IMPACT OF GENDER AND LINGUISTIC BACKGROUND ON ENGLISH LANGUAGE ART TEST: DIFFERENTIAL ITEM FUNCTIONING

by

Zahya Fawzi Ahmed

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Literacy Studies

Middle Tennessee State University

August 2019

Dissertation Committee:

Dr. Jwa K. Kim, Co-chair

Dr. Eric Oslund, Co-chair

Dr. Amy Elleman

Dr. Ying Jin

I dedicate this work to my parents and my family for their support, sacrifice, and endless love.

Acknowledgement

I would like to express my deepest appreciation to my chairs Dr. Kim and Dr. Oslund and my committee members Dr. Elleman and Dr. Jin for their guidance and their patient. I would also like to extend my deepest gratitude to my supervisor Dr. Kim who believed in me and gave me a chance to achieve my goals. He played a decisive role in guiding me through my entire journey. I feel fortunate to have the chance to learn from him and choose the path of statistics.

I would also thank all my friends and colleagues who helped me learn how to work as a team and share knowledge. This four-year journey has been full of joy, fear, and anxiety; therefore, I am so thankful to be surrounded by such supportive family, friends, and professors who helped me through it, and I gratefully acknowledge their assistance.

Abstract

The purpose of this study was to investigate the impact of gender and ELL status in students' performance, and to compare the consistency of the results of the differential item functioning (DIF) determining from applying Mantel-Haenszel (MH) procedure and the DIF results determining from applying item response theory (IRT)-likelihood ratio statistics. In literature, DIF has been applied as a statistical tool to investigate bias items against subgroups in a particular population and to evaluate the fairness and validity of the educational and/or psychological assessments. Gender and different linguistic backgrounds are the most studied variables in DIF literature. A version of an English language art test was examined for evidence of DIF based on gender difference and different linguistic backgrounds (ELL vs. non-ELL). DIF analyses were implemented through six sets: (1) gender differences (male vs. female), (2) ELL status (non-ELL students vs. ELL students), (3) gender within non-ELL group (male non-ELL vs. female non-ELL), (4) gender within ELL group (male ELL vs. female ELL), male cross linguistic background groups (male non-ELL vs. male ELL), and female cross linguistic background groups (female non-ELL vs. female ELL). The sample of this study consisted of students from 7th (*N* = 12,658) grade from 11 states who took the standardized English Language Arts (ELA) test based on common core state standards (CCSS) at the beginning of the 2014-2015 academic year. Data were analyzed using classical test theory (CTT) and IRT to detect DIF.

The results revealed one DIF item when the MH procedure was applied. On the other hand, IRT-likelihood ratio flagged 4 items out of 34 test items. The results from both detection methods were inconsistent. Implications of these analyses were discussed in accordance with the previous findings for providing linguistically diverse students in regard to their gender with effective literacy programs to meet their academic needs.

Keywords: Differentiate Item Function (DIF), Item Response Theory (IRT), Mantel-Haenszel (MH), likelihood ratio, English Language Learners (ELL), gender, common core state standards (CCSS)

TABLE OF CONTENTS

LIST OF TABLES x

LIST OF FIGURES xi

CHAPTER I: INTRODUCTION 1

Background of the Study 1

Gender 3

English as a Second Language 4

Common Core State Standards (CCSS) 5

Psychometrics 6

Classical test theory (CTT). 7

Item response theory (IRT). 8

Differential item functioning (DIF). 10

Historical development of DIF 11

Methods of detecting DIF. 12

Statement of the Problem 13

Research Questions 13

Significance of the Study 14

CHAPTER II: LITERATURE REVIEW 16

High-stakes Educational Tests 16

The efficiency and utility of high-stakes testing. 19

Test fairness and psychometrics. 20

DIF as a Tool of Examining Test Fairness 20

DIF detection methods. 21

CTT-based methods (non-IRT). 22

IRT-based methods. 26

DIF types. 34

Uniform. 34

Nonuniform. 34

Gender and Linguistic Background as a Function of Differential Test Performance... 35

Gender differences in standardized assessment. 35

Linguistic background. 37

Conclusion 39

CHAPTER III: METHOD 41

Participants 41

Materials 42

Procedures 44

Matching criterion. 44

Dimensionality. 45

Model-fit analysis. 45

DIF analysis 46

Mantel-Haenszel procedure. 47

Likelihood ratio statistics. 48

Calculating the DIF effect size. 49

CHAPTER IV: RESULTS 51

Introduction 51

Descriptive Statistics 52

Dimensionality 53

CTT Analysis 54

IRT Analysis 55

DIF analysis 57

The MH procedure. 58

IRT-likelihood ratio. 58

Gender groups (male vs. female). 59

Male cross non-ELL and ELL groups (male non-ELL vs. male ELL). 59

Female cross non-ELL and ELL groups (female non-ELL vs. female ELL). 60

ELL status (Non-ELL vs. ELL). 60

ELL -gender (Male ELL vs. Female ELL). 61

Non-ELL -gender (male non-ELL vs. female non-ELL). 62

The Agreement between MH Procedure and Likelihood Ratio 64

Summary 65

CHAPTER V: DISCUSSION 67

Introduction 67

Discussion of the Findings 68

Research question one 68

Research question two 69

Research question three 69

Research question four 70

Research question five 70

Implications for Practice 72

Limitation of the Study 73

Recommendations for Further Research 74

Conclusion 74

REFERENCES 77

APPENDICES 99

APPENDIX A: IRB APPROVAL 99

LIST OF TABLES

Table 3. 1 Item categories based on CCSS 43

Table 4. 1 Descriptive statistics for each set of DIF analysis 52

Table 4. 2 CTT and IRT parameter estimations for the whole test.................................... 56

Table 4. 3 DIF results for all six comparison sets of analysis using likelihood ratio. 59

Table 4. 4 parameter estimations for references groups and focal groups in each comparison set 64

LIST OF FIGURES

Figure 4. 1 the scree plot for the eigenvalues of the factors which clearly starts from the second eigenvalue. 54

Figure 4. 2 Item response function for male cross non-ELL and ELL groups. 60

Figure 4. 3 Item response function for ELL status. 61

Figure 4. 4 Item response function for gender within ELL group. 62

Figure 4. 5 Item response function for gender within non-ELL group............................. 63

CHAPTER I: INTRODUCTION

Background of the Study

The era of high-stakes standardized assessments has started early in the twentieth century when the educators and policy-makers tried to gather information about students' academic achievement for instructional and educational assistance to ensure students' success (Walsh & Betz, 1985). During the second half of the twentieth century, there were profound educational changes resulting in the high-stakes standardized tests become a trend of collecting data and information nationwide and making the comparison between different states (Marzano, 2018). The high-stakes assessment is a broad term used to describe the mechanism of (1) appraising performance of students, teachers, and school systems through a sequence of standardized assessments, (2) allowing public comparison, and (3) measuring accountability (Jones, Jones, & Hargrove, 2003). In 1969, the National Assessment for Education Statistics (NAEP) was founded to evaluate students' academic competence and to monitor their progress for better educational outcomes at the national, state, regional, and district levels (NAEP, 2009). NAEP provides annual report cards and statistics regarding students' progress based on high-stakes assessment results. Therefore, the reliance on these tests to make high-stakes decisions serves many public goals such as raising the academic standards, holding students and teachers accountable, and increasing school enrollment rates (Marzano, 2018).

Salvia, Ysseldyke, and Witmer (2012) discussed the meaning of assessment in a broad sense as a process of different data-collection techniques such as testing, observations, interview, and personal judgment. They referred to the word assessment as a general concept of collecting data from students' performance in different evaluating sets to make decisions on proper learning instructions and educational assistance to ensure students' success. Also, Reynolds and Livingston (2012) referred to the consequences of using high-stakes testing on the accountability system and depending on test outcomes in order to evaluate to which extent students are achieving their learning goals, and teachers are teaching what they are supposed to teach. Therefore, the word test or testing will be used throughout this study. Educational testing has always been a part of evaluating students' achievement in the educational process; however, there is a growing concern for understanding the results of these assessments (Bott, 1996).

Translating the outcomes of any test into comprehensible information can be very beneficial for the learning process. Each group of students has specific learning needs that may require precise learning instructions (Reynolds & Livingston, 2012). However, before using test data to make any decision, the quality of the test should be validated by applying statistical and psychometric procedures. Consequently, validity and reliability are used in terms of test quality in developing, adapting, and/or translating any measure (Walsh & Betz, 1985). Validity refers to that the test is measuring what it is designed to measure (Reynolds & Livingston, 2012). Reliability refers to the consistency of the test scores if it is re-administered under the same conditions to the same group of examinees (Gall, Gall, & Borg, 2007). Johnson and Johnson (2002) assumed that the high-stakes tests underestimate a certain group of students due to other factors rather than their academic competence such as their socioeconomic status, linguistic background, ethical group, and/or gender. Therefore, studying the impact of these variables on test results may give a deeper understanding of the response patterns.

Gender

Gender is a preferable tested variable because, in the validity studies, the researchers usually apply some psychometric procedures to examine the invariance in the instrument in regard to gender, ethnic, cultural, and/or language variables. Gender was always related to measurement differences. A recent report released by the U. S. National Assessment of Educational Progress (NAEP, 2017) indicated that there was a consistent gap between males and females across all grade levels. Therefore, investigating gender bias has been an important part of the validation process undertaken for any instrument. Theoretically, Abedlaziz, Ismail, and Hussin (2011) devoted the reasons beyond gender differences to the unfamiliarity of some topics on the test, offensive topics, and/or the role of the dominant stereotype during the test setting. However, many studies investigated gender differences using statistical procedures without analyzing the test content.

Gender test bias is a very well-addressed issue. The postulated difference in response patterns between males and females has been detected more likely in aptitude tests than in educational tests (Tanner, 2001). The gender achievement gap has been documented through many research studies (Hope, Adamson, McManus, Chis, & Elder, 2018; Kurt, Karakaya, Safaz, & Ates, 2014; Reardon, Kalogrides, Fahle, Podolsky, & Zarate, 2018). The advantages of one gender type over the other are also significantly manifested in aptitude and educational tests; however, the reasons for different response patterns are still unclear (Reynolds & Livingston, 2012). Discussing these reasons is beyond the scope of this research study.

English as a Second Language

Currently, U.S. classrooms are very diverse. Students are multilingual, multiethnic, and multicultural; consequently, they have different learning styles (Marzano, 2018). The changes that occur to the students' population in school bring new challenges to test makers to design a fair and reliable test for all students (Walsh & Betz, 1985). English language learners (ELL) are one of the most growing minority groups in the U.S. public schools. Based on the U. S. Department of Education report of 2000-2015, the percentage of ELL students who have enrolled in public elementary and secondary schools is about 10%, and this number is growing fast as a large number of new immigrants have arrived to the U.S. in the last few years (U. S. Department of Education, 2018). The U.S. Department of Education classifies the students who have limited English proficiency as ELL students based on English language proficiency test results. These students are receiving language assistance programs to attain English proficiency and meet the academic standards that all non-ELL students are supposed to meet (Reynolds & Livingston, 2012).

Limited English proficiency could result in poor educational outcomes which affect the students' educational evaluation (Lambert, Garcia, January, & Epstein, 2017). Reynolds and Livingston (2012) suggested some strategies that might be used in the case of assessing ELL students. They listed three possible alternatives: (1) using translated versions of the English test, (2) using nonverbal test, or (3) providing a qualified bilingual examiner or a translator. However, these suggestions have a negative impact on test validity and reliability, they are very expensive to implement, and they are time-consuming process (Reynolds & Livingston, 2012). Improving ELL educational achievement has been a concern since the beginning of the current century. The NAEP stated that there is a significant achievement gap between non-ELL and ELL students, and this gap increases in higher grade levels. August, Shanahan, and Escamilla (2009) reviewed and summarized a synthesized work sponsored by Office of English Language Acquisition (OELA), U. S. Department of Education, and Institute for Educational Sciences (IES) to deliberate five domains related to teaching and learning English as a second language. One of these domains was to investigate and develop the language and literacy assessment of ELL students. Therefore, the main focus of this study was to evaluate high-stakes test results in order to improve the understanding and interpretation of these results.

Common Core State Standards (CCSS)

The challenge of global competition in the field of education has led to developing internationally benchmarked standards that provide guidelines for teachers to ensure students’ success at the end of the academic school year (Haynes, 2011). An educational initiative of CCSS has been sponsored by the National Governors Association (NGA) and the Council of Chief State School Officers (CCSSO) to provide an equal learning opportunity for all students nationwide and to formulate educational standards that align with the modern technology of the twenty-first century (CCSSO, 2010). The k-12 CCSS were released in 2010 to create consistent standards for English language arts and literacy in history/social studies, science, and technical subjects that guarantee access to high-quality education and prepare students for college and careers

(Morrow, Tracey, & Del Nero, 2011). Evans, Evans, and Mercer (1986) referred to learning language arts as a fundamental stage to learning other skills since language arts consist of communication skills of listening, speaking, writing, and reading skills that are essential for other disciplines. Under the No Child Left Behind Act (NCLB, 2002), all students should be afforded equal opportunities for high-quality education based on the state requirements that meet the federal standards. The NCLB act is considered as a shift towards ranking students based on norm-referenced tests. Norm-referenced tests refer to measurements that compare students' scores to a hypothetical average score (Gall et al., 2007). Since then, about forty-eight states have adopted these academic standards to achieve their specific learning goals (CCSS Initiative, 2010).

Phillips and Wong (2010) argued that to bring the CCSS to the next level, the testing system should also be improved to align with these standards. Therefore, in recognition of the importance of the high-stakes tests that reflect the CCSS guidelines, analyzing data collected from these tests should be done with respect to certain psychometric criteria to ensure the fairness of the test to all students.

Psychometrics

There is a growing concern among educators in terms of test reliability and validity (Reynolds & Livingston, 2012). Test validation or test fairness is a widely used term to refer to psychometric properties that should be applied to ensure test validity and reliability (DeMars, 2010). Therefore, psychometric theories can provide an adequate interpretation of the test results and the measurement quality (Furr, 2011). There are two major psychometric theories in psychological and educational fields that have been used for data analysis. These two theories are classical test theory (CCT) and item response theory (IRT).

Classical test theory (CTT). The CTT or true score theory is based on the examinee's observed score as the unit of focus represented by a true score added up to an error score, and it is considered as a test-dependent theory (De Ayala, 2009). CTT estimates the examinees' latent trait as a function of his performance in test items as the true score if the examinee takes an infinite number of the same measure. Therefore, the true score cannot be obtained from the test data; as a result, the error score related to it is unobtainable as well, which leads to the difficulty of attaining the model-fit analysis to the data set (Lord, 1980). Applying CTT to test data requires estimating the true score and the measurement error score to compare them with the observed performance on a given test (Fan, 1998). The true score is a hypothetical unobservable score that would be equal to the observed score if the measurement error is zero (Reynolds & Livingston, 2012). The above information can be expressed mathematically with the following equation,

$$X = T + E, \tag{1}$$

where X is the observed score, T is the true score, and E is the error. Moreover, CTT is based on three major assumptions: (1) the true score and the associated measurement error are uncorrelated, (2) the measurement errors of one test are uncorrelated with measurement errors of parallel form, and (3) the true score of one test is uncorrelated with measurement error of another test (De Ayala, 2009). According to CTT, test parameters are dependent on examinees' characteristics. CTT difficulty index, which denotes by *p*, is obtained by calculating the proportion of test takers who answer the item correctly. The discrimination index is computed by attaining the total point-biserial correlation of test items that are answered correctly (DeMars, 2010). Among many theories developed in terms of psychometric properties, CTT had been the most dominant one over one-hundred years. Although CTT has been used for a long time, it has received many criticisms.

The CTT opponents claim that it is based on weak theoretical assumptions which make it an unfalsifiable model. Hambleton and Swaminathan (1985) discussed five common shortcomings of CTT: (1) test statistics indices depend heavily on the nature of the examinee sample, (2) the ability comparison between examinees is based on their final score on the test, ignoring the item level analysis, (3) the reliability index is obtained by administering two parallel forms, which is hard to achieve, (4) CTT does not provide any information about examinees behavior in terms of each test item, and (5) CTT assumes the homogeneity of variance in estimating the error of measurement. Therefore, another psychometric theory was introduced as an alternative for CTT during the second half of the twentieth century, which is known as item response theory (IRT).

Item response theory (IRT). IRT overcomes all theoretical and practical shortcomings of the CTT which makes it preferable by many researchers. IRT is a psychometric paradigm for analyzing test items and explaining the relationship between the examinee’s response and the underlying ability being tested (Hambleton, Swaminathan, & Rogers, 1991). Moreover, IRT is based on the probability of answering the item correctly as a function of an underlying latent trait under two major assumptions: unidimensionality and local independency (De Ayala, 2009). The unidimensionality assumption presumes that there is only one dominant factor which explains students' performance in the test. On the other hand, local independency assumption means that students' responses to the test items are independent of each other given the ability of students (DeMars & Jurich, 2015).

