can create meaning. Whereas, Solano-Flores and Wang (2011) describe the many features that visual aids may contain. Specifically, visual aids can contain features that indicate movement or provide depth and perspective using background features. Currently, research is attempting to identify the features of a visual aid that are the most appropriate for types of test items in order to scaffold reading. Sometimes it is easily apparent why some visual aids are less helpful than others, for example, if the visual aids lack the subjects being described in the text. For those ELLs who need more information to make sense of those test items, this oversight may result in not passing the items. On the other hand, including a detailed representation of the object described in the text in the visual aid does not necessarily solve this problem due to the likelihood of differences in interpretation of the visual aid, an increased cognitive load or when ELLs do not have the content knowledge required to correctly answer the test item.

Creating visual aids for test items must be a careful constructive process that takes into account cultural differences. The results of this dissertation do not provide detailed information about when to use visual aids or the different types of interpretations that can be made from visual aids. It wouldn't be possible to make pronouncements on the effectiveness of specific visual aids for test use without directly asking ELLs for their interpretations and comparing visual aid features from different cultures. Research that

specifically focuses on visual aids for culturally varied groups will likely be able to more clearly prescribe the types of images and styles of imagery that can be effective for use on science tests. At this time, insufficient empirical evidence has been gathered to guide test developers in knowing when visual aids are needed and when they are not.

Linguistic Complexity Coding

With regard to the coding of linguistic complexity there are limited findings. The presence of low frequency words in one item was the only easily identifiable linguistic problem for the ELLs in this study. Although there were a couple of other items with low frequency words, they were not flagged for DIF against ELLs. Only the Birds2 item requires students to know a key word that is not pictured or labeled, so the ELLs who are not familiar with that low frequency word may not be able to receive even partial credit on that test item.

There was no strong pattern for linguistic complexity in items that were flagged for DIF disfavoring ELLs. As was surmised, personal pronouns, academic phrasing and complex verb phrases simultaneously were found in certain test items. However, none of these linguistic features were consistently found in items that disfavored ELLs or that were not flagged for DIF. The results did not support recommendations for cautious use of conditional tense because it was present in items that displayed negligible or low effect size DIF favoring non-ELLs, but it was also present in items without DIF. It is possible that many ELLs had already learned to recognize these types of linguistic complexity and academic phrasing and that appear so often in the test items. Alternately, the linguistic

complexity may not have hindered ELLs understanding of the question being asked of them.

Sentences comprehension can sometimes depend on many factors working together such as sentence structure, word choice, and context. Prepositions that are specifically used with certain verbs or nouns are one example of how sometimes one word can change the whole meaning of the sentence, making it more difficult to interpret in some sentences if other linguistic complexity is present. At other times, when the full meaning of the sentence is understood, perhaps conditional tense or prepositions will not make much difference in comprehension. This supposition would need to be examined more in isolation to determine for which students and in which context each linguistic feature is potentially confusing or problematic.

The results indicated that most of the Spanish-speaking ELLs were able to understand the majority of the features of linguistic complexity examined in these test items. This is likely due to the linguistic conventions that dictate how academic language is used in science and the type of language that can be expected. The science lesson content sets the parameters for the language used in the items, and because science has its own terminology and ways of using language, test items tend to contain more complex language and terminology than other subject areas. This could explain why raters had difficulty, at times, distinguishing between the linguistic complexity and the item difficulty. Scientific terminology and phrasing may seem linguistically complex, but there may be very few ways of expressing the same information in simpler terms while maintaining the original meaning consistent with scientific practice.

Cognates. The presence of Spanish cognates did not explain the pattern of DIF results favoring ELLs, or a lack of DIF in items with many cognates. Most items incorporate several Spanish-English cognates, but the presence of so many cognates did not prevent the item from being problematic for ELLs. Still, being aware of Spanish-English cognates may be the most helpful asset on science tests for Spanish speakers. It could be that in this dissertation, as was found in Martiniello's (2008) study, many ELLs with low English proficiency levels could make the connection between the cognates and what they reference. These tests use a lot of scientific language, which does not seem to differentially be a hindrance to understanding the test items for ELLs. Most likely the language on the test may mirrors the language used in fifth grade science classes.

Reading. Reading comprehension skills do not account for English proficiency level, but the correlation with struggling readers and low SES sheds light on DIF for items with some features of linguistic complexity. Logistic regression results suggest that reading proficiency has an influence on students' performance on approximately half of the test items. This is an interesting finding because many of the items included common terms for some of the scientific language, most likely to accommodate struggling readers.

It must be stated that such categorical distinctions of low SES and poor reading skills are too broad to clarify what aspects of the item are confusing to ELLs. Some ELLs scored high on the reading proficiency test and others received low scores on the reading proficiency test. It is possible that some of the ELLs who scored low on reading in English may also score low on a Spanish reading test. However, those who do and don't read in Spanish are still working toward improving their English proficiency.

Lack of opportunity to learn is likely the main cause of the DIF identified; however, reading skills may explain DIF for some of the items as well, bringing into question whether the lack of adequate scaffolding for struggling readers is of greater concern. It could be that the DIF on the science test more strongly related to reading level than to not being a native English speaker. The interactions terms between reading scores, group and total science score and the interaction between total science score and reading scores for the items with the DIF against ELLs are good examples of the complexity of determining sources of DIF.