IRT is applied based on estimating the test item parameters and the examinee parameters; consequently, it can be applied in many different models. The most used IRT models are the dichotomous models of the one-parameter logistic model (1-plm), two-parameter logistic model (2-plm), and three-parameter logistic model (3-plm) (De Ayala, 2009). The simplest model is the 1-plm, because it estimates only the difficulty parameter (i.e., *b*-parameter) assuming that the discrimination (i.e., *a*-parameter) and the guessing (i.e., *c*-parameter) parameters are invariant across test items. It is mathematically expressed by the following formula:

$$P_{ij}\,(\theta)=\frac{e^{(\theta-b_i)}}{1+e^{(\theta-b_i)}}, \qquad (2)$$

where $P_{ij}\,(\theta)$ is the probability of answering item *i* correctly by an examinee *j* with ability level at θ, *e* is a transcendental number which equals 2.718, and b_i is the difficulty parameter of item *i.* The *b*-parameter is the item difficulty parameter. It also provides information about the difficulty level of a given item. The more difficult the item, the higher ability is required by the examinee to answer the item correctly (Hambleton et al., 1991). In theory, the value of *b*-parameter ranges from -∞ to +∞; however, the practical value usually ranges between -2 to +2. According to the 2-plm, the *a*-parameter is allowed to vary. It is expressed in the following equation

$$P_{ij}\ (\theta) = \frac{e^{Da_i(\theta - b_i)}}{1+e^{Da_i(\theta - b_i)}}, \qquad (3)$$

where a_i is the discrimination parameter for item i, and D is the scaling factor with a value of 1.7. The a-parameter refers to the steepness of the slope of the item characteristic curve (ICC). It allows differentiation between examinees from low ability group and other examinees in the high ability group (Camilli & Shepard, 1994). A large value of the a-parameter indicates that the item strongly discriminates between examinees on the ability level. The 3-plm can be applied to data set by adding the c-parameter to the equation (3), given

$$\mathrm{P}_{ij}\,(\theta)= c_i + (1 - c_i)\,\frac{\mathrm{e}^{Da_i(\theta\text{-}\mathrm{b_i})}}{1+\mathrm{e}^{Da_i(\theta\text{-}\mathrm{b_i})}}\ , \qquad (4)$$

where c_i is the guessing or pseudo-chance parameter for item i. This parameter estimates the probability of answering the item correctly by chance.

CTT and IRT shape most of the statistical and theoretical procedures that are applied to examine the psychometric properties for any measure; however, IRT remains the most usable for its outstanding attributes to address modern problems (Finch, French, & Immekus, 2014). Form both CTT and IRT perspectives, many statistical models and applications have been developed. One of the implications of these psychometric theories is the differential item functioning analysis.

Differential item functioning (DIF). DIF is another psychometric tool that is used to detect test items that function differently between two equal ability groups drawn from the same population (Furr, 2011). DIF means that two groups (e.g. gender) who have equal underlying ability may have a different probability of answering an item correctly due to other nuisance variables (Gall et al., 2007). Two forms of DIF have been distinguished in literature: uniform DIF, and nonuniform DIF. The uniform DIF refers to the consistency of DIF in item performance on the ability level, which means that the slopes of two subgroups do not intersect across the ability level. Whereas, nonuniform DIF refers to differences in performance on the ability level, which means that the slopes of two different groups intersect at some point at the ability level (Swaminathan & Rogers, 1990).

Historical development of DIF. The history of studying item bias can be traced back to the beginning of the twentieth century when Binet and Simon noticed that some items in an intelligence measurement were culturally biased against a group of students who belonged to a low socioeconomic class and they conducted their research in this area since the beginning of the twentieth century (Binet & Simon, 1973). Since Binet and Simon raised the concern about potential bias items in some high-stakes tests, various sophisticated statistical methods have been developed to detect item bias (Camilli & Shepard, 1994). For decades, the term item bias or item impact were used interchangeably to refer to any item in educational or psychological measurements that function differently between two groups from the same population (De Ayala, 2009). In 1960s, the terminology changed to DIF instead of item bias or item impact (Holland & Thayer, 1988).

Zumbo (2007) discussed three generations of DIF. He described the first generation as the starting point of investigating item bias by comparing the scores of two groups (focal group and reference group). During this generation, item bias or item impact was the formal term used in the literature. The second generation began when the term DIF replaced item bias and was widely accepted as a term referred to item bias. As Zumbo (2007) stated, the second generation was characterized by focusing on developing more sophisticated statistical methods for detecting DIF and distinguishing between item bias and item impact. Ackerman (1992) conducted a study to differentiate between item bias and item impact through the methodology used to detect each type. Item bias is detected by matching the two groups of interest on the latent ability to control the differences in performance on the ability variable, whereas item impact is considered the group differences on the measured ability (Osterlind & Everson, 2009). The third generation as described by Zumbo (2008) is the future of DIF. The reasons beyond the occurrence of DIF still needs more investigating which leads to the third generation of DIF (Zumbo, 1999). Historically, the importance of investigating test items that function differently among two different groups arises from the importance of the inference and interpretation of test results (Holland & Thayer, 1988). Therefore, as Zumbo (2008) emphasized, during the second generation of DIF the focus was on developing the methodology of investigating item bias. In this study, potential DIF items will be discussed and examined in terms of the DIF second generation.

Methods of detecting DIF. There is a variety of statistical methods developed to detect test items that function differently between two groups of examinees with equal ability level. These methods are classified as CTT-based methods and IRT-based methods. Atalay Kabasakal, Arsan, Gök, and Kelecioglu (2014) claimed that the CTT-based methods such as Mantel-Haenszel (MH), Logistic Regression (LR), and the SIBTEST were based mainly on comparing distributions and calculating the subgroups invariances. On the other hand, IRT-based methods such as Lord’s chi-square test, Raju’s area measure, and likelihood ratio are based on parameter and/or model comparison. However, the IRT-based methods are more powerful due to its strong assumptions that underline data analysis (San Martin, 2016).

Statement of the Problem

Considering the demand on a valid educational test, this study was designed to examine high-stakes test outcomes for any potential DIF items. The researchers investigated the gender-related DIF, gender differences within the ELL and non-ELL groups in an English language art measure based on CCSS, and then compare each gender group in the non-ELL group with the corresponding ELL gender group. Many research studies based on DIF analysis showed that DIF items, which are identified as bias against a subgroup in a particular population, do not mean that they impact all members of that group at the same level (Grover & Ercikan, 2017).

Previous studies have examined DIF through a range of different lenses; however, there was no study used data from CCSS-based tests. This study aimed at expanding the empirical DIF literature to consider gender difference within and cross two groups from different linguistic backgrounds (ELL vs. non-ELL) to develop a better understanding of how gender may affect students’ performance in high-stakes testing and to examine the consistency of DIF results between and within these groups. It also aimed at comparing the consistency of the DIF results obtained from the CTT-MH procedure and IRT-likelihood ratio.

Research Questions

1. Is there any gender-related DIF on the 7th grade CCSS ELA items (male vs. female)?

2. Is there any ELL-related DIF on the 7th grade CCSS ELA items (non-ELL vs. ELL)?
3. Is there any gender-related DIF within ELL and non-ELL the 7th grade CCSS ELA items (male ELL vs. female ELL, and male non-ELL vs. female non-ELL)?
4. Is there any gender-related DIF across ELL and non-ELL groups (male ELL vs. male non-ELL, and female ELL vs. female non-ELL)?
5. Is there any difference between the DIF results of MH procedure and IRT-likelhood raion method?

Significance of the Study

The shift towards evaluating students on high-stakes tests and comparing them on a national norm brings challenges for designing fair tests. In theory, high-stakes testing is run through strict statistical procedures to ensure the statistical characteristics of these tests such as validity and reliability. However, many studies showed differences in test performance between two groups of equivalent ability from the same sample based on advanced psychometric DIF analysis (Aryadoust, 2012; Grover & Ercikan, 2017; Hamilton, 1999; Innabi & Dodeen, 2006; Kan & Bulut, 2014; Young & Sudweeks, 2005).

Therefore, investigating DIF in a standardized test may provide more detailed information about students’ different performance that could help improve learning instructions at schools and provide the educators, policymakers, and practitioners with accurate guidance for test result interpretations. Also, data used in this study were from 7th grade population, which is classified as adolescent age. In addition, literacy skills and instructions begin to develop significantly during early adolescence; therefore, studying students from this age group may be especially enlightening. Also, most DIF studies that focused on this age group used math and/or science test items, because adolescent students start to develop more complicated cognitive skills, which makes more interesting for the researchers to investigate the differences between adolescent subgroups (Ong, Williams, & Lamprianou, 2015; Wang & Lane, 1996; Young & Sudweeks, 2005). On other hand, there are few studies examined DIF language items among adolescent subgroups (Koo, Becker, & Kim, 2014). Consequently, studying gender in relation to their linguistic background may account for the variations in their responses to a particular test item.

CHAPTER II: LITERATURE REVIEW

This chapter presents the rationale for examining gender-related DIF and the ELL status on high-stakes tests using two DIF detection methods. The following review of literature represents the literature pertinent to this research study. Specifically, this chapter consists of three sections. It starts with the importance of studying the high-stakes educational tests' outcomes which were used as the data of this study. This is followed by a general discussion of DIF detection methods with a detailed discussion of the two methods used in this study to analyze the data. The last section highlights the gender and linguistic background as a function of differential test performance, which were considered as the studied variables.

High-stakes Educational Tests

High-stakes tests refer to a measure that is designed to evaluate and compare students' achievement against specific standards and learning goals that are prerequisite for students' success (Jones et al., 2003). The high-stakes educational testing is a major and effective component in evaluating and understanding the learning and teaching processes. It is used as a tool of distinguishing between different ability levels, predicting students' future academic achievements, and comparing between students' performances, schools, districts and states (Reynolds & Livingston, 2012). During the past century, the high-stakes educational and psychological tests were introduced to gather information about the school system and to find ways to fix the issues related to education in general (Marzano, 2018). A substantial research literature has studied the results of using high-stakes testing on improving educational outcomes. Fuchs, Fuchs, Hamlett, and Stecker (1991) demonstrated that systematic monitoring of students' achievement using curriculum-based measurement (CBM) can help improve their performance and adjust the classroom instructions accordingly. High-stakes tests are also used in literature to examine the eligibility for special education referral (Phillips, 1993; Thurlow, Ysseldyke, & Silverstein, 1995). Most of the educational reform efforts in the American public school system were instigated by information driven from high-stakes testing that measure students' academic growth (Braun & Matthias, 2017). Therefore, what is taught in classrooms and what is assessed should be aligned very well in order to make sure that the learning objectives have been achieved. The initiative of No Child Left Behind Act (NCLB, 2002) highlighted a need for adopting nationwide standards that require an evaluation mechanism to accommodate these standards. Therefore, designing a test that measures these standards or reflects them becomes an urgent need for the new educational reform that has started by the beginning of the current century.

The high-stakes testing proponents call for adopting these tests for the following reasons: (1) high-stakes tests hold teachers and students accountable and keep them motivated, (2) they are a proper diagnostic tool to measure the curricula, (3) they allow to compare scores among teachers, students, and schools through a comprehensible comparison that is easy for parents to understand, and (4) high-stake tests' outcomes provide evidence for students professional development which can help improve the teaching instructions (Jones et al., 2003). However, Amrein and Berliner (2002) reviewed scientific research studies to examine the validity of the above statements. They found that high-stakes testing domain does not provide valid evidence for authentic learning; consequently, scores obtained from these tests are incorrectly interpreted in most of the cases.

There is enormous literature in the field of educational assessment that focuses on how to interpret the outcomes (Braun & Matthias, 2017; Bulut, Quo, & Gierl, 2017; Hickey & Zuiker, 2005; Immekus & McGee, 2016). However, these studies provide more evidence supporting the claim of having troubles in making relatively coherent inferences from the high-stakes testing that becomes an essential part of the educational system. The interpretations and inferences of test outcomes are a very paramount and crucial step in terms of making educational decisions (Walsh & Betz, 1985). Therefore, the consequential implications of the test outcomes reflect the importance of reaching the academic goals and applying the standards. The inferences can be strengthened within a psychometric framework by ensuring the validity of the test and the reliability of its scores (Linn, Baker, & Dunbar, 1991). Zumbo (1999) referred to validity and reliability as traditional methods to evaluate test results. He recommended using more sophisticated psychometric procedures to improve research-based analysis and decision making. However, ensuring the test validity and reliability remains an issue among many teachers and practitioners. Elliott, McKevitt, and Kettler (2002) reviewed the results of four research studies conducted to examine test accommodations for disabled students and validity issues. They remarked that educators who are responsible for decision-making are less likely aware of the validity of test accommodations. Therefore, applying psychometric procedures are crucial to making educational decisions regarding minority groups in schools.

Fixing the assessment issues has always been a concern for educators, policymakers, teachers, and practitioners. Linn (2000) reviewed the five waves of educational reform based on the assessment outcomes and how educational assessment contributed to the field of education during the last half of the past century. He discussed the last wave of educational reform as a stage of adopting common standards and high-stakes accountability system that increase the demand on developing valid assessments. Adopting a standardized assessment can have a negative impact if other factors such as school effect, ethnic groups, and/or gender have been ignored (Linn, 2000). Therefore, spending too much time teaching only the test content will have a consequential negative impact on the validity of the test as well as on the interpretation of its outcomes (Linn et al., 1991).

The efficiency and utility of high-stakes testing. Several decisions can be made according to the score reports provided by districts. Low scores can lead to school closure resulting in many teachers lose their jobs, on the other hand, high scores are considered as an indicator to a good education (Brigance & Hargis, 1993). Also, these decisions are not only affecting students' academic life but also are somehow related to their personal life. Families, for example, choose where to live based on school rankings to provide their kids with better education opportunities (Amrein & Berliner, 2002). Scoring system and comparison on high-stakes tests may motivate teachers and students to improve their scores on high-stakes tests to be able to compete nationally and internationally. Consequently, high-stakes tests become a popular source of accountability, which also leads to a high-stakes accountability system (Wixson & Carlisle, 2005).

Porter, McMaken, Hwang, and Yang (2011) conducted a study to compare the CCSS-based curriculum content with the state standards-based curriculum content in mathematics and English language arts and reading (ELAR) to measure the degree of agreement between both contents. They found that the correlation between the two contents is low to moderate at all grade levels. Moreover, Taylor (2004) documented how these tests negatively affect the minority groups in schools. However, she focused only on the low socioeconomic status students and at-risk students. Other minority group students such as ELLs were not discussed in her book.

Test fairness and psychometrics. The rapid improvement in assessment practice and policies requires more analysis to understand its outcomes (Linn et al., 1991). Therefore, the test designers and testing companies are carefully examining the psychometric properties of test items to ensure test fairness (Brigance & Hargis, 1993). The diversity in the classrooms nationwide put them in a big challenge to provide a test that has no bias. Test bias has been recognized in the literature as a problem that results in unfair evaluation to the examinees' subgroups (Shealy & Stout, 1993). Test bias also can occur when another nuisance factor is tested besides the main tested factor (van de Vijver, Fons, & Poortinga, 2005). Late in the last century, test bias had been replaced by a more proficient term which has been known as DIF (Sireci, Patsula, & Hambleton, 2005). Therefore, DIF is a relatively new term used in literature.

DIF as a Tool of Examining Test Fairness

Test developers apply several statistic procedures to ensure the quality of the test items in terms of test fairness for all examinees. One of the commonly used statistical procedure is DIF which is used to identify potentially biased items across different groups of examinees (Camilli & Penfield, 1997). DIF is referred to as a statistical technique that is applied to detect test items that function differently between two or more groups of test takers due to the variables not directly related to the cognitive components of examinees (Dorans & Holland, 1993). A relatively enormous psychometric studies were conducted in the domain of high-stakes tests; however, there was no study that used data from high-stakes standardized tests that were designed to measure the standards of the CCSS. Furthermore, despite the importance of high-stakes tests, there has been little research devoted to this topic in DIF literature. Camilli (2006) claimed that, in terms of test fairness, high-stakes tests are not well examined.

DIF detection methods. During the last century, a variety of methods has been developed to detect DIF for both dichotomously and polytomously scored items. However, the prominent methods in the literature are based on IRT (Ahmadi & Thompson, 2012; Grover & Ercikan, 2017), the MH approach (Allalouf & Abramzon, 2008; Andrich & Hagquist, 2012; Atalay Kabasakal et al., 2014; Beaver, French, Finch, & Ullrich-French, 2014; Camilli & Penfield, 1997; Innabi & Dodeen, 2006), Rasch model (Alavi & Bordbar, 2017; Andrich & Hagquist, 2012; Aryadoust, 2012), logistic regression (Abedlaziz, et. al., 2011; Arikan, van de Vijver, Fons & Yagmur, 2018; Bastug, 2016; Camilli & Congdon, 1999; Fidalgo, Tenenbaum, & Aznar, 2018), and likelihood ratio test (Cohen, Kim, & Wollack, 1996; Kim & Cohen, 1995).

Fan (1998) compared CTT and IRT in terms of their person parameter estimates, item parameter estimates, and the degree of invariance between item statistics obtained from both methods through a correlation analysis. He used data from a high-stakes test from a statewide assessment program administered to evaluate students reading, math, and writing abilities. The results showed that the correlation between CTT and IRT parameter estimates were from high to moderate, which indicated that CTT is as strong as IRT in estimating person and item parameters. However, Fan did not expand his research to detect DIF using both methods. He only studied the how each theory can provide accurate estimation to the test and examinees parameters. The DIF detection methods can be classified as CTT-based methods and IRT-based methods. In the following sections, the advantages and disadvantages of each method will be highlighted.

CTT-based methods (non-IRT). CTT is one of the commonly applied psychometric methods. There are many approaches based on the CTT that are used in the literature to examine DIF. The major aspect of all CTT-based methods is that they are based on distribution comparisons between the focal and the reference groups. The next sections discuss the most common CTT- based methods.

Logistic Regression. Logistic regression is one of the widely used non-IRT approaches for detecting DIF items across groups. Logistic regression was proposed by Swaminathan and Rogers (1990). It refers to how to make predictions from one or more quantitative and/or qualitative variables about a binary variable (de Ayala, 2009). In the DIF context, logistic regression can be used to predict the behavior of two groups of examinees as a unit of analysis from their group membership, their ability level, and the interaction between these two variables (Camilli & Shepard, 1994).

There are many studies that have applied logistic regression to detect DIF (Cheema, 2017; Doğan, Hambleton, Yurtcu, & Yavuz, 2018; Kim & Oshima, 2013; Kim, 2001; Kurt, Karakaya, Safaz, & Ates, 2015; Lambert et al., 2017; Miller & Spray, 1993; Ong et al., 2015; Wang & Lane, 1996; Wedman, 2017; Woods & Harpole, 2015). However, most of these studies used simulated data to compare logistic regression with other DIF detection methods. One of the major advantages of applying logistic regression as a DIF detection method is that it has a superior power to detect uniform and nonuniform DIF types (Camilli & Shepard, 1994). On the other hand, logistic regression has a high rate of occurring the Type I error due to its sensitivity to the small differences between the studied groups (Jodoin & Gierl, 2001). Also, there is no effect size measure associated with logistic regression analysis (Cohen & Bolt, 2005).

Standardization. Standardization method of detecting DIF is similar to MH in that both methods are based on testing the null hypothesis of having an equal expectation for two subgroups of examinees on the same ability level (Dorans & Holland, 1993). However, unlike MH, standardization method computes the proportion correct statistics for all available data in the reference and the focal groups after matching them on their total score (Dorans & Kulick, 1983). Standardization method is rarely used in DIF literature.

Simultaneous Item Bias Test (SIBTEST). Shealy and Stout (1993) established another procedure to assess test items that function differently among subgroups of examinees. They claimed that it is a proper statistical procedure to detect test bias, DIF, differential bundle functioning (DBF), and/or differential test functioning (DTF) simultaneously. DBF is another form of detecting test bias by grouping the test items into bundles or more precisely classifying them into categories (Camilli, 2006). They also distinguished between test bias and DIF in terms of test construct validity. Simultaneous Item Bias Test (SIBTEST) is a nonparametric model using multidimensional item responses data to detect DIF in dichotomous data after matching the reference and focal groups on their ability variable (Atalay Kabasakal et al., 2014). SIBTEST is designed to accommodate multidimensional data.

Transformed item difficulty (TID). Transformed item difficulty (TID) or delta-plot is one of the first methods used to detect item bias. As a method based on CTT, the TID method calculates the item difficulty index by computing the proportion of the examinees who answer the item correctly. TID was first introduced by Angoff (1972) when he studied the cultural difference between the two groups. TID method is based on calculating item difficulty index (p-value) for both studied groups on each test item, and then converting the p-values to a normal deviate z corresponding to the $(1 – p)$ the percentile with a mean of 13 and standard deviation of 4.

Abedlaziz et al. (2011) utilized the TID methods to identify DIF items on mathematical ability test across 10-grade males (n = 380) and females (n = 420) groups. They compared the degree of agreement between the TID and logistic regression on identifying DIF. They found that seventeen items out of forty exhibited DIF. However, their findings revealed that the TID and logistic regression showed low agreement in the DIF items.

Mantel-Haenszel (MH). The MH procedure is a χ^2 test of the DIF null hypothesis which states that the odds of answering a given item correctly in both groups are equal using χ^2 distribution with one degree of freedom (de Ayala, 2009). The MH procedure was first suggested by Scheuneman (1979); however, Scheuneman's statistical procedure received criticism in terms of its methodology (Camilli & Shepard, 1994). Later, Holland and Thayer (1988) developed a statistical approach based on the Mantel and Haenszel (1959) procedure to detect DIF. They claimed that the MH procedure for testing the DIF null hypothesis is the most powerful test among other DIF detection methods that are based on CTT. Holland and Thayer (1988) provided two statistical procedures to detect the common odds-ratio between the focal and the reference groups which denotes by α_{MH} and the magnitude of the DIF item which denotes by Δ_{MH}. The Δ_{MH} is used to calculate the differences between the focal and reference groups in the difficulty parameter, and it is referred to as MH odds ratio (Camilli & Shepard, 1994). The MH χ^2 procedure is based on the calculation of a binary item response with the group membership which is added up into a series of 2 * 2 * *k* contingency tables for each interval on the matching variable which is the ability level, where *k* is the number of groups (de Ayala, 2009). The baseline of the MH procedure is that the focal and the reference groups are compared on the odds-ratio after they are matched on the proficiency of interest which is usually their total raw score. In other words, MH is studying the relationship between two variables (the group membership and the item response) after controlling the third variable which is their total raw score (Sireci, et al., 2005).

The advantages of applying the MH procedure as a method of identifying DIF are because it does not require expensive software to be computed, and it also does not require a large sample size to obtain valid results. However, the MH procedure has a disadvantage of not detecting nonuniform DIF. It is used only for dichotomously scored items because it requires a binary response for the data analysis. Many scientific studies examined the occurrence of DIF by applying the MH procedures (Andrich & Hagquist, 2012; Atalay Kabasakal et al., 2014; Bastug, 2016; Camilli & Penfield, 1997; Dorans, 1989; Hambleton & Rogers, 1989; Hidalgo-Montesinos & Gómez-Benito, 2003; Kim & Oshima, 2013; Woods & Harpole, 2015). Allalouf and Abramzon (2008) investigated DIF using the MH procedure in a language test for a large sample size of more than

30,000 examinees who come from different countries. For the purpose of their study, they divided the study into two parts: in part one, they examined a pilot group of examinees (2,200 Arabic speakers and 1,500 Russian speakers) in order to detect the items that exhibit DIF; in part two, they conducted a validation study using a small sample of the examinees used in part one. Allalouf and Abramzon (2008) conducted their study to put recommendations on how to reduce DIF items in second language assessment domain. The findings showed that eighteen items out of 44 (42%) of the test items were flagged as DIF; however, the pilot study reduced the number of the DIF items.

Another pertinent study was conducted by Beaver et al. (2014) to assess test items for potential gender-related DIF in a psychological scale that assessed the social and emotional development for children age 0 to 7. They used the MH procedure to detect gender-related DIF among young children and found that four items out of fifty exhibited large DIF. Their results indicated that sex differences can lead to misinterpretation to development delay using this measure. Oliveri, Lawless, Robint, and Bridgeman (2018) analyzed 320 GRE test items from the verbal and quantitative sections between U.S. citizens group (*n* = 300) and non-U.S. citizen group (*n* = 300). They found that DIF items disadvantaged the non-U.S. citizen group are testing the knowledge of English idioms. The MH procedure was used in this study for its advantages.

IRT-based methods. IRT is one of the most frequently used methods to detect DIF for a long time. A numerous number of IRT-based methods have been used in DIF literature. Parameter comparison is the cornerstone for all the IRT models (Hambleton et al., 1991). Fundamentally, the parameters of any chosen model are estimated then compared. If the parameters equally fit the data for both groups (the focal and the reference groups) then there is no DIF, if the parameters of one of the groups do not fit the data, then there is a sign for a DIF (Camilli & Shepard, 1994). Practically, there are several statistical procedures to test the model-data fit from an IRT perspective. The most used ones in the DIF literature are: Lord's $\boldsymbol{\chi^2}$ (Lord, 1980), Exact signed and unsigned area (Raju, 1988), the likelihood ratio (Thissen & Steinberg, 1988), Rasch model (Rasch, 1960), Mixture IRT (Cohen & Bolt, 2005), and graded response model (GRM) (Andrich, 1978; Masters & Wright, 1984).

Lord's χ^2. In the context of IRT, Lord (1980) developed the 2PL IRT model which was based on estimating two parameters instead of only one. The *a*-discrimination parameter was added to the 1PL IRT model. In detecting item bias, Lord suggested that item bias occurred if the ICCs of the two groups differ significantly. Using an asymptotic χ^2 test, Lord evaluated the null hypothesis which stated that *b* parameter and *a* parameter of the focal and reference groups are equal. He also proposed a test of significance to assess DIF. Lord's χ^2 is similar to the χ^2 distribution with two degrees of freedom (Thissen, Steinberg, & Wainer, 1993).

McLaughlin and Drasgow (1987) were one of the first researchers who used Lord's chi-square to identify item bias. They studied a pooled sample of 1000 and 250 and 50 test items in simulated data which is the minimum number of sample size and test items suggested by Lord (1980) to get valid parameter estimation. McLaughlin and Drasgow (1987) generated the data using 2PL and 3PL models at different level of significance. They found out that using Lord's χ^2 can increase the rate of Type I error due to the inflation of the chi-square value. They also tested the Lord's χ^2 statistic under different methods of θ estimation and found out that Lord's χ^2 can give misleading results if maximum likelihood estimation is used. Angoff and Cook (1988) also applied Lord's χ^2 to an empirical data to compare between the Scholastic Aptitude Test (SAT) and the Prueba de Aptiud Academica (PAA) Spanish-language version. The purpose of their study was to find a correspondent score to in the Spanish translated version. Other research studies applied Lord's χ^2 in purpose of comparing it with other statistical methods (Chan, Drasgow, & Sawin, 1999; Ellis & Kimmel, 1992; Kim, Cohen, & Park, 1995; Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001).

However, despite the strong mathematical equations that underline Lord's χ^2, there have been many criticisms which makes it unpreferable by other researchers in recent years. The main criticism against Lord's χ^2 test was that it was not powerful enough to detect any significant difference between ICCs of each group even if there were different parameter estimations between the studies groups (Camilli & Shepard, 1994). Another Lord's χ^2 disadvantage is that the *c*-parameter is not included in the equation of DIF detection because it is poorly estimated in the 3PL model (Sireci et al., 2005).

Exact signed and unsigned area. Raju (1988) developed a new approach based on IRT to overcome Lord's χ^2 shortcomings; this method is so-called Exact Signed and Unsigned Area. Raju studied the area between the ICCs of two groups for any significant differences. This method is known as a practical difference between the two groups in a given item. This method is applicable for dichotomous and polytomous items (Sireci et al., 2005). The null hypothesis states that the area between two ICCs is identical after estimating the parameters and put them on a common scale (Hambleton et al., 1991). Raju (1990) also proposed a significant test associated with the magnitude of the DIF item to show that if the DIF is large, then there is a significant difference between the ICCs of the two groups. Raju (1988) developed a mathematical formula to measure the signed and unsigned area between two ICCs. The signed area formula is used when the ability level of the two groups does not intersect with each other. It is based on subtracting the probability of answering an item correctly of the focal group from the probability of answering the item correctly of the reference group.

On the other hand, if the ability level of the focal group is inconsistent across the ability level, the ICCs of two groups will intersect and results in canceling each other. Therefore, the formula in the unsigned area differs slightly in terms of using the square root of the outcome to avoid canceling the value of the difference (Camilli & Shepard, 1994).

Many studies in the DIF literature have applied the area index for DIF detection (Bastug, 2016; Benson, Donnellan, & Morey, 2017; De Beer, 2004). Teresi, Kleinman, and Ocepek-Welikson (2000) compared the DIF results using Lord's χ^2 and signed/ unsigned indices to evaluate the IRT-based methods results in terms of DIF detection. They used data from a cognitive screening test to investigate group invariance in terms of their education level, ethnicity, and race (n = 924). The analysis consisted of nineteen items fourteen of which were dichotomously scored. Their results aligned with other research findings, revealing that IRT-based methods of detecting DIF give relatively similar outcomes. Raju's area index was not recommended by many researchers because of the misleading results if the majority of the examinees were at the either side of the ability continuum (Camilli & Shepard, 1994). Moreover, there is no test of significance associated with the area index (Swaminathan & Rogers, 1990).

Rasch model. Rasch model is the most powerful DIF detection methods used for high-stakes assessment data (Wyse & Mapuranga, 2009). It is basically based on comparing the item difficulty parameter in both groups under study. Mathematically, Rasch model is similar to the 1PL IRT model; however, it was developed separately by Rasch (1960) in which the test takers are grouped based on their raw scores.

Rasch model has been applied in the DIF literature for dichotomous test items (Andrich & Hagquist, 2012; Cauffman & MacIntosh, 2006). Alavi and Bordbar (2017) investigated DIF in a high-stakes assessment in Iran from a sample of 5000 examinees who took a National University Entrance Exam for Foreign Language (NUEEFL). They used Rasch model as a method to detect DIF between males (*n* = 1665) and females (*n* = 3335). After the model met the unidimensionality and local independence assumptions, the researchers checked their data for the goodness-of-fit index by estimating the infit and outfit indices using the mean-square fit values (MNSQs) and the standardized z values (ZSTDs). They investigated the interaction between the person ability and the item difficulty across males and females at 0.5 logits or larger. Alavi and Bordbar (2017) found out that 40 items out of 95 items in the test were identified as DIF items. However, the researchers recommended to break the sample down into subtests in order to understand the area of weakness among the students. DeMars and Jurich (2015) studied DIF results obtained from Rasch model after matching the examinees on their ability level. The purpose of their study was to evaluate the effect of the Rasch model on DIF detection when the ability levels of the two compared groups are remarkably different. The results revealed that the Type I error was inflated and the DIF effect size was biased when the two groups were very different. They recommended not to apply Rasch model when the two groups are very different on their ability level due to the large effect of the guessing parameter on the data. Camilli and Shepard (1994) believed that Rasch model has limited application since it provides estimation to only one parameter.

Mixture IRT. Recently, a new IRT-based method was proposed by Cohen and Bolt (2005) to detect DIF within a multidimensional framework. It is called the mixture IRT. Yalcin (2018) studied gender difference on an international mathematical assessment among 4th grade students (n = 1166). The purpose of his study was to detect DIF items in the context of multidimensional measure. He found no significant differences between males and females in the mathematical measure. The findings revealed that using mixture IRT models with multidimensional measure may help interpret the gender difference as ability differences rather than DIF. Mixture IRT model is applied only for multidimensional measures.

Likelihood ratio. Likelihood ratio is one of the most widely used IRT-based methods in the recent DIF literature. As all other DIF detection IRT-based methods, likelihood ratio is mainly based on parameter comparison. Thissen and Steinberg (1988) developed a mathematical equation to detect DIF items by testing the null hypothesis which assumes that the parameters of the focal and reference group are the same. To test this hypothesis, two IRT models are tested for the goodness-of-fit index. It is basically based on model comparison where the first model or no-DIF model keeps all the parameters in all test items constrained. On the other hand, the second model or the DIF model still keep all the parameters constrained except for the parameters of the item under studying. Then the likelihood ratio test statistics (TSW-ΔG^2) is applied to compare between these two models with a degree of freedom equal to the difference between the parameters in the two models. If the test yields a significant result, then the item exhibits DIF. The likelihood ratio can be obtained from the following formula:

$$LR = -2LL_c - (-2LL_A), \tag{5}$$

where $-2LL_c$ is the compact model or the no-DIF model, and $-2LL_A$ is the augmented model or the DIF model. Likelihood ratio test for DIF is based on two nested model comparison as De Ayala (2009) stated. The likelihood ratio test statistics is originally used to evaluate the model fit and model comparison within the IRT framework (Millsap, Gunn, Everson, & Zautra, 2014). The baseline of the likelihood ratio test for DIF is to compare between two models with different number of parameters to see if adding more parameters lead to any significant results (De Ayala, 2009). In this case, the constrained compact model will have fewer parameters than the augmented model (i.e. small model and large model). For using this method as a DIF detection method, two different subgroups from the same population are compared instead of comparing two models and then a χ^2test of significance is applied to test if the difference between the parameter estimations of both groups is significant (De Ayala, 2009).

Many studies applied likelihood ratio test statistics to compare between two groups for any DIF potential items (Albano & Rodriguez, 2013; Atalay Kabasakal et al., 2014; Cohen & Bolt, 2005; Elosua & Wells, 2013; Kim & Yoon, 2011; Stark, Chernyshenko, & Drasgow, 2006; Woods & Harpole, 2015). The likelihood ratio method is the most prevalent IRT-based method in the literature because of its power of controlling the Type I error. The null hypothesis of the likelihood ratio assumes that the test parameters in the reference and the focal groups are invariant. Kim (2001) compared the DIF results based on the likelihood ratio test to the logistic regression method. He studied polytomously scored responses of European languages students (*n* = 571) and Asian languages students (*n* = 467) who took the Speaking Proficiency English Assessment Kit (SPEAK), which showed that the two detection methods revealed similar results. Barnes and Wells (2009) studied a 4-point Likert scale survey to identify any potential DIF items between gender and ethnicity groups using the likelihood ratio as a detection method. They found that seven items out of forty-eight functioned differently between males and females, on the other hand, only one item exhibited DIF among the ethnicity groups.