DIF Replication and Methods Comparison

One research question was concerned with uncovering DIF in science tests over two years of testing and using two DIF methods. The replication of the DIF pattern for items on both tests was relatively reliable, such that the majority of items flagged for DIF in the first test were also flagged in the second test. This suggests that the items function consistently; those items with DIF against ELLs in the first sample maintained their DIF and the items with DIF against non-ELLs in the first sample maintained their DIF.

There was a slight decrease in DIF identified in the second year of testing. For SIBTest the effect sizes decreased when the item was flagged for DIF in the second administration. Of the common items, the three items with low or negligible DIF on the 2006 test did not exhibit DIF on the 2007 test. The total proportion of practically significant DIF was lower on the second test, as well. While 11 of the 15 items flagged for DIF in 2006 were practically significant using SIBTest only 7 of the 13 items

flagged for DIF in 2007 were practically significant. The great reduction in effect size on the 2007 test, especially for those items that appear on both tests is of note, in part because both samples are quite large. Sample characteristics are also likely mitigating factors on the DIF effect sizes reported. In the 2006 sample, where low SES ELLs were the overwhelming majority of ELLs, the SIBTest DIF effect sizes were greater. In the 2007 test, where there was a significantly lower proportion of low SES ELLs, DIF effect sizes decreased. The differences in the samples' demographics may partly explain these issues with effect size reduction. First, mean test scores for ELLs were higher in 2007 and second there was a much greater proportion of high SES ELLs in 2007. A possible outcome from the increase in ELL mean scores was that there was slightly more overlap in the distribution of scores for the focal and reference group. This may go further in explaining both the decrease in effect size and the decrease in items identified with practically significant DIF in 2007.

It is important to consider the contribution of the non-common items in each test and their effect on the total score estimates. The second test was likely to have more accurate total scores due to the negligible or low effect size of DIF in the non-common items. These data suggest that, due to approximately half as much practically significant DIF on the 2007 test, the total test scores were more accurate estimates of ELLs' true scores. The conclusion that adverse impact on the 2006 test may have increased the effect sizes should not be overlooked. Adverse impact is important for ELL educators to contemplate for the incoming ELLs in the following year, particularly for low SES students and at low SES area schools.

Comparison of Methods

It can be difficult to estimate the amount of error present in the data from empirical studies. Thus, researchers suggest applying more than one DIF method to increase confidence in the results (Fidalgo, Ferreres & Muñiz, 2004; Hambleton & Jones, 1994; Shealy & Stout, 1993). The two DIF methods in this dissertation are generally in agreement as to which items and testlets favor either group. Logistic regression identified 1 more item on each test for DIF than SIBTest, but the main difference in the outcome of the two methods relates to effect size. SIBTest flagged DIF items with moderate and high effect sizes; however, logistic regression flagged those same items for DIF with negligible effect sizes. When the methods identified the same items for significant DIF the group favored was in agreement. On the 2006 test there were 5 items where one method identified an item for significant DIF that the other method did not, but on the 2007 test LR identified only one more item than SIBTest.

SIBTest differential bundle functioning (DBF) of the testlets revealed a balance of testlet DIF. There were three significant testlets (Weather, The Birds and Compost) that favored non-ELLs, and two testlets (Heat and Boiling) and the bundled Independent items that favored ELLs. The testlets that favored ELLs were scientific investigation scenarios that measures scientific thinking and reasoning. In contrast, Weather, The Birds, and Compost testlets asked students to generalize their content knowledge to new systems. Logistic regression confirmed this pattern of differential testlet functioning for all but The Birds testlet.

Differential testlet functioning suggests the possibility of a second dimension on which the two groups differ, and that a multidimensional construct is being assessed (Camilli & Shepard, 1994). ELLs and non-ELLs are matched on total test scores; however, the total scores do not account for differences in ability or knowledge on the second dimension. Items that display DIF are sensitive to a second dimension, on which ELLs and non-ELLs differ, that affects the likelihood of responding correctly to the science items. This second dimension could be the specific content assessed in the systems scenarios. The DBF on the Weather and the Compost testlets in 2007 is most likely due to differences in content knowledge.

The testlets that favored ELLs, Heat and Boiling on the 2007 test, indicate that ELLs tend to do better on scientific investigation scenarios. This is true with respect to the constructed response items asking for conclusions and designing new experiments that also favor ELLs. Presumably, if ELLs had learned about the Weather and Compost systems, then constructed response items that require them to generalize their knowledge about this content may have favored them. The Independent items also favored ELLs on both tests, but the reason for this is not clear. The response types of Independent items were approximately half multiple choice and half constructed response. These items do not reference a scenario and are unrelated in content to each other.

The lack of meaningful effect sizes for the DIF identified with logistic regression in the first test, where SIBTest identified a few items with high DIF, is only somewhat surprising. In other studies that use logistic regression, there has been little more than negligible DIF identified. In particular, Hauger and Sireci (2008) used logistic regression

to compare eighth-grade science items from the Trends in International Mathematics and Science Study (TIMSS); their results comparing dual language testing in three different countries revealed one item with moderate DIF from the 48 multiple choice items examined.