Other IRT-based models are also used in literature for the polytomously scored items. The most used method is the graded response model (GRM) proposed by Samejima (1969). Many research studies applied the GRM to a polytomous data to detect DIF (Elosua & Wells, 2013; Lambert, January, Cress, Epstein, & Cullinan, 2017; Wei, Chesnut, Barnard-Brak, Stevens, & Olivárez, 2014; Yau, Wong, Lam, & McGrath, 2015).

Budgell, Raju, and Quartetti (1995) compared a French translated version of two standardized Canadian Numerical Test (15-item) and Reasoning Test (18-item) with its English version to detect DIF items in the translated version. They applied three IRT-based methods (signed area, unsigned area, and Lord χ^2) and one CTT-based method (MH) to detect DIF in a translated assessment instrument. The other purpose of their study was to evaluate using MH as a DIF detection method compared with the three IRT methods. Their finding revealed that there was remarkable consistency in identifying DIF using both IRT-based methods and MH.

DIF types. DIF items could also be classified as being uniform or nonuniform. Both IRT-base methods and CTT-based methods were used in DIF literature to detect both types of DIF.

Uniform. Uniform DIF occurs when the magnitude and direction of the DIF item favor only one group (usually the reference group) along the theta continuum (Swaninathan & Rogers, 1990). Investigating uniform DIF leads to the ICCs of both groups having equal *a*-parameter (Maller, French, & Zumbo, 2011).

Nonuniform. Nonuniform DIF occurs when the DIF item favors one group at some points at the theta continuum; however, the ICCs of both groups from IRT perceptive cross at a certain point at the ability level (Maller, French, & Zumbo, 2011). Nonuniform DIF can exhibit DIF with an interaction between the group membership and the ability level which indicates that the differences in the likelihood of the correct response on a given item depend on the ability level regardless the group membership (Camilli & Shepard, 1994).

Cheema (2017) investigated the occurrence of uniform and nonuniform DIF in a high-stakes assessment between two gender groups from high economic countries (OECD countries) and low economic countries (non-OECD countries). He applied binary logistic regression to examine test items from a pool of 50 dichotomously scored items. He found that nonuniform DIF was larger than uniform DIF in the OECD countries rather than the non-OECD countries. Cheema (2017) argued that the economic, cultural, social, and political differences between countries contribute to the differences between two gender groups.

Swaminathan and Rogers (1990) also distinguished between uniform and nonuniform DIF using logistic regression in simulated data. They compared between logistic regression and MH methods in their sensitivity of detecting the uniform and nonuniform DIF. Their findings indicated that logistic regression was more accurate in detecting both types of DIF. Another study was conducted by Wiesner, Windle, Kanouse, Elliott, and Schuster (2015) to examine only uniform DIF across gender and ethnic groups using exploratory factor analysis. They analyzed data from a predictive scale for disorder.

Gender and Linguistic Background as a Function of Differential Test Performance

Gender differences in standardized assessment. Gender is the most studied variable in DIF literature. Among most of the studies conducted DIF analysis, gender was either the only (Fidalgo et al., 2018; Wetzel & Hell, 2013; Yalcin, 2018), or one of the variables studied as a comparing variable (Hope et al., 2018; Kunnan & Weinstein-Shr, 1990; Wiesner et al., 2015). Many research studies have been devoted to investigating gender differences in educational assessments. Most of these studies found that males tend to have advantages over females in the probability of answering test items that evaluate the critical thinking and scientific topics (Carlton & Harris, 1992; Kunnan & Weinstein-Shr, 1990). On the other hand, test items that are designed to measure social studies and verbal learning are more likely to favor females over males (Carlton & Harris, 1992; Kurt et al., 2014). Yet, selecting a particular gender as a reference group was not well determined in the literature, therefore; in this study the male group was set as a reference group for all the gender comparison sets.

For a long time, research studies have reported gender differences in standardized educational assessments (Kunnan & Weinstein-Shr, 1990; Murray, Booth, & McKenzie, 2015). However, the findings of these studies are not consistent. Some of them found females outperformed males (Aryadoust, 2012; Kan & Bulut, 2014), and others found that males outperformed females (Grover & Erickan, 2017; Hamilton, 1999). Moreover, many other DIF studies showed that males have advantages over females on some items and have disadvantages on others (Innabi & Dodeen, 2006). Le (2006) studied gender differences, test language, and country of test in an international test for any DIF items. For gender variable, his focus was on item focus, context, competency, background knowledge, item format and score point. His findings revealed that males outperformed females when the focus of the item was global, when the competency and background knowledge were required. Males also did better than females when the item format was multiple choice. However, items with different score point favored females when the score point was 2. Gender has also been used in most of the methodological and validation studies (Breidenbach & French, 2010; Cohen & Bolt, 2005; Zwick & Ercikan, 1989).

In addition to studies that provided evidence for gender differences, Reardon et al. (2018) studied the relationship between gender achievement gap and test items formats. They analyzed data from a standardized math test and a standardized ELA test. Their sample was drawn from 4th grade students (*n* = 794) and 8th grade students (*n* = 665) who took the ELA test in 47 states. Math data also contained 4th graders (*n* = 777) and 8th graders (*n* = 696). The test items were classified into two categories: constructed-response items and multiple-choice items. They found that the gender gap was larger in the ELA test and favored the male group. Moreover, the multiple-choice test format better discriminated between both genders than the constructed-response item format. They also argued that almost 25% of the gender gap phenomenon is due to test item formats used in the test.

A longitudinal qualitative case study was conducted by Kanno and Kangas (2014) to evaluate the access to advanced college preparation courses for the ELL students in high-grade levels. They reviewed the educational practices at a suburban public high school. They found that ELL students lag behind dramatically during the high-grade levels comparing with their non-ELL peers, which impacted their academic achievement on standardized tests. Therefore, studying this particular population may contribute to understanding the achievement gap of the ELL students at the college level. Although gender differences have been studied in many researches, no DIF research has been conducted to examine gender difference within and cross two subgroups in the same population.

Linguistic background. Many scientific studies which have been conducted to investigate DIF with linguistically and/or culturally diverse groups claimed that first language interference could be a possible reason for exhibiting DIF in some items of the test (Akour, Sabah, & Hammouri, 2015; Aryadoust, 2012; Barnes & Wells, 2009; Ismail & Koch, 2012; Wei et al., 2014). These studies used different statistical procedures to examine the DIF items in a given test. Using a confirmatory approach, Abbott (2007) studied an English version reading subtest of the Canadian Language Benchmarks Assessment (CLBA). The advantage of this approach is that it can be used to study individual test items and groups of items (Roussos & Stout, 1996). He studied binary reading strategy used in second language scheme. This strategy is known as bottom-up and top-down strategy. Abbott (2007) compared two groups of non-English-speaker examinees, who were Arabic native speakers (*n* = 250) and Mandarin native speakers (*n* = 250). He analyzed each test item using SIBTEST based on the bottom-up and top-down reading strategy. Roughly 53% of the test items exhibited moderate to large DIF favored the Arabic speakers on some items and the Mandarin speakers on other items. His main findings were that there was a systematic difference between the two studied groups, which indicated that first language and culture affected students' performance on the second language test.

Elder (1996) applied the MH odd ratio procedure to investigate the first language impact on reading and listening standardized English test and whether there were any DIF items in the test. His sample consisted of Chinese (*n* = 1,176), Italian (*n* = 4,463), and Modern Greek (*n* = 1,224) speakers. Each group in his sample was divided into two groups; background speaker (BS) who speak their native language at home, and non-background speakers (NBS) who less likely speak their native language at home. Elder (1996) then compared BS versus NBS for each language group at three levels of proficiency. He found out that students with first language exposure performed better than those who speak English at home.

Koo et al. (2014) examined DIF patterns between ELL and non-ELL students in third and eighth grades. They analyzed high-stakes reading assessment data from Florida Comprehension Achievement Test (FCAT) in third (*n* = 173,737) and eighth grades (*n* = 160,391). The instrument measures four major reading comprehension areas; phrase-in-context, main idea, cause and effect, and evaluation. They found that third grade ELL students outperformed the non-ELL group in phrase-in-context test items. On the other hand, non-ELL students outperformed the ELL students in evaluation skills in eighth grade. Many other research studies investigated DIF in psychological assessments (Lambert et al., 2018), math and science assessments (Wei et al., 2014) to detect DIF among ELL samples. However, the results were not consistent. DIF detection studies used data from non-educational measurements not always revealed measurement invariance between compared groups.

The current study is focusing on studying DIF using data from educational high-stakes testing from 7th grade. Chae, Kim, and Han (2012) studied test items from a high-stakes assessment to evaluate DIF across five accommodated test forms (no accommodation, behavior accommodation, English audio accommodation, foreign language accommodation, and scribe accommodation). Their sample consisted of students from 3rd grade (*n* = 129,321), 5th grade (*n* = 127,840), and 7th grade (*n* = 133,444). Their results revealed that DIF was exhibited across groups and grades; however, the lower-grade level had more DIF than the higher-grade level. Therefore, examining DIF in high-level grades may lead to some insight into the response behavior among these students in terms of other variables.

Conclusion

Recently examining the high-stakes tests' outcomes using DIF has become a prominent topic. However, most of the DIF studies compared between two subgroups of examinees. This study contributed to the field of DIF by extended the analysis to break down the gender and linguistic background groups into smaller groups. This purpose was achieved by examining gender within each linguistic background groups and cross both linguistic background groups. This chapter provided a review of the DIF literature in the light of high-stakes educational tests. In addition, findings from other related studies that support the research questions of this study were discussed and reviewed.

CHAPTER III: METHOD

The primary purpose of this study was to examine any potential DIF items in a high-stakes educational test and whether the gender and ELL status impact the result of DIF. To achieve this goal, two DIF detection methods were applied on six set of comparisons. Therefore, this chapter presented the methodology that was employed to answer the research questions. The chapter was organized into three sections: selection of the participants of the study, the material and how data were obtained, and the procedure of data analysis.

Participants

A commercial testing company provided the original data. The participants of the present study were selected from the original sample of 35,800 students who took a version of the standardized English Language and Art (ELA) test at the beginning of the 2014-2015 academic year from 11 states. However, due to missing data on various demographics and test scores from the original data, only 12,658 students were included for further analysis. The sample of the study included 8,466 Caucasian (66.9 %), 1,802 African American (14.2 %), 335 Asian (2.6 %), 580 American Indian (4.6 %), 147 Hispanic (1.2 %) and 529 multi-ethnic (4.2 %) groups. All students' data in the analysis were only from 7th grade students.

For the purpose of the study, the participant was first divided into two groups: male reference group (*n* = 5705) and female focal group (*n* = 5603), and non-ELL reference group (*n* = 11,310) and ELL focal group (*n* = 1,348). Then six sets of comparisons were implemented with equal number of examinees in each group, (1) male (*n* = 316) and female (*n* = 316), (2) male non-ELL (*n* = 340) and male ELL(*n* = 340), (3) female non-ELL (*n* = 502) and female ELL (*n* = 502), (4) non-ELL (*n* = 280) and ELL (*n* = 280), (5) male ELL (*n* = 402) and female ELL (*n* = 402), and (6) male non-ELL (*n* = 831) and female non-ELL (*n* = 831). The sample size for each set of analysis was selected according to the Educational Testing Service (ETS) guideline. The minimum sample size as recommended by the ETS is at least 200 for the focal group and at least 500 for the reference group (Zwick, 2012). Therefore, the number of students was selected randomly from the pool of 12,658 for each comparison set, and they were matched on their total raw scores.

Materials

Item-level data were utilized from an archival benchmark standardized English Language Art assessment (form A) previously administered to measure students' language proficiency. The measure was designed to align with CCSS. The measure consisted of thirty-four dichotomously scored items. All test items were included for analysis in this study.

According to the CCSS guidelines, this standardized test is designed to measure students' proficiency in reading (informational text and literature) and English language arts (Language and writing) skills. Because of copyright limitations, a detailed description of the test items and examples could not be released. Additionally, test items were designed to match the seventh-grade level of difficulty. Table 3.1 illustrated the skills that were measured in the CCSS test and the number of items assigned for each area. They fell into two categories: English language arts and reading. Each category is designed to measure two areas. English language arts category has two subject areas which are language and writing. Reading category has also two subject areas which are informational text and literature. Eighteen items measured the reading skill: context clues, explicit details, text structure, main idea, compare/ contrast historical fiction to reality, theme/ summarize, point of view, story elements, use multiple sources, and meaning of words. The remaining sixteen items measured English language art: word meaning, convey information clearly, capitalization/ punctuation/ spelling, plan/ revise/ edit, organize text, write narratives, arguments, conduct research, figurative language/ nuance, specialized vocabulary, and gather information.

Table 3.1

Item categories based on CCSS

	area	standard	# of items
ELA	Language	Word meaning	3
		Capitalization/ punctuation/ spelling	1
		figurative language/nuance	1
		specialized vocabulary	1
	Writing	Convey information clearly	1
		plan/ revise/ edit	4
		Organize text	1
		write narratives	1
		Arguments	1
		Conduct research	1
		gather information	1
Reading	Informational text	Context clue	3
		Main Idea	2
		Meaning of words	1
		Explicit details	2
		use multiple	1
		sources Text	1
	Literature	theme/ summarize	2
		story elements	1
		point of view	2
		context clues	1
		text structure	1
		compare/ contrast historical fiction to reality	1
Total			34

The standardized test was administered at the students' schools, then scored by the testing company. The students' identities and any other personal information were not included at any stage of the study. Neither the testing company nor participants were compensated for their participation in this study.

Procedures

Descriptive statistics were obtained for the whole sample and for each comparison set in the analysis using SPSS software. The item responses included in each set were chosen from a pool of examinees (*n* = 12,658). The data of this study were subject to CTT and IRT analysis using Xcalibre software 4.1 version. DIF analysis was conducted using Xcalibre software 4.1 version. The reliability of the measure was also calculated using Cronbach's coefficient α. For the purpose of this study, reference groups and focal groups for each comparison set were matched on their ability level. Then the assumption of dimensionality was checked. After that the CTT and IRT analysis were implemented. Finally, DIF analysis was conducted for each comparison set.

Matching criterion. In distinguishing between DIF and the real group differences, matching criterion was proposed by Holland and Thayer (1988). The implication of using the matching criterion is to assure that both groups have equal ability distributions and that the difference in item responses should be resulted from bias items or other factors rather than their ability levels. Methodologically, matching criterion works as the internal analysis for any potential DIF items (Camilli & Shepard, 1994). For the purpose of this study, the matching variable was the total raw score of the ELA test. Equal number of examinees were selected randomly for the reference group and the focal group from each total raw score. This process was repeated for each comparison set. The ability level which was estimated from the total raw scores was used for the parametric IRT analysis. The Xcalibre software 4.1.8 was used to estimate the ability level for the IRT analysis.

Dimensionality. The unidimensionality assumption is prerequisite for IRT DIF analysis. There are many methods used in literature to test the unidimensionality of the given scale. However, in this study, unidimensionality was assessed using exploratory factor analysis (EFA), eigenvalue ratio, and scree plot. The EFA was implemented in order to ensure that there was one dominant factor explains most of the variances in the model (Hambleton et al., 1991). Eigenvalue ratio between the first factor to the second one was checked as well. To obtain the eigenvalue ratio, Lord' criteria of the difference between the first and second eigenvalues divided by the difference between the second and third eigenvalues was applied (Lord, 1980). The larger the value obtained from this calculation, the more accurate the claim to meet the unidimensionality assumption. Also, unidimensionality of the measure was checked visually by using scree plot test (Hambleton & Swaminathan, 1985). Unidimensionality is also implicitly required for CTT analysis; therefore, checking this assumption was obtained for the analysis of both methods.

Model-fit analysis. In any application of the IRT model, testing the model fit is very important to check if the selected IRT model is valid for further analysis. The misfit between the data and the selected IRT model may lead to invalid parameter estimates (Fan, 1998). Theoretically, the 3PL model usually shows better model fit than the 1PL model and the 2PL model when data are dichotomously scored; however, the three models were tested for the fit to data and the results were compared. The appropriate

IRT model that adequately represented the observed data was selected for data analysis based on the likelihood ratio test statistics G^2. Because the underlying model was chosen based on IRT, three IRT models (1-plm, 2-plm, and 3-plm) were tested to obtain evidence supporting model-data fit. The G^2 is a fit index that is used to check how well the selected IRT model fits the response data. In theory, fitting an IRT model to a given data means that the underlying latent construct explains all the covariance among the test items (Steinberg & Thissen, 2006). The likelihood ratio test for model comparison was calculated using the following equation:

$$\Delta\mathrm{G}^2 = -2\,In(L_R) - \left(-2\,In(L_F)\right) = G_R^2 - G_F^2, \qquad (6)$$

where L_R is the maximum likelihood of the reduced model and L_F is the maximum likelihood of the full model. Then the relative improvement in the proportion of variability accounted for by one model over the other was assessed using R_Δ^2, which can be obtained from the following equation:

$$R_\Delta^2 = \frac{(G_R^2 - G_F^2)}{G_R^2} \qquad (7)$$

DIF analysis. Typically, DIF analysis is based on a comparison between two groups; the reference group or the group to which it is compared, and the focal group or the group of primary interest. In this study, there were a focal group and a reference group for each set of DIF analysis. For gender comparisons, male was set as the reference group and female was the focal group. Whereas, in the linguistic background comparisons, non-ELL students were set as the reference group and the ELL students were set as the focal group. In order to answer the research questions, six sets of DIF analysis was conducted. The first set identified gender-related DIF regardless their linguistic background (male vs. female). The second set examined the male group between non-ELL and ELL groups (males non-ELL vs. males ELL). The third set examined the female group between non-ELL and ELL groups (females non-ELL vs. females ELL). The fourth set examined DIF between non-ELL and ELL groups after matching them on their ability level (non-ELL vs. ELL) regardless their gender. The fifth set investigated gender-related DIF within ELL group (males ELL vs. females ELL). Finally, the sixth set examined gender-related DIF within non-ELL group (males non-ELL vs. females non-ELL).

For each set of the DIF analysis, two DIF detection methods were used to compare between-group and within-group for any potential DIF items. One method was chosen based on CTT (the MH procedure), and the other method was chosen based on IRT (the likelihood ratio statistics). In the next section, the implementation and the underlying equations used in the analysis for each method was discussed in detail followed by how the DIF magnitude or the effect size was calculated for each DIF method.

Mantel-Haenszel procedure. The MH procedure was used for the data analysis because it has been applied commonly in the literature of DIF. The MH procedure was calculated using Xcalibre version 4.1.8 software. The coefficient is a weighted average of the odds ratios for each level of θ. If the odds ratio is less than 1, then the item is more likely to favor the reference group than the focal group. Similarly, if the value of odds ratio is greater than 1, then the item favors the focal group. For each set of comparison, the response matrix was calibrated after matching the examinees on their ability level using the following equation:

$$\alpha_k = \frac{c_{R_K}\, I_{F_K}}{c_{F_K}\, I_{R_K}}, \qquad (8)$$

where *C* and *I* are the correct and incorrect responses, respectively. The *R* denotes to the reference group and *F* denote to the focal group. However, the MH procedure is not sensitive to the nonuniform DIF, therefore; nonuniform DIF items could not be identified using this method. Holland and Thayer (1988) proposed measuring the amount of DIF or DIF effect size in a delta scale as suggested by the Educational Testing Service.

Likelihood ratio statistics. Likelihood ratio was implemented using Xcalibre version 4.1.8 software for parameter estimations. After fitting an IRT model to the data, parameters were estimated for each item in the test for both groups while keeping the parameters of the rest of the items constrained to be equal across two groups (DIF model). Next step was to fit the same IRT model to both groups while all item parameters were constrained for both groups including the studied item (no-DIF model). The last step in implementing likelihood ratio was to calculate the $TSW\text{-}\Delta G^2$:

$$\text{TSW-}\Delta\text{G}^2 = \text{G}_2^2 - \text{G}_1^2, \qquad (9)$$

where G_2^2 is the no-DIF model and G_1^2 is the DIF model. The likelihood ratio is similar to χ^2 distribution with degree of freedom equals to the difference in the number of parameters estimated in the two models. In this study, the parameter estimates for all 34 items in the no-DIF model were constrained to be equal for both the reference and focal groups. In the DIF model, all item parameters were set to be equal for both groups except for the studied item. The constrained items were referred to as anchor items during the data analysis. For example, item 1 was unconstrained in the reference and focal groups while item 2 through 34 were set as anchor items. The same step was repeated for all the 34 test items. Therefore, in this study, there were six DIF analyses. Each analysis had one no-DIF model calibration run and 34 DIF model run, which leaded to a total of 35 calibration runs for each set of the analysis to estimate the likelihood ratio statistics. For the six pairs of reference and focal groups, 210 separate model calibration runs were needed. The last step was to compare the likelihood ratio statistics of the no-DIF model with all the 34 DIF models for each set of analysis. The item parameters were estimated for the reference and the focal groups separately. For each studied item, the item parameters were estimated for both reference group and focal group separately. Camilli and Shepard (1994) suggested 7 steps to examine DIF for likelihood ratio statistics. They suggested creating two items from the studied item, one for the reference group and the other for the focal group. Therefore, a response matrix was created for each studied item in all the comparison sets where the reference and focal groups had different column for the same item.

Calculating the DIF effect size. Effect size or the direction of the DIF was calculated using the natural logarithm of the odds ratio divided by its standard error which transforms the value into a standardized matrix. As recommended by the ETS, DIF is classified into six categories: A category which refers to a negligible DIF favoring the focal group, B category which refers to a intermediate DIF favoring the focal group, C category which refers to a large DIF favoring the focal group, – *A* category which refers to a negligible DIF favoring the reference group, – B category which refers to a intermediate DIF favoring the reference group, and – C category which refers to a large DIF favoring the reference group. In terms of the MH procedure, the magnitude of DIF is classified as negligible or *A* category if $|\Delta_{MH}|$ < 1, moderate or *B* category if 1 ≤ $|\Delta_{MH}| > 1.5$, and large or *C* category if $|\Delta_{MH}|$ ≥ 1.5. The MH D-DIF is obtained from the following index:

$$\text{MH D-DIF} = - 2.35 \text{ In } (\hat{\alpha}_{MH}) \qquad (10)$$

In the likelihood ratio, effect size was calculated using Cohen's G^2 statistics in which the negligible level A is obtained if 3.84 < G^2 < 19.4, the moderate level B is obtained if 9.4 < G^2 < 41.9, and large level C is obtained if G^2 > 41.9 when α = .05. It also considers the negative sign as an opposite direction of the DIF item, which means that the item favors the focal group.

This chapter stated the methodology of how the data would be analyzed and interpreted. The septs of analyzing the data were clarified. First, the sample size was chosen for each set of analysis from a pool of 7th grade students who took a high-stakes standardized test based on CCSS. Second, descriptive statistics were obtained for the whole test and for each set of comparison to check how representative the sample was. Third, the assumptions required for CTT and IRT analysis such as unidimensionality and matching criteria were examined before conducting any further analysis. The last step was to conduct the DIF analysis based on the CTT and IRT methods. Also, the underlying equations were listed and clarified. The results obtained from this analysis were represented in detail in the following chapter.

Chapter IV: Results

Introduction

As the world is moving fast toward development in all domains, education is always considered as one of the most significant indicators for the growth of any society. Evaluating the quality of education goes through several steps one of these steps is to evaluate the educational outcomes. As Tanner (2001) stated educational tests are the tool that educators and policy makers use to evaluate and understand the teaching and learning processes. Therefore, the importance of the educational assessment cannot be ignored.

This study was designed to investigate gender-related DIF, linguistic background DIF, and gender-DIF cross and within different linguistic background groups using data from a high-stakes test based on CCSS. The data were analyzed by implementing two different DIF detection methods to check the consistency of the results between these two methods. For the purpose of this quantitative study, data were first analyzed to evaluate the general psychometric properties of the test items. Then two different psychometric methods were used to detect DIF. One is based on CTT (the MH procedure) and the other one is based on IRT (likelihood ratio statistics). Therefore, this chapter presented a detailed report of the data analysis for the stated research questions. The descriptive statistics were first reported followed by checking the unidimensionality assumption and the model-fit statistics. The MH procedure for detecting DIF was reported, then IRT-likelihood ratio statistics for examining DIF was reported. The last section of this chapter summarized the overall results from this study through the lens of literacy and psychometrics.

Descriptive Statistics

As indicated in Table 4.1, a summary of the descriptive statistics presented the characteristics of the subgroups used in the analysis. The mean and standard deviation (SD) of the test for all gender groups, male cross non-ELL and ELL groups, female cross non-ELL and ELL groups, ELL status, ELL-gender, and non-ELL gender are listed, respectively. It is clear that the mean was high in the gender group and the non-ELL students. When ELL students were included in estimating the average total score, the average score dropped.

Table 4. 1

Descriptive statistics for each set of DIF analysis

The set of analysis	reference and focal groups	*N*	*M*	*SD*	α
gender groups	male	316	18.37	9.18	.93
	female	316			
male cross non-ELL and ELL groups	non-ELL	340	15.54	7.05	.86
	ELL	340			
female cross non-ELL and ELL groups	non-ELL	502	15.93	6.70	.85
	ELL	502			
ELL status	non-ELL	280	17.07	8.23	.91
	ELL	280			
ELL -gender	male	402	15.48	6.47	.83
	female	402			
non-ELL-gender	male	831	18.93	8.37	.91
	female	831			

The reliability coefficient for the whole test and all subgroups was checked using the Cronbach's alpha coefficient. The results showed that $\alpha = .73$ for the whole test and above $\alpha = .80$ for subgroups, which indicated that the items are consistently measuring the same construct. The α was considered acceptable comparing with many commercially available tests in which α is roughly .90.

Dimensionality

Whereas there are a number of approaches that are used to assess the dimensionality assumption, there is no clear cutoff point to determine the number of factors that should be retained (Yau et al., 2015). However, there is a sufficient consensus among researchers to use eigenvalue ratio as an indicator for unidimensionality. The unidimensionality assumption of the test was obtained by applying EFA. The results revealed that the percentage of variance explained by the first factor was 18.14 %, and the variance explained by the second factor was 3.88 %. The first eigenvalue was 6.17, the second eigenvalue was 1.32, and the third eigenvalue was 1.04. The ratio of the first eigenvalue to the second one was 17.32, which indicated that the eigenvalue of the first factor is larger enough than the next two eigenvalues. In addition, the inspection of the scree plot was checked visually to confirm the unidimensionality of the test. This method is usually used to evaluate the unidimensionality assumption by checking if the drop between the first and second eigenvalues is trailed off like the scree at the foot of a mountain, which indicates that the eigenvalues before the drop represent the number of dominant factors (DeMars, 2010). Figure 4.1 showed the graphed eigenvalues which represented a big drop between the first and second eigenvalues. Therefore, the results of the EFA analysis suggested that the unidimensionality assumption appeared to be met across the whole test.

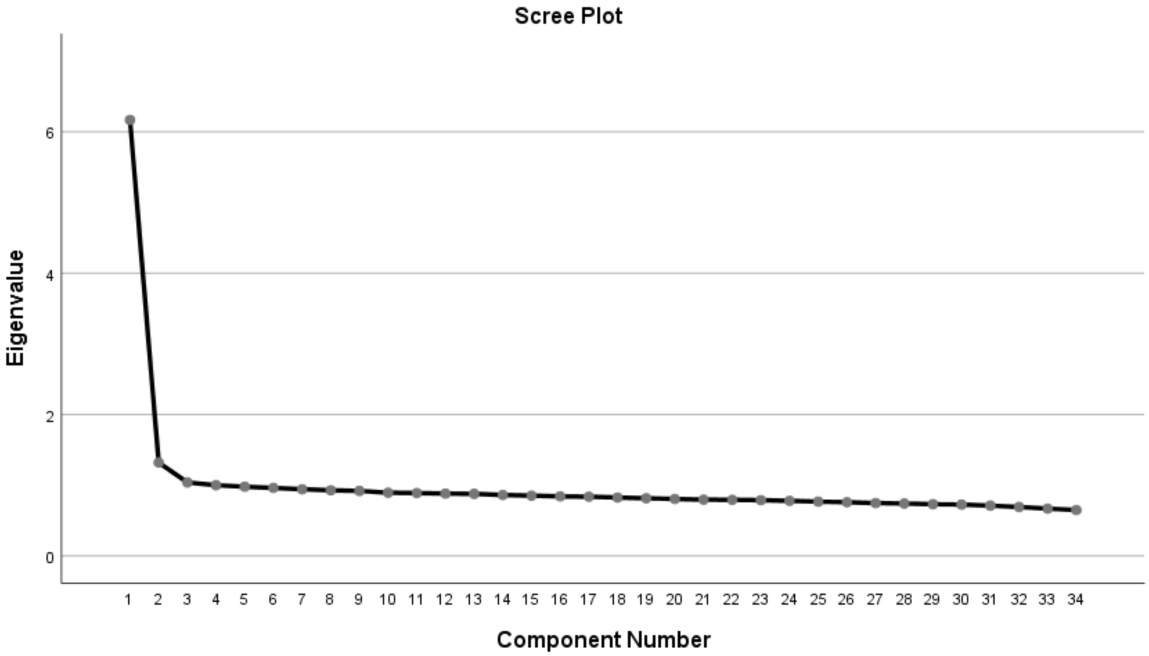


Figure 4. 1 the scree plot for the eigenvalues of the factors which clearly starts from the second eigenvalue.

CTT Analysis

The psychometric properties were evaluated based on CTT. The *P* value or the difficulty parameter was measured by calculating the proportion of examinees who answered the item correctly. The *P* values range from 1.0 to 0.0. The values close to 1.0 are indicative of an easy item. In addition, the item point-biserial correlation with total score or the classical discrimination parameter was calculated. The value of the classical discrimination parameter ranges from – 1.0 to 1.0. However, for dichotomously scored items, it is very rare to obtain values above 0.50 (Camilli & Shepard, 1994). A positive high value represents a better item to differentiate between examinees. Table 4.2 summarized the values of both parameters as estimated by the CTT. The results showed that the classical item difficulty parameters were ranged from easy to moderate. Item 19 was the most difficult item with P = 0.85, and item 5 was the easiest with P = 0.28. Whereas, the classical item discrimination parameters showed a relatively good discrimination strength to differentiate between high-ability examinees and low-ability examinees. Item 3 showed the strongest discrimination ability comparing with all the other items in the test r = 0.49, while item 19 had the lowest value, r = 0.15.

IRT Analysis

In order to conduct IRT analysis, first we had to choose from the three logistic IRT models the best model that fits our data. The model comparison revealed a significant improvement in fit by the 2plm over the 1plm, $\Delta G^2(34) = 12{,}641.92$. An analogous comparison between the 2plm and the 3plm showed also a significantly improvement in the overall fit between the two models, ΔG^2 $(34) = 1{,}875.85$. The R^2_Δ statistics indicated that 2plm resulted in 2.6% improvement in fit over the 1plm. In addition, the R^2_Δ between the 2plm and 3plm showed that the 3plm resulted in about 0.4% improvement in the model fit. In general, more complex models tend to fit the data better than less parameter models. The above analysis revealed that 3plm was the best model fit the data. Therefore, 3plm IRT was chosen as the model for the data analysis in this study. Data were calibrated to estimate IRT item parameters based on the 3plm-IRT. Item parameters a, b, and c were estimated. Typically, a-parameter values less than 0.5 are considered low and the item does not differentiate strongly between examinees. As represented in Table 3, all a-parameters showed good ability to discriminate between high ability and low ability examinees. Item 32 had the larger value which indicated that it differentiated very well between ability levels. Whereas, item 22 had the least discrimination ability comparing with all test items. The *b*-parameter values showed in Table 4.2 indicated that most of the test items showed moderate level of difficulty. However, the values of items 19 and item 27 exceeded 2.0, which indicated that these two items were difficult, while item 8 was the easiest.

Table 4. 2

CTT and IRT parameter estimations for the whole test

	CTT		*IRT*		
Item #	*P* value	*Pearson point-biserial correlation*	*a*	*b*	*c*
1	0.8405	0.3471	0.8066	-1.4749	0.2068
2	0.6477	0.2899	0.4884	-0.3238	0.2208
3	0.5911	0.4936	1.1094	-0.0376	0.1871
4	0.7816	0.3836	0.8357	-1.0069	0.2255
5	0.8539	0.363	0.9414	-1.4884	0.1967
6	0.5484	0.3381	0.6182	0.3184	0.2139
7	0.6548	0.3358	0.5877	-0.3722	0.2096
8	0.5818	0.3574	0.6006	-0.0176	0.176
9	0.377	0.2747	0.777	1.4014	0.2207
10	0.5273	0.3208	0.5438	0.3719	0.1844
11	0.5118	0.4581	1.0403	0.3514	0.1926
12	0.7806	0.3161	0.5811	-1.3093	0.1884
13	0.577	0.3859	0.6935	0.0237	0.1816
14	0.7237	0.4811	1.1495	-0.6612	0.1818
15	0.5755	0.4517	0.8324	-0.0699	0.1459
16	0.6912	0.4033	0.7679	-0.5561	0.1983
17	0.3449	0.2876	1.0086	1.4216	0.2131
18	0.475	0.367	0.7347	0.623	0.1904
19	0.2836	0.1543	1.0256	2.0974	0.2222
20	0.4918	0.3755	0.7992	0.5769	0.2107
21	0.442	0.4213	0.9161	0.6618	0.1718

	CTT		*IRT*		
Item #	*P* value	*Pearson point-biserial correlation*	*a*	*b*	*c*
22	0.4844	0.2635	0.4135	0.6996	0.1623
23	0.4804	0.3794	0.7411	0.5589	0.1827
24	0.5406	0.4187	0.7803	0.191	0.1747
25	0.3608	0.2767	0.8592	1.4474	0.2195
26	0.3224	0.371	1.1135	1.203	0.1584
27	0.3356	0.1761	0.7732	2.0284	0.2471
28	0.5319	0.4303	1.0222	0.3635	0.2274
29	0.4185	0.3354	0.615	0.9114	0.1564
30	0.3764	0.4086	1.0075	0.93	0.1585
31	0.4895	0.4714	1.0672	0.4087	0.1778
32	0.4327	0.3371	1.1763	1.0351	0.2653
33	0.3915	0.3274	0.8956	1.1277	0.2079
34	0.5573	0.3802	0.7419	0.2462	0.218

DIF analysis

The study was designed to examine DIF between different groups and through two different methods: the CTT-based MH procedure and IRT-based likelihood ratio. The DIF analysis was also carried out in six sets: (1) gender groups (male vs. female), (2) male cross non-ELL and ELL groups (male non-ELL vs. male ELL), (3) female cross non-ELL and ELL groups (female non-ELL vs. female ELL), (4) ELL status (non-ELL vs. ELL), (5) gender within non-ELL group (male non-ELL vs. female non-ELL), and (6) gender within ELL groups (male ELL vs. female ELL.) The last step was to compare the consistency of the results obtained from both DIF detection methods. DIF results for each set of analysis were discussed in detail in the following sections for each detection method.