As data from this dissertation includes polytomous items, logistic regression was used because it measures non-uniform DIF for polytomous items and has the additional benefit of permitting additional predictors to be included in the DIF analysis. Although logistic regression is a flexible method it presents a set of difficulties. Researchers have mentioned that it is cumbersome to make interpretations for polytomous items (French & Miller, 1996; Miller & Spray, 1993; Su & Wang, 2005). Many researchers do not include polytomous items in their DIF analyses, likely because it is time consuming and complicated.

Logistic regression DIF simulation studies with samples sizes over 5,000 comparing results where there are differences in group ability distributions are difficult to find. Effect sizes for logistic regression were particularly suggested to avoid negligible DIF that would be significant in very large sample sizes. Fidalgo, Ferreres and Muñiz (2004) define *conservative criterion* as dismissing items not identified for DIF with more than one method. This criterion increases the likelihood of Type II errors but minimizes the likelihood of Type I errors. SIBTest's high level of agreement with logistic regression seems to indicate that the majority of items identified for DIF by logistic regression should have at least small effect sizes. Instead, the logistic regression effect sizes calculated were all well below practical significance. Judging from these results

and those of other studies that compared DIF methods, the effect sizes determined for logistic regression may be too conservative in decreasing Type I error at the cost of increasing Type II errors with a sample containing these characteristics.

Another reason for the use of effect sizes is that unequally distributed ability groups can lead to Type I errors. There have been mixed results regarding the effect of unequal group ability differences on significant DIF (Demars, 2010; Jodoin & Gierl, 2001; Kristjansson et al., 2005). When examining only item significance levels, Kristjansson et al. (2005) found that with logistic regression Type I error was more affected by increased item discrimination and unequal group sample sizes than by unequal between group ability. Joidon and Gierl (2001) were mostly concerned that large sample sizes would yield Type I error due to the use of the chi-square statistic flagging small differences which are not likely practically significant. They found a trend such that as sample size and the percentage of DIF on the test increased the Type I error increased more when group abilities differed than when they did not differ. However, when the DIF was balanced for the groups being compared Type I error was not as affected by unequal between group abilities.

French and Maller (2007) concluded that more research guidelines are needed for logistic regression effect size estimates. They also found that purification did not improve Type I error. In simulation studies comparing the results of DIF methods, Hidalgo and López-Pina (2004) reveal that logistic regression identifies more items with DIF than the MH method but that the logistic regression effect size estimates are not sensitive to DIF that was present. Acar and Kelecioğlu (2010) who performed DIF

analyses using 25 Turkish science items reported that while other methods flagged items for DIF with moderate effect sizes logistic regression revealed only negligible effect sized DIF.

Limitations

One of the limitations of this study is that, unlike simulation studies, using real data means that the true score for any student is unknown. Logistic regression uses the total score as a substitute estimate for true content knowledge in order to compare the two groups of subjects. This can be problematic when the ability distributions of the two groups being compared are not equal because then there is a high probability that impact is present. The effect of adverse impact can be seen in the difference in the strength of the DIF for the replicated items in the second year of testing when score distributions were more similar and the testlets about living systems which favored non-ELLs.

One reason it is difficult to uncover the sources of the DIF favoring non-ELLs is that many items have a unique combination of features. The probabilistic nature of language referred to by Solano-Flores (2008) is the underlying reason that, even when test items are fully dissected, making general statements about which parts of speech, sentence length, or verb tenses to avoid cannot be recommended to eliminate the DIF against ELLs. Researchers are limited in how much we can know about how students make sense of the test material due to the language style or use, as a result of cultural differences. It is likely that many of the causes of DIF against ELLs could be tied to the huge amount of diversity and complexity in the background characteristics of ELLs that are incalculable or unavailable to researchers. In particular, the result cannot be accepted

to apply to all ELLs due to the high degree of variation represented. As stated previously, even though this study examined only one language group, any group of Spanish-speaking ELLs comprise a wide variety of cultural, demographic, and experiential differences. Therefore the results of this study can only generalize to Spanish speaking ELLs in Washington state. Figuring out and removing language that causes difficulty in comprehension for some ELLs does not resolve the problem for all ELL students (Solano-Flores, 2008).

For example, the unreleased item Independent3 on the 2006 test contains a pictorial visual aid of snack foods and favors non-ELLs. At first glance, it would seem the two main properties of this item predispose it to favor ELLs because it is an independent item not related to a scenario and is a constructed response type. There are several possible explanations for the DIF against ELLs on this independent item. It is possible that ELLs' culture made it difficult to interpret the labels and pictures. Solano-Flores and Wang (2011) found that pictures are also open to different cultural interpretations and as such may be misinterpreted. In addition, Independent3 is the only item on the test that covers this science knowledge, leaving open the possibility that adverse impact could have also contributed significant variance to the DIF identified. In this case, there are several explanations for what aspect of the item contributes to the DIF including cultural or experiential reasons, limited non-schematic pictures, opportunity to learn or some combination of these factors.