The MH procedure. Using the MH technique, the X^2 values were obtained from the Xcalibre software 4.1.8 to identify which items functioning differently in all the six sets. According to the MH analysis, all the subgroups analysis revealed non-significant DIF between groups except the gender analysis within the non-ELL group. Only item 15 was flagged as bias against male group of non-ELL students (see Table 2), which indicated that less than 3 % of the test items contained DIF. The DIF effect size analysis revealed that item 15 favored females over males with two standard units in difference Δ_{MH} = - 2.01, which is considered as large DIF or category - C DIF. Negative sign indicates that the item favors the focal group which is the female group in this study. This means that females were about 2 times more likely to answer the item correctly than males. Based on the CCSS categories, item 15 evaluates students' writing ability in organizing the written text. In this version of the test, item 15 was the only item in the test which was designed to measure the organizing texts under the writing category.

IRT-likelihood ratio. IRT-likelihood ratio statistics was distributed as X^2 with 12 degree of freedom. For all the 34 items examined for DIF in all sets of the analysis, DIF was exhibited in only four sets of the analysis. The direction of DIF varied in all the comparison sets with DIF items. Some items favored the reference group and other favored the focal groups. DIF was designated in 6 items which presented about 17.6 % of the test items. Based on Cohen's G^2 statistics, the values of likelihood ratio presented moderate, and negligible magnitude level of DIF. The results of DIF analysis were discussed in detail for each set of analysis in the following sections. See Table 4.3 for summary of the DIF results obtained from the IRT-likelihood ratio.

Table 4. 3

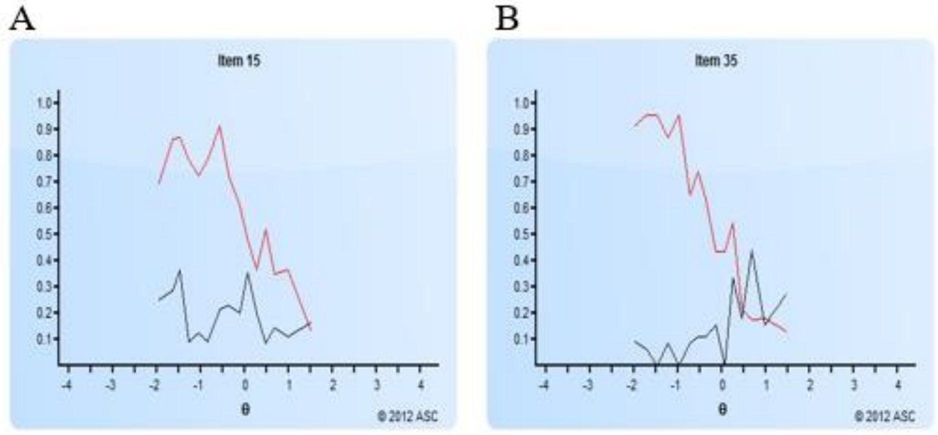
DIF results for all six comparison sets of analysis using likelihood ratio.

	item	ΔG^2	*Against*	*Effect size*
Gender groups	No DIF			
male cross non-ELL and ELL groups	15	-30.96	ELL	moderate
Female cross non-ELL and ELL groups	No DIF			
ELL status	26	22.86	ELL	moderate
ELL -gender	26	26.88	female	moderate
	30	23.74	female	moderate
non-ELL-gender	2	-22.61	male	moderate
	3	-44.63	male	severe
	14	-37	male	moderate
	15	-36.70	male	moderate

P = .05

Gender groups (male vs. female). The general comparison between male and female students revealed no DIF, which was consistent with the MH procedure results. There was no difference between the performance of males and females regardless their linguistic background.

Male cross non-ELL and ELL groups (male non-ELL vs. male ELL). The comparison between the same gender (e.g., male) cross two different linguistic background groups showed one DIF item (item 15). This item favored the non-ELL students over the ELL. Item 15 was designed to measure the ability of organizing a text under the writing category, and DIF effect size was moderate. Figure 4.2A showed the response pattern of male ELL in answering item 15. While Figure 4.2B represented the response pattern of the male non-ELL students in answering the same item. Table 4.4 displayed item parameter estimations for both male non-ELL and male ELL students which indicated that there were slight differences in the discrimination parameter and the difficulty parameter between both groups.



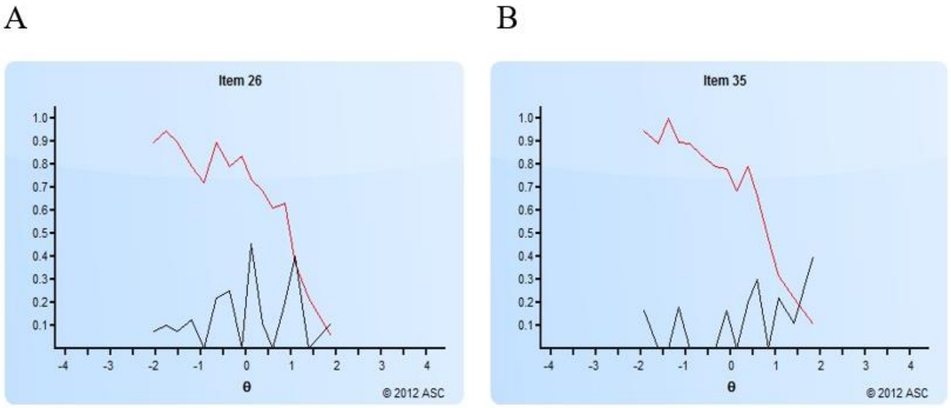
Item 15/ ELL *Item 15/ non-ELL*

Figure 4. 2 Item response function for male cross non-ELL and ELL groups.

Note. The black line represented the studied IRF of the male ELL students, while the red line represented the fit line.

Female cross non-ELL and ELL groups (female non-ELL vs. female ELL). In the comparison between female non-ELL and female ELL students, there were no DIF items. Female students from both linguistic groups had the same responses pattern to all the test items, which indicated that female students seemed to respond equally to all the test items.

ELL status (Non-ELL vs. ELL). DIF analysis between two different linguistic background students showed only one DIF item (item 26) favoring the non-ELL students. Based on the CCSS manual, item 26 was designed to measure how 7th grade students can contrast and compare fictional portrayal to reality and how authors used history in their work. Table 4.4 represented the item parameters for item 26 of non-ELL and ELL students. Figure 4.3 displayed the ICC for both groups.

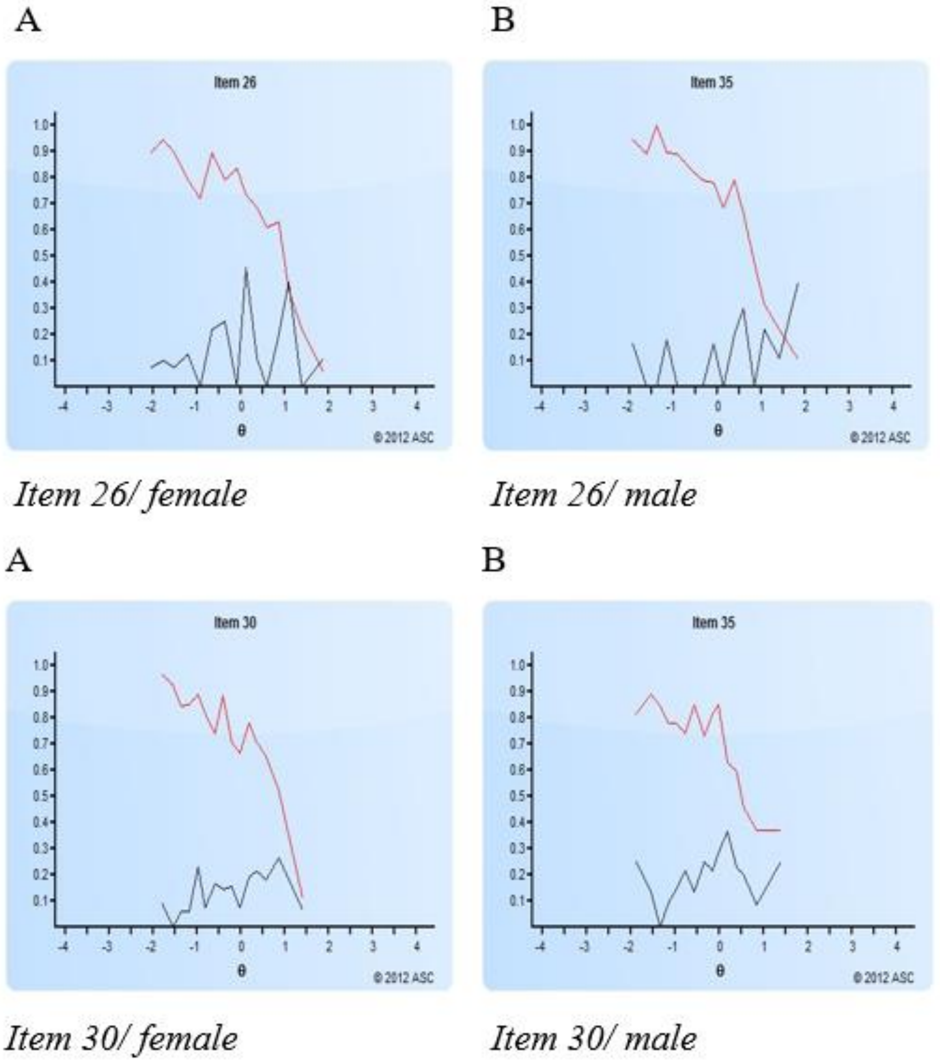


Item 26/ ELL *Item 26/ non-ELL*

Figure 4. 3 Item response function for ELL status.

Note. The black line represented the studied IRF, while the red line represented the fit line

ELL -gender (Male ELL vs. Female ELL). Two DIF items (item 26 and item 30) were detected in this comparison. Item 26 was designed to measure “compare and contrast historic fiction to reality under literature category”, whereas item 30 was designed to measure students’ proficiency of eliciting information from a written text by analyzing the main ideas. Both items favored male students over female students. Figure 4.4A illustrated the ICC for the DIF item of the female group, on the other hand, Figure 4.4B displayed the ICC for the male students. Also, Table 4.4 represented the item parameter estimations for male and female students for both DIF items.



Item 26/ female

Item 26/ male

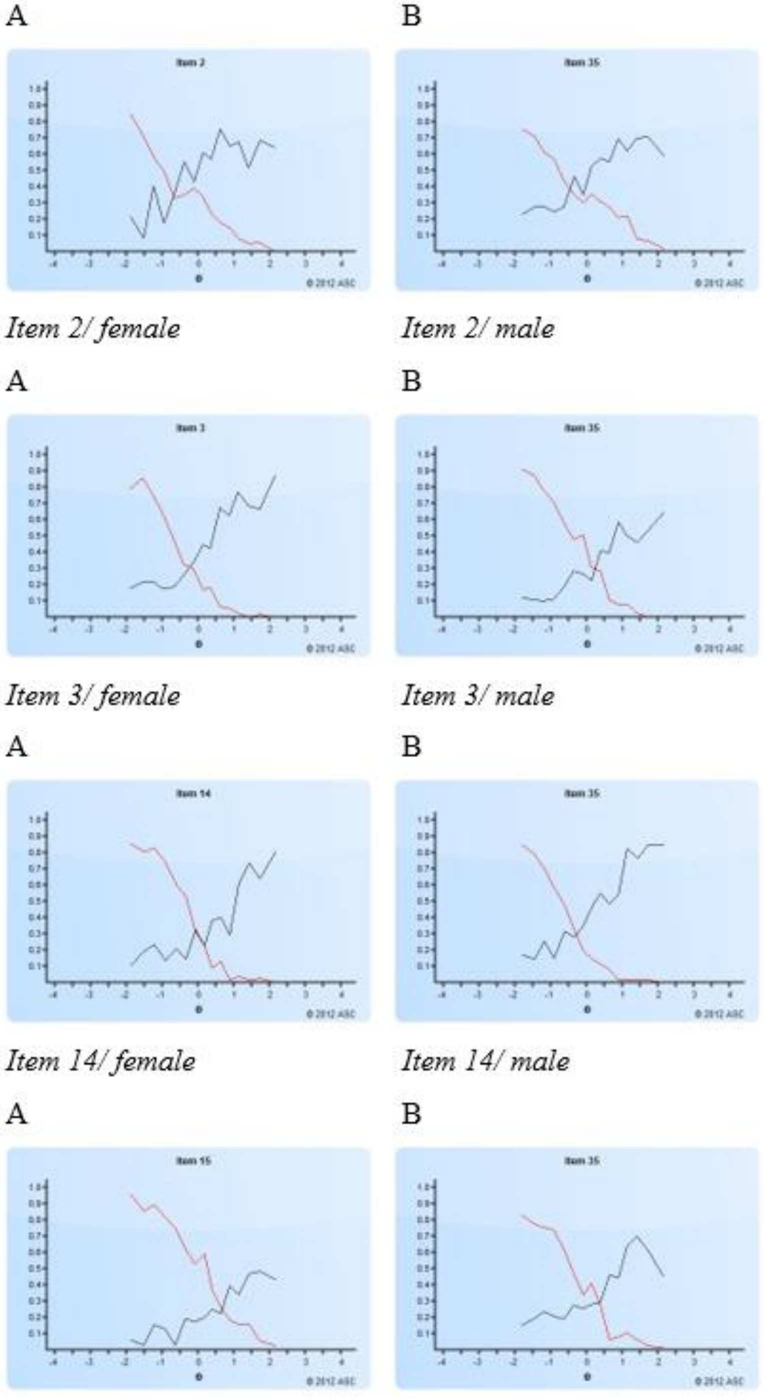
Item 30/ female

Item 30/ male

Figure 4. 4 Item response function for gender within ELL group.

Note. The black line represented the studied IRF, while the red line represented the fit line

Non-ELL -gender (male non-ELL vs. female non-ELL). There were four gender DIF items among the non-ELL students (items 2, 3, 14, and 15). All of them favored female students over male students. This set of comparison revealed more DIF items than other sets.



Item 2/ female

Item 2/ male

Item 3/ female

Item 3/ male

Item 14/ female

Item 14/ male

Item 15/female

Item 15/ male

Figure 4. 5 Item response function for gender within non-ELL group.

Note. The black line represented the studied IRF, while the red line represented the fit line.

Based on the CCSS manual, items 2 and 3 were designed to measure word meaning, item 14 was designed to measure plan, revise and edit, and item 15 was designed to measure organize text. Items 2 and 3 was classified under the language category. While items 14 and 15 were under the writing category. As illustrated in Figure 4.5As, the female ICCs were demonstrated, while Figure 4Bs displayed the male ICCs for the DIF items.

Table 4. 4

parameter estimations for references groups and focal groups in each comparison set

Comparison set		*a*	*b*	*c*
Male cross non-ELL and ELL groups				
Item 15	non-ELL	1.100	0.346	0.190
	ELL	0.950	0.745	0.210
ELL status				
Item 26	Non-ELL	1.298	1.209	0.240
	ELL	1.275	1.277	0.252
ELL-gender				
Item 26	male	1.298	1.209	0.240
	female	1.275	1.277	0.252
Item 30	male	1.289	1.271	0.246
	female	1.140	1.340	0.266
Non-ELL gender				
Item 2	male	0.773	-0.278	0.244
	female	0.677	-0.044	0.248
Item 3	male	1.197	0.362	0.229
	female	1.231	0.020	0.215
Item 14	male	1.362	0.133	0.217
	female	1.266	0.410	0.234
Item 15	male	1.058	0.450	0.189
	female	1.064	0.005	0.225

As it is depicted in the above figures, the IRF or ICC for each DIF item exhibited nonuniform DIF except for item 15 in the sixth comparison. To get more information of the differences between groups, item parameters were estimated separately. Table 4.4 summarized the 3PL model item parameters for each comparison set with DIF items. As we can see, the item difficulty parameters of the disadvantaged groups are larger than the advantaged groups in each comparison set.

The Agreement between MH Procedure and Likelihood Ratio

In order to inspect consistency between the CTT-MH procedure and IRT-likelihood ratio results, the percentage of pairwise agreement was calculated. Although, both methods agreed upon only one item, the percentage of agreement was about 16.7 %. There were discrepancies between the two DIF analyses results. The agreement between the MH procedure and likelihood ratio was only upon one item (item 15), which was the only item flagged in the MH analysis. Likelihood ratio revealed more DIF items than the MH procedure. Moreover, the DIF analysis using likelihood ratio showed DIF in four comparison sets out of six, whereas MH procedure flagged only one item in one comparison set.