Lack of opportunity to learn some areas of focus, may be the cause of DIF, particularly for those items that were about natural systems. If the content is not taught in

school, the student cannot be expected to know the material. Due to the majority of ELLs that are low SES in the 2006 test, it is likely that impact is an explanation for some of the DIF against ELLs. These results show that the same items might not contain practically significant DIF when there is a change in the sample's demographics. The decrease in DIF effect sizes for the replicated items in the 2007 test together with the increase in high SES ELLs and higher mean ELL scores supports the idea that adverse impact contributed to some of the variance in the DIF identified against ELLs on the 2006 test.

Impact signifies that the item might actually be measuring true group differences due to unequal opportunities to learn the tested material. In this case, the testlet bundling analyses provide evidence for adverse impact because the scenarios focused on the content knowledge were more difficult for ELLs. If groups of students do not have the opportunity to learn the content knowledge about a particular scientific phenomenon they are likely to perform poorly on all the items that refer to that scenario.

One difficulty that has been mentioned often by researchers is the need to have a well-defined sample in order to make generalizations and to find patterns in understanding. When performing item analyses it is important to focus on word frequency and familiarity to groups with different language proficiency and how they can contribute to DIF (Kok, 1992). This study is limited by the fact that little is known about the cultural breakdown of these Spanish-speaking ELLs, for instance their Spanish dialects and ancestral countries of origin.

Another important limitation regarding the sample is the possibility of misclassification of ELL status, specifically English proficient students classified as

ELLs. The operational definition of Spanish speaking ELLs in this study was confined to those whose primary language spoken at home was Spanish. While level of English reading proficiency on a measure was used as a covariate, ELLs' levels of English proficiency were unknown. In some cases, some ELLs who spoke only Spanish at home may have transitioned out of English as a second language classes. It could be that many of the ELLs may have spent years in mainstream English classrooms so as to have reached an equivalent level of English proficiency as their native English speaking peers. This will limit the application of these results in recognizing which item features are linguistically complex for Spanish speaking ELLs who have a lower level of English proficiency.

ELLs' second language knowledge is affected by a multitude of difficult to measure factors and therefore can be quite inconsistent. Solano-Flores and Li (2009b) found that the interaction for person, item and dialect accounted for the highest error variance in item scores. With so much linguistic and cultural variation it is difficult to generalize about the linguistic features of the items that are affecting the majority of Spanish-speaking ELLs. In other words, due to high within group variability it is difficult to pinpoint vocabulary or references that definitely lie within the scope of their experiences.

Implications

Using more than one method to analyze DIF is beneficial in confirming DIF results, specifically when the methods are different in their matching variable approach, type of procedure and calculations, as in this case with SIBTest and logistic regression

which are employed for this study. Logistic regression has been accused of over-identifying items with DIF and therefore requiring conservative effect size estimates that reduce the likelihood of Type I error (Hidalgo & López-Pina, 2004; Jodoin & Gierl, 2001). The results from the current study suggest that logistic regression does not identify more items for DIF than does SIBTtest when sample sizes are large and a quadratic term is included. Logistic regression's current effect size guidelines may not be sensitive enough to for comparisons to DIF identified by other methods. Still, logistic regression is instructive if information about several predictors can be utilized to help explain the DIF identified.

A further recommendation is for researchers to incorporate more constructed response items in the tests. Most but not all constructed response items flagged for DIF favored ELLs. Constructed response items may provide more accurate information regarding whether students know the material being tested or can think scientifically. However, a test of solely constructed response items would not necessarily eliminate DIF. Moreover, it is important that more DIF analyses include polytomous items in ELL studies, paying particular attention when these items do not favor ELLs. Constructed response items that contain DIF favoring non-ELLs may reveal more information about the sources of DIF. In this study, constructed response items that did not favor ELLs tended to be part of a testlet that contained DIF due to adverse impact. Even though writing skills add another complexity to the item analysis, constructed responses can provide more information about what is understood and what needs clarification.

There is a limited scope of generalizations that can be drawn from the results concerning the linguistic features that might be the source of DIF against ELLs due to the probabilistic nature of language and the lack of linguistic complexity in the items. This is an obstacle in being able to provide precise recommendations about how to phrase the question within an item or the types of complex sentences that especially confuse ELLs. To uncover the best way to word an item, ELLs would need to be involved in the test development process. For instance ELLs can be interviewed and asked to explain why the pilot items confused them. To investigate this question, students with a variety of dialectical, national and cultural backgrounds should be represented. Using more technical English words which look, sound and have similar meaning to Spanish words may aid in understanding, but cognates are not always as accommodating as might be expected.

Future Directions

Future research on DIF for ELLs will need to ensure that the sample is representative of the variety of ELLs, including high SES ELLs, and in doing so special attention must be paid to the effect a mainly low SES ELL sample can have on DIF that is identified. Researchers have stressed the need to examine DIF for low SES students who have been shown to have lower overall test scores (Abedi & Lord, 2001; Krashen & Brown, 2005). Results from a sample that compares primarily low SES ELLs to high SES non-ELLs will need to consider that high DIF could be confounded with SES.

In this study, low SES ELLs are the lowest scoring group, as well as by far the majority of ELLs in the first sample in particular. Accounting for SES as a covariate

could be one way to clarify whether or not the source of DIF is impact due to SES. Further investigating only those DIF items suspected for impact by examining SES at the school level may shed more light on the issue of impact. Another solution to confirm findings when the variance in SES is uncontrolled may be to replicate DIF results with a smaller SES balanced sub-sample.

An additional direction for future research would be to validate the adverse effect of linguistically complex features in items on students' overall test performance. In this study, the DIF methodology uncovered which items were differentially more difficult for ELLs. The next step is to outline the impact of potentially biased test items, in particular, whether ELL status influences the pass-fail rate on the entire test. A few questions arise in consideration of the practical significance of uncovering linguistic complexity: What percentage of ELLs is adversely affected by linguistically complex features, and which ELLs are most affected? Do the number of items that are differentially more difficult for ELLs lead to lower test scores for the entire group?

The complexity involved in generalizing to ELLs' knowledge and experiences cannot be overstated; therefore, not involving ELLs in the test development process is dismissive of their needs and their growing numbers in many states. Regardless, what has most often been mentioned in DIF research comparing ELLs and non-ELLs is that they are an often a forgotten or neglected set of participants.

More research is needed to determine the sources of DIF that has already been uncovered in state tests, primarily concentrating on items that might be expected to favor the focal group but contain DIF against the focal group, and those same items when they

do not contain DIF. Test developers must include ELLs in the process of creating the test items from the very beginning and recruit ELLs when they pilot those test items. They can begin by investigating the high DIF constructed response items that favor non-ELLs and the responses to those items. This research should address better question development by comparing ELLs' responses and questions, as well as uncovering causes in the gaps in understanding.

References

Abbot, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing, 24*(7), 7-36.

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher, 33*, 4-14.

Abedi, J. (2008). *Linguistic modification: Part I—Language factors in the assessment of English language learners: the theory and principles underlying the linguistic modification approach.* Washington, DC: LEP Partnership.

Abedi, J. (2009). Computer testing as a form of accommodation for English language learners. *Educational Assessment, 14*, 195-211.

Abedi, J., Courtney, M., Leon, S., Kao, J., & Azzam, T. (2006*). English language learners and mathematics achievement: A study of opportunity to learn and language accommodation* (CSE Tech. Rep. No. 720). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Gandara, P. (2006). Performance of English language learners as a sub-group in large scale assessment: Interaction of research and Policy. *Educational Measurement: Issues and Practice, 25*(4), 36-46.

Abedi, J., & Herman, J. (2010). Assessing English language learners' opportunity to learn mathematics: Issues and limitations. *Teachers College Record*, *112*(3), 723-746.

Abedi, J., Hofstettler, C., Baker, E., & Lord, C. (2001). *NAEP performance and test accommodations: Interactions with student language background* (CSE Tech.

Rep. No. 536). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*, 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* (CSE Tech. Rep. No. 429). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Bailey, A., 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 Tech. Rep. No. 663). Los Angeles, CA: UCLA Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

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.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Aguirre-Muñoz, Z., & Baker, E. L. (1998). *Improving the equity and validity of assessment-based information systems* (CSE Tech. Rep. No. 462). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Aguirre-Munoz, Z., Parks, J., E., & Benner, A. (2006). *Consequences and validity of performance assessment for English language learners: Conceptualizing & developing teachers' expertise in academic language* (CSE Tech. Rep. No. 700). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Angoff, W. H. (1989). Context bias in the test of English as a foreign language. Educational Testing Service.

August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English language learners. *Learning Disabilities Research & Practice, 20*(1), 50-57.

Beretvas, S., & Walker, C. M. (2012). Distinguishing differential testlet functioning from differential bundle functioning using the multilevel measurement model. *Educational and Psychological Measurement*, *72*(2), 200-223.

Betts, J., Bolt, S., Decker, D., Muyskens, P., & Marston, D. (2009). Examining the role of time and language type in reading development for English language learners. *Journal of School Psychology, 47*, 143-146.

Bolt, D. M. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement, 37*, 307-327.

Bolt, D. & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTest detection procedure. *Behaviormetrika, 23*(1), 67-95.

Boscardin, C., Aguirre-Munoz, Z., Stoker, G., Kim, J., Kim, M., & Lee, J. (2005). Relationship between opportunity to learn and student performance on English and algebra assessments. *Educational Assessment*, *10*(4), 307-332.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Carlo, M. S., August, D., McLaughlin, B., Snow, C., Dressler, C., Lippman, D. N., Lively, T. J., & White, C. E. (2004). Closing the gap: Addressing the vocabulary needs of English-language learners in bilingual and mainstream classrooms. *Reading Research Quarterly, 39*(2), 188-215.

Chang, H.-H., Mazzeo, J., & Roussos, J. (1996). Detecting DIF for polytomously scored\ items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33,* 333-353.