Summary

In this chapter, an introduction was given to highlight the purpose of the study. After data were analyzed using two different DIF detection methods, MH procedure and IRT-likelihood ratio, the results were reported in detail. Six sets of comparison were applied. Then the DIF results were compared between the both detection methods for any agreement. Although the agreement was upon one item, this item was the only one flagged by applying the MH procedure. IRT-likelihood ratio analysis revealed more DIF items cross most of comparison sets. The impact of gender on the non-ELL and ELL groups were clear in the results obtained from this study. The results of this study were discussed in detail in the next chapter in the light of the research questions.

CHAPTER V: DISCUSSION

Introduction

Evaluating the psychometric properties of test items and the examinees' responses should be an essential step of developing a new measure and/or adopting an existing one. In the psychometric realm, there are many methods and approaches used for examining and evaluating these properties. The most commonly applied methods are CTT and IRT. CTT represents the traditional psychometric method whereas IRT is considered as a relatively new method. Each method has advantages and disadvantages that make it preferable by many researchers. Despite the weak theoretical assumptions underline the CTT, it is a reasonably simple method to apply and to interpret. Unlike CTT, IRT is based on very strong assumptions that can be falsified or proven (Fan, 1998). Under CTT and IRT, there are many other psychometric applications that can be applied to conduct further analysis. One of these applications is the DIF analysis. DIF is detected when items on a given test do not function equivalently between two or more subgroups driven from the same population after matching them on their ability level (Angoff, 1993). In this study, DIF analysis was applied under CTT and IRT to examine a high-stakes test for any DIF items between two gender groups, gender within non-ELL and ELL groups, gender cross non-ELL and ELL groups, and between non-ELL and ELL groups.

In the preceding chapter, the analysis and presentation of data were reported in detail. To the interest of the researcher, the following research questions were raised to examine a high-stakes test for any potential DIF items:

1. Is there any gender-related DIF on the 7th grade CCSS ELA items (male vs. female)?
2. Is there any ELL-related DIF on the 7th grade CCSS ELA items (non-ELL vs. ELL)?
3. Is there any gender-related DIF within ELL and non-ELL the 7th grade CCSS ELA items (male ELL vs. female ELL, and male non-ELL vs. female non-ELL)?
4. Is there any gender-related DIF across ELL and non-ELL groups (male ELL vs. male non-ELL, and female ELL vs. female non-ELL)?
5. Is there any difference between CTT and IRT DIF results?

The current chapter presented detailed discussion of the findings in the light of the above research questions. This chapter consisted of discussion of the findings, implications for practice, limitation of the study, and recommendations for further research.

Discussion of the Findings

DIF analysis is a key component of the process of checking the validity and fairness of any measurement. In this study, DIF analysis was conducted to compare the performance of gender groups, gender within non-ELL and ELL groups, gender cross non-ELL and ELL groups, and non-ELL and ELL groups using the MH procedure and IRT-likelihood ratio statistics. Gender-related DIF and different linguistic background were discussed in literature and results revealed similar conclusions. The findings were discussed in the light of each research question.

Research question one: Is there any gender-related DIF on the 7th grade CCSS ELA items (male vs. female)?

Based on the MH procedure, there were no DIF items in the comparison between males and females. Also, IRT-likelihood ratio statistics revealed no gender-related DIF items. Ryan and Bachman (1992) found no DIF between male and female when they examined the gender-related DIF in the Test of English as a Foreign Language (TOEFL) using the MH procedure. Their results aligned with what was revealed by this study about the gender-related DIF. On the other hand, gender-related DIF was found in many other research studies. Aryadoust (2012), for example, found one DIF item favored male students when he analyzed data obtained from IELTS listening subtest using Rasch model as a detection method.

Research question two: Is there any ELL-related DIF on the 7th grade CCSS ELA items (non-ELL vs. ELL)?

The results obtained the MH procedure showed no DIF between non-ELL and ELL. However, IRT likelihood ratio method revealed one DIF item favored the non-ELL students with moderate DIF level. The item measures historical fiction under the literature category which may require the students to have robust background knowledge of the historical characters, places, and/or time period.

Research question three: Is there any gender-related DIF within ELL and non-ELL the 7th grade CCSS ELA items (male ELL vs. female ELL, and male non-ELL vs. female non-ELL)?

Regarding this question, the MH procedure identified only one DIF item. IRT-likelihood ratio identified four items. Item 15, which was the common DIF item from both DIF detection methods, favored the female non-ELL students on the non-ELL gender comparison set and favored the male non-ELL when DIF was examining between male ELL and male non-ELL students. It was designed to measure the students' writing ability and how they can write a clear and coherent paragraph in which the organization, development, and style are appropriate to the writing task and the audience. Item 14 was also classified under the same CCSS category. It also favored female ELL students. However, item 2 and 3 were classified under language and word meaning standard. They mainly measured the knowledge of vocabulary.

Alavi and Bordbar (2017) found gender-related DIF in a National University Entrance Exam for Foreign languages. All the students who participated in the study and took the test to enroll in college were ELL students. They found 40 DIF items out of 95 items in which some favored males and other favored females. Therefore, the results were inconsistent across gender groups. Also, they found vocabulary items favored female over male students, which seemed be consistent with the overall findings of this study.

Research question four: is there any gender-related DIF across ELL and non-ELL groups (male ELL vs. male non-ELL, and female ELL vs. female non-ELL)?

The comparison between gender cross linguistic background groups revealed differences between gender when male non-ELL students were compared with their ELL peers. Male non-ELL students outperformed male ELL students on one item out of 34 items. On the other hand, there were no DIF items when female non-ELL students were compared with their ELL peers.

Research question five: Is there any difference between CTT and IRT DIF results?

Despite the results obtained from the MH procedure and IRT-likelihood ratio were not consistent in the number of DIF items detected by both methods, there were relative agreement upon the DIF items. Both methods detected the same item (item 15) as DIF. the MH procedure does not detect nonuniform DIF, which may explain this disagreement between both methods. These results aligned with the claim of the superiority of IRT-based methods on detecting DIF. Also, the results were consistent with a study conducted by Acar (2012). He compared the DIF results obtained from a CTT-based approach using logistic regression technique and an IRT-based approach using likelihood ratio. He detected less DIF item when the logistic regression technique was applied, which indicated that IRT-likelihood ratio is more powerful in detecting DIF than other methods based on CTT. In addition, it was clear from the figures that all the DIF items showed nonuniform DIF which cannot be detected by the MH procedure. That can explain the disagreement between these two methods. Another reason for the lack of agreement between the MH procedure and the likelihood ratio in the identification of DIF was because the MH procedure cannot detect uniform type of DIF. It was observable from the figures that all the items response functions were crossing at a certain point at the ability continuum.

In this study, DIF analysis was conducted in six comparison sets. Item 15 was detected as DIF item in two comparison sets. It was found as DIF item when male students were compared cross non-ELL and ELL groups favoring male non-ELL, on the other hand, it was identified as DIF item when male non-ELL students were compared with female non-ELL students favoring female non-ELL with a moderate level of DIF in both comparison sets. Item 26 was also identified as DIF item in two comparison sets.

Non-ELL students outperformed their ELL peers on item 26, while male ELL outperformed female ELL on the same item with a moderate level of DIF in both comparison sets. Obviously, the comparison between non-ELL and ELL on item 26 revealed bias against ELL group. When the ELL group was broken down into two gender groups, the analysis showed that female students were the disadvantaged group. This result indicated that gender may have an impact on students' performance on the test.

Implications for Practice

The era of high-stake tests in the United States has started since the middle of the last century (Marzano, 2018). Since then depending on these tests outcomes has been the guideline for most of the educational reforms. Therefore, the findings of this study could be useful in interpreting the educational tests outcomes that affect the whole educational system. There were only four DIF items detected in this study and the DIF magnitude was negligible in one item and moderate in the other three, which indicated that these items can be used in the test and may not need further examination (Zwick, 2012). In DIF analysis, all DIF items should be reported regardless their magnitude. However, in practice, only severe DIF items should be deleted from the test or be considered for further review. Moderate DIF items may need more expert' revision. Negligible DIF items are not considered as problematic; therefore, they are kept in the test with no further reviewing. These results implied that this high-stakes test outcome can be trustworthy.

An important way in which this study extended the previous DIF research is by examining impact of gender on student LEA performance from different linguistic background and the impact of gender within each language group. The findings of this study have important implications on DIF research and literacy studies. These results suggested that in order to achieve a valid and fair test, test developers should consider the content of each item and the sensitivity of each content to DIF. The results of this study also provided insight into how gender differences impact ELL status. It appeared likely gender played a role on how students responded to some certain items that belong to categories of the CCSS. Therefore, not only students' linguistic background should be taken into consideration, but also their gender when the classroom instructions are implemented. From a practical point of view, practitioners and policymakers need to be aware of gender differences when the population is broken down to subgroups. Also, data used in this study were collected from adolescent students in 7th grade which can contribute to the field of DIF because there are no DIF study used this particular age group.

Limitation of the Study

Although the findings of this study were informative for test developers and educators, there are several limitations associated with this research study. Zumbo (2007) claims that it is time for the third generation of DIF, which focuses on investigating the reasons beyond DIF items. However, in this study, the researchers do not have full access to the test items used in the analysis. Only item descriptions were provided by CCSS categories and the testing company. Therefore, the reasons of manifesting potential DIF items in the test cannot be fully explained.

Another limitation was that the level of students' proficiency in their 1st language is not controlled. The recommendations to control the first language proficiency level is to test them in their native language or at least ensure that they have finished a sufficient level of proficiency (Abbott, 2007).

Recommendations for Further Research

The purpose of this study was to evaluate a high-stakes standardized test for any potential gender-related and linguistic background DIF items. For this purpose, data were analyzed using CTT and IRT psychometric methods, then the MH procedure and IRT-likelihood ratio were implemented to detect DIF. The data were only analyzed quantitatively. Using mixed method data analysis may lead to a robust framework and better understanding to the DIF statistics. The qualitative data analysis could be applied by using experts review who have enough knowledge about the test content, structure, and format to undertake a qualitative investigation (Aryadoust, 2012).

Comparing parameter estimation between CTT and IRT also could be an extension to this research study. A correlation study could be conducted to check the level of consistency between the parameter estimations under each method. The results of this study are of primary importance to the test developers, educators, policy-makers, and researchers. The results revealed that the gender of the ELL students can impact their performance in the test.

Conclusion

This study aimed at investigative the occurring of gender-related DIF and the impact of the ELL status in a high-stakes test designed to measure the CCSS categories. Two DIF detection methods were used, the MH procedure and the IRT-likelihood ratio statistics. Based on the findings of this study; it could be concluded that there was gender-related DIF in the standardized test that measures the CCSS categories when the responses of each gender group analyzed separately with regarding to their linguistic background. In addition, the findings of this study suggested that interpretation the outcomes of a high-stakes standardized test based on CCSS can be relatively accurate. The results were satisfactory to the researchers. Since the data used in this study were provided from a high-stakes test, they were supposed to be driven from a well-designed test. The findings indicated the importance of considering gender-related DIF within and cross non-ELL and ELL students. While gender-related DIF was not exhibited in gender analysis regardless their linguistic background, male and female students seemed to respond differently in some items when their linguistic background was taken into consideration. Although there were four DIF items detected in the analysis, the results were very satisfactory to the researchers because high-stakes tests are designed with great care that only the best functioning items were used.

The DIF items detected in the test inconsistently favored one group over the other. They favored males when the comparison was within the ELL group. On the other hand, they favored females when the comparison was within the non-ELL group. In general, the DIF analysis pointed to the conclusion that were more likely related to the content of the test items rather than the group memberships. The ELL status also seemed to have minor impact on the examinees' response. Wyse and Mapuranga (2009) argued that native speakers do not always outperformed foreign students when their study revealed that students who took the PISA test in their native languages had DIF items in some versions of the test.

Therefore, it could conclude that IRT has a powerful and comprehension analysis, which leads to more accurate insight of the analyzed data (Thissen & Steinberg, 1988).

In addition, the advantage of IRT invariant parameter estimations has been claimed to provide less measurement error than CTT (Teresi et al., 2000). In this study, results obtained from IRT included more DIF items

REFERENCES

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing*, *24*(1), 7-36.

Abedlaziz, N., Ismail, W., & Hussin, Z. (2011). Detecting a gender-related DIF using logistic regression and transformed item difficulty. *Online Submission,* 734-744. Retrieved from https://ezproxy.mtsu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED527685&site=ehost-live&scope=site

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.

Acar, T. (2012). Determination of a differential item functioning procedure using the hierarchical generalized linear model: A comparison study with logistic regression and likelihood ratio procedure. *SAGE Open*, *2*(1) doi:10.1177/2158244012436760

Ahmadi, A., & Thompson, N. A. (2012). Issues affecting item response theory fit in language assessment: A study of differential item functioning in the Iranian national university entrance exam. *Journal of Language Teaching & Research, 3*(3), 404-412.

Akour, M., Sabah, S., & Hammouri, H. (2015). Net and global differential item functioning in PISA polytomously scored science items: Application of the differential step functioning framework. *Journal of Psychoeducational Assessment, 33*(2), 166-176. doi:10.1177/0734282914541337

Alavi, S. M., & Bordbar, S. (2017). Differential item functioning analysis of high-stakes test in terms of gender: A Rasch model approach. *Malaysian Online Journal of Educational Sciences, 5*(1), 10-24. Retrieved from https://ezproxy.mtsu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1125062&site=ehost-live&scope=site

Albano, A. D., & Rodriguez, M. C. (2013). Examining differential math performance by gender and opportunity to learn. *Educational & Psychological Measurement, 73*(5), 836-856. doi:10.1177/0013164413487375

Allalouf, A., & Abramzon, A. (2008). Constructing better second language assessments based on differential item functioning analysis. *Language Assessment Quarterly, 5*(2), 120-141.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing & student learning. *Education Policy Analysis Archives, 10*(18). Retrieved from http://epaa.asu.edu/epaa/v10n18/.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics, 37*(3), 387-416. Retrieved from https://ezproxy.mtsu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ968401&site=ehost-live&scope=site http://dx.doi.org/10.3102/1076998611411913

Angoff, W. H. (1972). *A technique for the investigation of cultural differences*. Paper presented at American Psychological Association meeting, Honolulu, HI.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Angoff, W. H., & Cook, L. L. (1988). Equating the scores of the Prueba de Aptitud Académica and the scholastic aptitude test. *ETS Research Report Series, 1988*(1), 18.

Arikan, S., van de Vijver, Fons, & Yagmur, K. (2018). Propensity score matching helps to understand sources of DIF and mathematics performance differences of Indonesian, Turkish, Australian, and Dutch students in PISA. *International Journal of Research in Education and Science, 4*(1), 69-81.

Aryadoust, V. (2012). Differential item functioning in while-listening performance tests: The case of the international English language testing system (IELTS) listening module. *International Journal of Listening, 26*(1), 40-60. doi:10.1080/10904018.2012.639649

Atalay Kabasakal, K., Arsan, N., Gök, B., & Kelecioglu, H. (2014). Comparing performances (type I error and power) of IRT likelihood ratio SIBTEST and mantel-haenszel methods in the determination of differential item functioning. *Educational Sciences: Theory and Practice, 14*(6), 2186-2193.

August, D., Shanahan, T., & Escamilla, K. (2009). English language learners: Developing literacy in second-language learners – report of the national literacy panel on language-minority children and youth. *Journal of Literacy Research, 41*, 432-452. doi:10.1080/10862960903340165

Barnes, B. J., & Wells, C. S. (2009). Differential item functional analysis by gender and race of the national doctoral program survey. *International Journal of Doctoral Studies, 4*, 77-96.

Bastug, O. Y. O. (2016). A comparison of four differential item functioning procedures in the presence of multidimensionality. *Educational Research and Reviews, 11*(13), 1251-1261. Retrieved from https://ezproxy.mtsu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1106185&site=ehost-live&scope=site

Beaver, J. L., French, B. F., Finch, W. H., & Ullrich-French, S. (2014). Sex differential item functioning in the inventory of early development III social-emotional skills. *Journal of Psychoeducational Assessment, 32*(8), 775-780. doi:10.1177/0734282914544924

Benson, K. T., Donnellan, M. B., & Morey, L. C. (2017). Gender-related differential item functioning in DSM-IV/DSM-5-III (alternative model) diagnostic criteria for borderline personality disorder. *Personality Disorders, 8*(1), 87-93. doi:10.1037/per0000166

Binet, A., & Simon, T. (1973). *Classics in psychology: The development of intelligence in children.* New York: Arno Press.

Bott, P. A. (1996). *Testing and assessment in occupational and technical education*. Allyn & Bacon: Needham Height, MA.

Braun, H., & Matthias, V. D. (2017). The use of test scores from large-scale assessment surveys: Psychometric and statistical considerations. *Large-Scale Assessments in Education, 5(1), 1-16*. doi:10.1186/s40536-017-0050-x

Breidenbach, D. H., & French, B. F. (2010). Ordinal logistic regression to detect differential item functioning for gender in the institutional integration scale. *Journal of College Student Retention: Research, Theory & Practice, 12*(3), 339-352. doi:10.2190/CS.12.3.e

Brigance, A. H., & Hargis, C. H. (1993). *Educational assessment: Ensuring that all students succeed in school.* Springfield, Il: Charles C Thomas, Publisher.