Chen, X., Ramirez, G., Luo, Y. C., Geva, E., & Ku, Y. (2011). Comparing vocabulary development in Spanish- and Chinese-speaking ELLs: The effects of metalinguistic and sociocultural factors. *Reading and Writing*,. doi: 10.1007/s11145-011-9318-7

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

Clauser, B. E., & Mazor, K. M., & Hambleton, R. K. (1993). The effects of purification of matching criterion on the identification of DIF using the Mantel-Haenszel procedure. Applied Measurement in Education, 6(4), 269-279.

Collier, V. P. (1987). Age and rate of acquisition of second language for academic purposes. *TESOL Quarterly, 21*, 614-671.

Collier, V. P. (1995). A synthesis of studies examining long-term language minority student data on academic achievement. *Bilingual Research Journal, 16,* 187-212.

DeMars, C. E. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement*, *70*(6), 961-972.

DeMars, C. E., & Wise, S. (2010). Can Differential Rapid-Guessing Behavior Lead to Differential Item Functioning? *International Journal of Testing*, *10*(3), 207-229.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and Mantel-Haenszel method. *Applied Measurement in Education, 2*(3), 217-233.

Douglas, J. A., Roussos, L. A., & Stout, W. (1995). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*(4), 465-484.

Engelhard, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education*, *3*(4), 347-60.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessment. *International Journal of Testing, 2*(3/4), 199-215.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*(3), 301-321.

Fidalgo, A. M., Ferreres, D., & Muñiz, J. (2004). Liberal and conservative differential item functioning detection using Mantel-Haenszel and SIBTEST: Implications for type I and type II error rates. *The Journal of Experimental Education, 73(1)*, 23-39.

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement, 67*, 565-582.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33,* 315-332.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection**.** *Educational and Psychological Measurement, 67*(3), 373-393.

Friend, J. I., & Degen, E. (2007). Middle-level reform: The introduction of advanced English and science courses. *Journal of Advanced Academics, 18*(2), 246-276.

Fry, S. (2007). How far behind in mathematics and reading are English language learners: A report. Washington, DC: the Pew Center. Retrieved from: http://www.pewhispanic.org/2007/06/06/how-far-behind-in-mathematics-and-reading-are-english-language-learners/

Gierl, M .J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement, 38*(2), 164-187.

Gierl, M. J., Gotzmann, A., & Boughton, K. A. (2004). Performance of SIBTEST when the percentage of DIF items is large. *Applied Measurement in Education, 17,* 241-264.

Golia, S. (2010). The effects of the presence of items affected by uniform differential item functioning on Rasch measure. Retrieved from: http://ssrn.com/abstract=1819087

Gómez-Benito, J., Hidalgo, M. D., & Padilla, J-L. (2009) Efficacy of effect size measures in logistic regression: An application for detecting DIF. *European Journal of Research Methods for the Behavioral and Social Sciences, 5*(1), 18-25.

Graves, M. F., Boettcher, J. A., Peacock, J. L., & Ryder, R. J. (1980). Word frequency as a predictor of students' reading vocabularies. *Journal of Reading Behavior, 12*(2), 117-127.

Griffin, K. A., Allen, W. R., Kimura-Walsh, E. (2007). Those who left, those who stayed: Exploring the educational opportunities of high-achieving Black and Latina/o students at magnet and nonmagnet Los Angeles high schools *Educational Studies: Journal of the American Educational Studies Association, 42*(3), 229-247.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Hambleton, R. K., & Jones, R, W. (1994). Comparison of empirical and judgmental procedures for detecting differential item functioning. *Educational Research Quarterly, 18*(1), 21-36.

Hauger, J. B., & Sireci, S. G. (2008). Detecting differential item functioning across examinees tested in their dominant language and examinees tested in a second language. *International Journal of Testing, 8*(3), 237-250.

Henderson, D. L. (2001, April). *Prevalence of gender DIF in mixed format high school exit examinations.* Paper presented at the Annual Conference of the National Council on Measurement in Education, Seattle, WA.

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,* 904-915.

Holland, W. P., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hunt, E., & Agnoli, F. (1991). The Whorfian hypothesis: A cognitive psychology perspective. *Psychological Review, 98*(3), 377-389.

Ilich, M. O. (2011, December). *English language learners and differentia item functioning on mathematics items.* Paper presented at the conference for the Washington Education Research Association, Seattle, WA.

Jodoin, G. M., & 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.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education and Praeger.

Klein, J. R., & Jimerson, S. R. (2005). Examining ethnic, gender, language, and socioeconomic Bias in oral reading fluency scores among Caucasian and Hispanic students. *School Psychology Quarterly, 20*(1), 23-50.

Klockars, A. J., & Lee, Y. (2008). Simulated tests of differential item functioning using SIBTEST with and without impact. *Journal of Educational Measurement, 45*(3), 271-285.

Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners*. Washington, DC: Council of Chief State School Officers.

Kopriva, R. J., Emick, J. E., Hipolito-Delgado, C. P., & Cameron, C. A. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educational Measurement: Issues and Practice, 26*(3), 11-20.

Kopriva, R., & Sexton, U. M. (1999). *Guide to scoring LEP student responses to open-ended science items.* Washington, DC: SCASS LEP Consortium Project, American Association for the Advancement of Science.

Kopriva, R. J., Wiley, D. E., & Emick, J. (2007). Status 2007: Inspecting the validity of large-scale assessment score inferences for ELLs and others under more optimal testing conditions—Does it measure up? Retrieved from ERIC database (ED497497)

Kristjansson, E., Aylesworth, R., Mcdowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*(6), 935-953.

Krashen, S., & Brown, C. L. (2005). The ameliorating effects of high socioeconomic status: A secondary analysis. *Bilingual Research Journal, 29*, 185-196.

Krippner, S. (1966). *Vocabulary guide of cognate words in Spanish and English*. Maimonides Hospital of Brooklyn, NY. ERIC Accession Number ED015088.

Kuhl, P. K. (2011). Early language learning and literacy: Neuroscience implications for education. *Mind, Brain and Education, 5*(3), 128-142.

Lawton, S. B. (2012). State education policy formation: The case of Arizona's English language learner legislation. *American Journal of Education, 118*(4), 455-487.

Lee, S. K. (2002). The significance of language and cultural education on secondary achievement: A survey of Chinese-American and Korean-American students. *Bilingual Research Journal*, *26*(2), 327-38.

Lee, Y., Cohen, A., & Toro, M. (2009). Examining type I error and power for detection of differential item and testlet functioning. *Asia Pacific Education Review*, *10*(3), 365-375.

Linacre, J. M., & Wright, B. D. (1989). Mantel-Haenszel DIF and PROX are equivalent! *Rasch Measurement Transactions, 3*, 52-53. Retrieved from: http://www.rasch.org/rmt/rmt32a.htm

Luykx, A., Lee, O., Mahotiere, M., Lester, B., Hart, J., & Deaktor, R. (2007). Cultural and home language influences on children's responses to science assessments. *Teachers College Record, 109*(4), 897-926.

Mahon**,** E. A. (2006). High-stakes testing and English language learners: Questions of validity. *Bilingual Research Journal, 30*(2), 479-497.

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*, 160-179.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1993). The effects of purification of matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6*(4), 269-279.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement 30*, 107-122.

Minicucci, C., & Olsen, L. (1992). *Programs for secondary limited English proficient students: A California study. FOCUS Number 5. Occasional papers in bilingual education*. National Clearinghouse for Bilingual Education, Washington, DC.

Nagy, W. E., Garcia, G. E., Durgunoglu, A. Y., & Hancin-Bhatt, B. (1993). Spanish-English bilingual students' use of cognates in English reading. *Journal of Reading Behavior, 25*, 241-259.

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

Pei, L. K., & Li, J. (2010). Effects of unequal ability variances on the performance of logistic regression, Mantel-Haenszel, SIBTEST IRT, and IRT likelihood ratio for DIF detection. *Applied Psychological Measurement, 34*(6), 453-456.

Rivera, C., & Stansfield, C. W. (2004). The effect of linguistic simplification of science test items on score comparability. *Educational Assessment, 9*(3&4), 79-105.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedure for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105-116.

Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalized formula for Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24*(3), 293-322.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*(4), 355-371.

Scheuneman, J. D., & Slaughter, C. (1991). Issues of test bias, item bias, and group differences and what to do while waiting for the answers. Educational Testing Service. Retrieved from ERIC database. (ED400294)

Schmidt, W. H., Cogan, L. S., & McKnight, C. C. (2011). Equality of educational opportunity: Myth or reality in U.S. schooling? *American Educator, 34*(4), 12-19.

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), 105-106.

Shaw, J. M. (1997). Threats to the validity of science performance assessments for English language learners. *Journal of Research in Science Teaching, 34*(7), 721-43.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias*. Baltimore, MD: Johns Hopkins University Press,.

Siegel, M. A. (2007). Striving for equitable classroom assessment for linguistic minorities: Strategies for and effects of revising life science items. Journal of Research in Science Teaching,44(6), 864-881.

Sireci, S. G., Han, K. T., & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment, 13,* 108-131.

Smith, W. F. (2009). Language-related DIF in the WASL mathematics test. (doctoral dissertation). University of Washington, Washington. Retrieved May 2, 2010, from Dissertations & Theses @ University of Washington WCLP. (Publication No. AAT 3370570).

Soares, T. M., Goncalves, F. B., & Gamerman, D. (2009). An integrated Bayesian model for DIF analysis. *Journal of Educational and Behavioral Statistics, 34* (3), 348-377.

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

Solano-Flores, G. (2012, April). *Probabilistic approaches to examining the impact of items' linguistic features on student performance.* Paper presented at the Annual conference of the National Council on Measurement in Education, Vancouver, Canada.

Solano-Flores, G., Lara, J., Sexton, U., & Navarrete, C. (2001). *Testing English language learners: A sampler of student responses to science and mathematics test items.* Washington, DC: Department of Education, Council of Chief State School Officers.

Solano-Flores, G., & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice, 25*(1), 13-22.

Solano-Flores, G., & Li, M. (2009a). Generalizability of cognitive interview-based measures across cultural groups. *Educational Measurement: Issues and Practice*, *28*(2), 9-18.

Solano-Flores, G., & Li, M. (2009b). Language variation in the testing of English Language Learners, native Spanish speakers. *Educational Assessment, 14,* 180–194.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*(5), 553-73.

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.

Solano-Flores, G., & Wang, C. (2011, April). *Conceptual framework for analyzing and designing illustrations in science assessment: Development and use in the testing of linguistically and culturally diverse populations.* Paper presented at the Annual Conference of the National Council on Measurement in Education, New Orleans, LA.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test-performance of African-Americans. *Journal of Personality and Social Psychology, 69*(5), 797-811.

Stoneberg, B. D., Jr. (2004). *A study of gender-based and ethnic-based differential item functioning (DIF) in the spring 2003 Idaho Standards Achievement Tests applying the simultaneous bias test (SIBTEST) and the Mantel-Haenszel chi square test.*

Unpublished manuscript, The University of Maryland-College Park and the National Center for Educational Statisitics.

Su, Y. H., & Wang, W. C. (2005). Efficiency of the Mantel, generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. *Applied Measurement in Education, 18,* 313-350.

Taylor, C. S., & Lee, Y. (2011). Ethnic DIF and DBF in Reading Tests with Mixed Item Formats. *Educational Assessment, 16,* 1-34.

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education, 25*(3), 246-280.

Wang, M., Park, Y., & Lee, K. (2006). Korean-English biliteracy acquisition: Cross-language phonological and orthographic transfer. *Journal of Educational Psychology*, *98*(1), 148-158.

Wang, C., & Solano-Flores, G. (2011, April). *Illustrations with graphic devices in large-scale science assessments: An exploratory cross-cultural study of students' interpretations.* Paper presented at the Paper presented at the Annual Conference of the National Council on Measurement in Education, New Orleans, LA.

Williams, J. A. (2001). Classroom Conversations: Opportunities to learn for ESL students in mainstream classrooms. *The Reading Teacher, Vol. 54*(8), 750-757.

Wolf, M. K., Kim, J., Kao, J. C., & Rivera, N. M. (2009). *Examining the effectiveness and validity of glossary and read-aloud accommodations for English language learners in a mathematics assessment* (CSE Tech. Rep. No. 766). Los Angeles,

CA: UCLA Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Wolf, M., & Leon, S. (2009). An investigation of the Language Demands in Content Assessments for English Language Learners. *Educational Assessment*, *14*(3/4), 139-159.

Woods, C. M. (2011). DIF testing for ordinal items with Poly-SIBTEST, the Mantel and GMH tests, and IRT-LR-DIF when the latent distribution is nonnormal for both groups. *Applied Psychological Measurement 35*(2) 145-164.

Wu, A. D., & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing, 6*(3), 287-300.

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

Young, J. W., Holtzman, S., & Steinberg, J. (2011). *Score comparability for language minority students on the content assessments used by two states.* Research Report ETS RR–11-27.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment, 9*(1/2), 61-68.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for*

binary and Likert-type item scores. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense. Retrieved from http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html

Zwick, R., & Thayer, D. T. (2003). *An empirical Bayes enhancement of Mantel-Haenszel DIF analysis for computer-adaptive tests.* Newtown, PA: Law School Admission Council.

Appendix A: Coding Scheme

Coding Features	Descriptions
Non-Verbal	Composite score from the following non-linguistic features:
Tables	Does the item refer to a table?
Diagrams	Does the item refer to a diagram
Schematic pictures	Does the item contain a picture representing a function, scheme or state of matter?
Non-schematic pictures	Does the item contain a picture of a non-schematic object
Linguistic complexity	Composite score from the following individually counted features:
Multi-meaning words, colloquialisms	Number of words that can have multiple meanings, idioms, colloquialism
Academic language	Number of academic words, not scientific/technical in nature
Relative pronouns	Number of relative pronouns used (e.g., "which", "whom")
Complex verb phrases	Presence and number of complex verb phrases (e.g., "have been going")
Passive voice	Presence of passive voice (e.g., "was given"),
Conditional clauses	Presence and number of conditional clauses (e.g., "if", "could", "would")
Cognates	Number of Spanish-English cognates
Structural component	
Independence of item	Independent item (not related to a scenario) or scenario-based item
Item response type	Multiple-choice or written (constructed) response

Appendix B: Linguistic Complexity Coding Examples

Coding for Weather5

4 The grass in each of the four cities is sometimes wet in the morning, even on days without rain. What is the source of the water on the grass?

- o **A.** Water evaporates from the grass.
- o **B.** Water in the air condenses on the grass.
- o **C.** Water on the grass is absorbed by the soil.

Multi-meaning words, colloquialisms: even on

Academic language: source

Relative pronouns: no

Complex verb phrases: no

Passive voice : no

Conditional clauses: yes

Cognates: evaporates, condenses, absorbed

Coding for Boiling6

12 Joel wants to boil water to make hot chocolate in the shortest amount of time. Based on the results of his investigation, what kind of water should Joel start with to make hot chocolate?

Be sure to:

- Choose **one** kind of water: ice water, cold water, or hot water.
- Explain your answer using data from Joel's investigation.

Multi-meaning words, colloquialisms: no

Academic language: based on

Relative pronouns: his

Complex verb phrases: be sure to

Passive voice: no

Conditional clauses: should

Cognates: investigation, explain, results, using, data