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19*(4), 309-321.

Bulut, O., Quo, Q., & Gierl, M. J. (2017). A structural equation modeling approach for examining position effects in large-scale assessments. *Large-Scale Assessments in Education,* 5(1), 1-20. doi:10.1186/s40536-017-0042-x

Camilli, G., & Shepard, L. A. (1994). *Method of identifying biased test items*. Thousand Oaks, California: Sage Publications, Inc.

Camilli, G. (2006). Test fairness. *Educational Measurement, 4*, 221-256.

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational & Behavioral Statistics, 24*(4), 323-341. doi:10.3102/10769986024004323

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Mantel-Haenszel statistics. *Journal of Educational Measurement, 34*(2), 123-139.

Carlton, S. T., & Harris, A. M. (1992). Characteristics associated with differential item functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons. *ETS Research Report Series, 1992*(2), 1-143.

Cauffman, E., & MacIntosh, R. (2006). A Rasch differential item functioning analysis of the Massachusetts youth screening instrument. *Educational & Psychological Measurement, 66*(3), 502-521. Retrieved from https://ezproxy.mtsu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=20948381&site=ehost-live&scope=site

Chae, S. E., Kim, D., & Han, J. (2012). Determinants of differential item functioning in an elementary mathematics test with accommodations. *IEEE Transactions on Education, 55*(2), 279-284. doi:10.1109/TE.2011.2170991

Chan, K., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. *Journal of Applied Psychology, 84*(4), 610.

Cheema, J. R. (2017). Cross-country gender DIF in PISA science literacy items. *European Journal of Developmental Psychology*, 1-16. doi:10.1080/17405629.2017.1358607

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*(2), 133-148.

Cohen, A. S., Kim, S. H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*(1), 15-26.

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

De Beer, M. (2004). Use of differential item functioning (DIF) analysis for bias analysis in test construction. *SA Journal of Industrial Psychology, 30*(4), 52-58.

DeMars, C. (2010). *Item response theory*. New York, NY USA: Oxford University Press.

DeMars, C. E., & Jurich, D. P. (2015). The interaction of ability differences and guessing when modeling differential item functioning with the Rasch model: Conventional and tailored calibration. *Educational & Psychological Measurement, 75*(4), 610-633. doi:10.1177/0013164414554082

Doğan, N., Hambleton, R. K., Yurtcu, M., & Yavuz, S. (2018). The comparison of differential item functioning predicted through experts and statistical techniques. *Cypriot Journal of Educational Sciences*, *13*(2), 375-384. doi:10.18844/cjes.v13i2.2427

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization; In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

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

Dorans, N. J., & Kulick, E. (1983). Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach 1. *ETS Research Report Series, 1983*(1), 14.

Elder, C. (1996). The effect of language background on “foreign” language test performance: The case of Chinese, Italian, and modern Greek. *Language Learning, 46*(2), 233-282.

Elliott, S. N., McKevitt, B. C., & Kettler, R. J. (2002). Testing accommodations research and decision making: The case of" good" scores being highly valued but difficult to achieve for all students. *Measurement and Evaluation in Counseling and Development, 35*(3), 153.

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology, 77*(2), 177.

Elosua, P., & Wells, C. (2013). Detecting DIF in polytomous items using MACS, IRT and ordinal logistic regression. *Psicologica: International Journal of Methodology and Experimental Psychology, 34*(2), 327-342.

Evans, S. S., Evans, W. H., & Mercer, C. D. (1986). *Assessment for instruction*. Boston: Allyn & Bacon.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-381.

Fidalgo, A. M., Tenenbaum, H. R., & Aznar, A. (2018). Are there gender differences in emotion comprehension? analysis of the test of emotion comprehension. *Journal of Child & Family Studies, 27*(4), 1065-1074. doi:10.1007/s10826-017-0956-5

Finch, H., French, B. F., & Immekus, J. C. (2014). *Applied psychometrics using SAS*. USA: Information Age Publishing Inc.

Fuchs, L. S., Hamlett, D. F. C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal*, *28*(3), 617-641.

Furr, M. (2011). *Scale construction and psychometrics for social and personality psychology.* SAGE Publications Ltd.

Gall, M. D., Gall, J. P., & Borg, W. R. (2007). *Educational research*. USA: Pearson Education, Inc.

Grover, R. K., & Ercikan, K. (2017). For which boys and which girls are reading assessment items biased against? detection of differential item functioning in heterogeneous gender populations. *Applied Measurement in Education, 30*(3), 178-195. doi:10.1080/08957347.2017.1316276

Hambleton, K. R., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA, USA: Sage Publications, Inc.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer-Nijhoff Publishing.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*(4), 313-334.

Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education*, *12*(3), 211-235.

Haynes, M. (2011). The federal role in confronting the crisis in adolescent literacy. *The Education Digest, 76*(8), 10.

Hickey, D. T., & Zuiker, S. J. (2005). Engaged participation: A sociocultural model of motivation with implications for educational assessment. *Educational Assessment, 10*(3), 277-305. doi:10.1207/s15326977ea1003_7

Hidalgo-Montesinos, M. D., & Gomez-Benito, J. (2003). Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment*, *19*(1), 1.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test Validity*, 129-145.

Hope, D., Adamson, K., McManus, I. C., Chis, L., & Elder, A. (2018). Using differential item functioning to evaluate potential bias in a high-stakes postgraduate knowledge-based assessment. *BMC Medical Education*, *18*(1), 64. doi:10.1186/s12909-018-1143-0

Immekus, J. C., & McGee, D. (2016). The measurement invariance of the student opinion survey across English and non-English language learner students within the context of low- and high-stakes assessments. *Frontiers in Psychology,7*. doi:10.3389/fpsyg.2016.01352/full; 10.3389/fpsyg.2016.01352

Innabi, H., & Dodeen, H. (2006). Content analysis of gender-related differential item functioning TIMSS items in mathematics in Jordan. *School Science & Mathematics, 106*(8), 328-337. doi:10.1111/j.1949-8594.2006.tb17753.x

Ismail, G., & Koch, E. (2012). Investigating item and construct bias in an English verbal analogies scale. *Southern African Linguistics and Applied Language Studies, 30*(3), 325-338.

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

Johnson, D. D., & Johnson, B. (2002). *High stakes: Children, testing, and failure in American schools*. Lanham, ML USA: Rowman and Littlefield Publishers, Inc.

Jones, M. G., Jones, B. D., & Hargrove, T. Y. (2003). *The unintended consequences of the high-stakes testing*. Lanham, ML USA: Rowman and Littlefield publishers, Inc.

Kan, A., & Bulut, O. (2014). Examining the relationship between gender DIF and language complexity in mathematics assessments. *International Journal of Testing*, *14*(3), 245-264. doi:10.1080/15305058.2013.87911

Kanno, Y., & Kangas, S. E. (2014). "I'm not going to be, like, for the AP" English language learners' limited access to advanced college-preparatory courses in high school. *American Educational Research Journal*, *51*(5), 848-878.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling, 18*(2), 212-228.

Kim, J., & Oshima, T. C. (2013). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement, 73*(3), 458-470.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, *18*(1), 89-114. Retrieved from https://ezproxy.mtsu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=mzh&AN=2001651469&site=ehost-live&scope=site

Kim, S., Cohen, A. S., & Park, T. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32*(3), 261-276.

Koo, J., Becker, B. J., & Kim, Y. (2014). Examining differential item functioning trends for English language learners in a reading test: A meta-analytical approach. *Language Testing, 31*(1), 89-109. doi:10.1177/0265532213496097

Kunnan, A. J., & Weinstein-Shr, G. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly, 24*(4), 741-746.

Kurt, M., Karakaya, I., Safaz, I., & Ates, G. (2015). Differential item functioning by education and sex in subtests of the repeatable battery assessment of neuropsychological status. *European Journal of Psychological Assessment, 31*(1), 11. doi:10.1027/1015-5759/a000198

Lambert, M. C., Garcia, A. G., Epstein, M. H., & Cullinan, D. (2018). Differential item functioning of the emotional and behavioral screener for Caucasian and African American elementary school students. *Journal of Applied School Psychology, 34*(3), 201-214. doi:10.1080/15377903.2017.1345815

Lambert, M. C., Garcia, A. G., January, S. A., & Epstein, M. H. (2017). The impact of English language learner status on screening for emotional and behavioral disorders: A differential item functioning (DIF) study. *Psychology in the Schools, 55*(3), 229-239. Retrieved from https://ezproxy.mtsu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1169669&site=ehost-live&scope=site http://dx.doi.org/10.1002/pits.22103

Lambert, M. C., January, S. A., Cress, C. J., Epstein, M. H., & Cullinan, D. (2017). Differential item functioning across race and ethnicity for the emotional and behavioral screener. *School Psychology Quarterly,* doi:10.1037/spq0000224

Le, L. T. (2006). Analysis of differential item functioning. In *annual meeting of American Educational Research Association, San Fracisco CA* (Vol. 8).

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. New Jersey, USA: Lawrence Erlbaum Associates, Inc.

Magis, D., & Facon, B. (2014). Delta plot R: An R package for differential item functioning analysis with Angoff's Delta plot. *Journal of Statistical Software*, *59*(1), 1-19.

Maller, S. J., French, B. F., & Zumbo, B. D. (2011). Item and test bias. In N. J. Salkind (Eds.), *Encyclopedia of measurement and statistics* (pp. 490-493). Retrieved from http://dx.doi.org/10.4135/9781412952644

Marzano, R. (2018). *Making classroom assessments reliable and valid*. Bloomington, IN USA: Solution Tree Press.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49*(4), 529-544.

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement, 11*(2), 161-173.

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30*(2), 107-122.

Millsap, R. E., Gunn, H., Everson, H. T., & Zautra, A. (2015). Using item response theory to evaluate measurement invariance in health-related measures. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. (pp. 364). New York & London: Routledge.

Morrow, L. M., Tracey, D. H., & Del Nero, J. R. (2011). Best practices in early literacy: Preschool, kindergarten, and first grade. *Best Practices in Literacy Instruction, 4*, 67-95.

Murray, A. L., Booth, T., & McKenzie, K. (2015). An analysis of differential item functioning by gender in the learning disability screening questionnaire (LDSQ). *Research in Developmental Disabilities, 39*, 76-82. doi:10.1016/j.ridd.2014.12.006

National Center for Education Statistics. (2017). *National Assessment of Educational Progress: An overview of NAEP*. Washington, D.C.: National Center for Education Statistics, Institute of Education Sciences, U.S. Dept. of Education.

National Center for Education Statistics. (2009). *National Assessment of Educational Progress: An overview of NAEP*. Washington, D.C.: National Center for Education Statistics, Institute of Education Sciences, U.S. Dept. of Education.

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common Core State Standards for mathematics: Kindergarten introduction.* Retrieved from http://www.corestandards.org/ELA-Literacy/

No Child Left Behind Act of 2001, P.L. 107-110, 20 U.S.C. § 6319 (2002).

Oliveri, M. E., Lawless, R., Robin, F., & Bridgeman, B. (2018). An exploratory analysis of differential item functioning and its possible sources in a higher education admissions context. *Applied Measurement in Education, 31*(1), 1-16. Retrieved from https://ezproxy.mtsu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1165314&site=ehost-live&scope=site http://dx.doi.org/10.1080/08957347.2017.1391258

Ong, Y. M., Williams, J., & Lamprianou, I. (2015). Exploring crossing differential item functioning by gender in mathematics assessment. *International Journal of Testing, 15*(4), 337-355. Retrieved from https://ezproxy.mtsu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1078290&site=ehost-live&scope=site http://dx.doi.org/10.1080/15305058.2015.1057639

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning.* Sage Publications.

Phillips, M. M. (1993). The absolute magnitudes of type IA supernovae. *The Astrophysical Journal, 413,* L108.

Phillips, V., & Wong, C. (2010). Tying together the common core of standards, instruction, and assessments. *Phi Delta Kappan*, *91*(5), 37-42.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common core standards: The new US intended curriculum. *Educational Researcher, 40*(3), 103-116.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*(4), 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197-207.

Rasch, G. (1960). Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests. Oxford, England: Nielsen & Lydiche.

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zarate, R. C. (2018). The relationship between test item format and gender achievement gaps on math and ELA tests in fourth and eighth grades. *Educational Research*, *47*(5). 284-294. doi:10.3102/0013189X18762105

Reynolds, C. R., & Livingston, R. B. (2012). *Mastering modern psychological testing: Theory and methods.* New Jersey: Pearson Education.

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

Ryan, K. E., & Bachman, L. F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing*, *9*(1), 12-29.

Salvia, J., Ysseldyke, J., & Witmer, S. (2012). *Assessment: In special and inclusive education.* USA: Cengage Learning.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4, Pt. 2), 100.

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16*(3), 143-152.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detect test bias/DIF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194.

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. 93-115. Mahwah, NJ: Lawrence Erlbaum Associates.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292.

Stark, S., Chernyshenko, O. S., Chan, K., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology, 86*(5), 943.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychological methods*, *11*(4), 402.

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

Tanner, D. E. (2001). *Assessing academic achievement.* Needham Heights., MA: Allyn and Bacon.

Taylor, G. S. (2004). *The impact of high-stakes testing on the academic futures of non-mainstream students*. Lewiston, New York: The Edwin Mellen Press.

Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine, 19*(11-12), 1651-1683.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104*(3), 385.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Thurlow, M. L., Ysseldyke, J. E., & Silverstein, B. (1995). Testing accommodations for students with disabilities. *Remedial and Special Education, 16*(5), 260-270.

U. S. department of education. (2018). *National center for education statistics.*

van de Vijver, Fons J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39-63). Mahwah, NJ: Lawrence Erlbaum Associates.

San Martín, E. (2016). Identification of item response theory models. In van der

Linden, W. J. (Eds.), *Handbook of Item Response Theory: Models, Statistical Tools, and Applications* (pp. 127-170). Vol. 2. Boca Raton, FL: Taylor & Francis Groups.

Walsh, B. W., & Betz, N. E. (1985). *Tests and assessment*. Englewood Cliffd, NJ: Prentice-Hall, Inc.

Wang, N., & Lane, S. (1996). Detection of gender-related differential item functioning in a mathematics performance assessment. *Applied Measurement in Education, 9*(2), 175-199.

Wedman, J. (2017). Reasons for gender-related differential item functioning in a college admissions test. *Scandinavian Journal of Educational Research, 12*. 1-12. doi:10.1080/00311831.2017.1402365

Wei, T., Chesnut, S. R., Barnard-Brak, L., Stevens, T., & Olivárez, A., Jr. (2014). Evaluating the mathematics interest inventory using item response theory: Differential item functioning across gender and ethnicities. *Journal of Psychoeducational Assessment, 32*(8), 747-761. Retrieved from https://ezproxy.mtsu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1045292&site=ehost-live&scope=site http://dx.doi.org/10.1177/0734282914540449

Wetzel, E., & Hell, B. (2013). Gender-related differential item functioning in vocational interest measurement: An analysis of the AIST-R. *Journal of Individual Differences, 34*(3), 170-183. doi:10.1027/1614-0001/a000112

Wiesner, M., Windle, M., Kanouse, D. E., Elliott, M. N., & Schuster, M. A. (2015). DISC predictive scales (DPS): Factor structure and uniform differential item functioning across gender and three racial/ethnic groups for ADHD, conduct disorder, and oppositional defiant disorder symptoms. *Psychological Assessment, 27*(4), 1324.

Wixson, K. K., & Carlisle, J. F. (2005). The influence of large-scale assessment of reading comprehension on classroom practice. In S. G. Paris & S. A. Stahl (Eds.), *Children's Reading Comprehension and Assessment*. (pp. 395-405). Mahwah, NJ: Lawrence Erlbaum Associates.

Woods, C. M., & Harpole, J. (2015). How item residual heterogeneity affects tests for differential item functioning. *Applied Psychological Measurement, 39*(4), 251-263. doi:10.1177/0146621614561313

Wyse, A. E., & Mapuranga, R. (2009). Differential item functioning analysis using Rasch item information functions. *International Journal of Testing, 9*(4), 333-357. doi:10.1080/15305050903352040

Yalcin, S. (2018). Determining differential item functioning with the mixture item response theory. *Egitim Arastirmalari - Eurasian Journal of Educational Research,* (74), 187-206. doi:10.14689/ejer.2018.74.10

Yau, D. T. W., Wong, M. C. M., Lam, K. F., & McGrath, C. (2015). Evaluation of psychometric properties and differential item functioning of 8-item child perceptions questionnaires using item response theory. *BMC Public Health, 15*, 792. doi:10.1186/s12889-015-2133-3

Young, E. L., & Sudweeks, R. R. (2005). Gender differential item functioning in the multidimensional self-concept scale with a sample of early adolescent students. *Measurement and Evaluation in Counseling and Development, 38*(1), 29-44.

Zumbo, B. D. (2008, July). Statistical methods for investigating item bias in self-report measures. *Florence Lectures on DIF and Item Bias*. Lectures Conducted from Universita degli Studi di Firenze, Florence, Italy.

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 (ordinal) item scores*. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233. doi:10.1080/15434300701375832

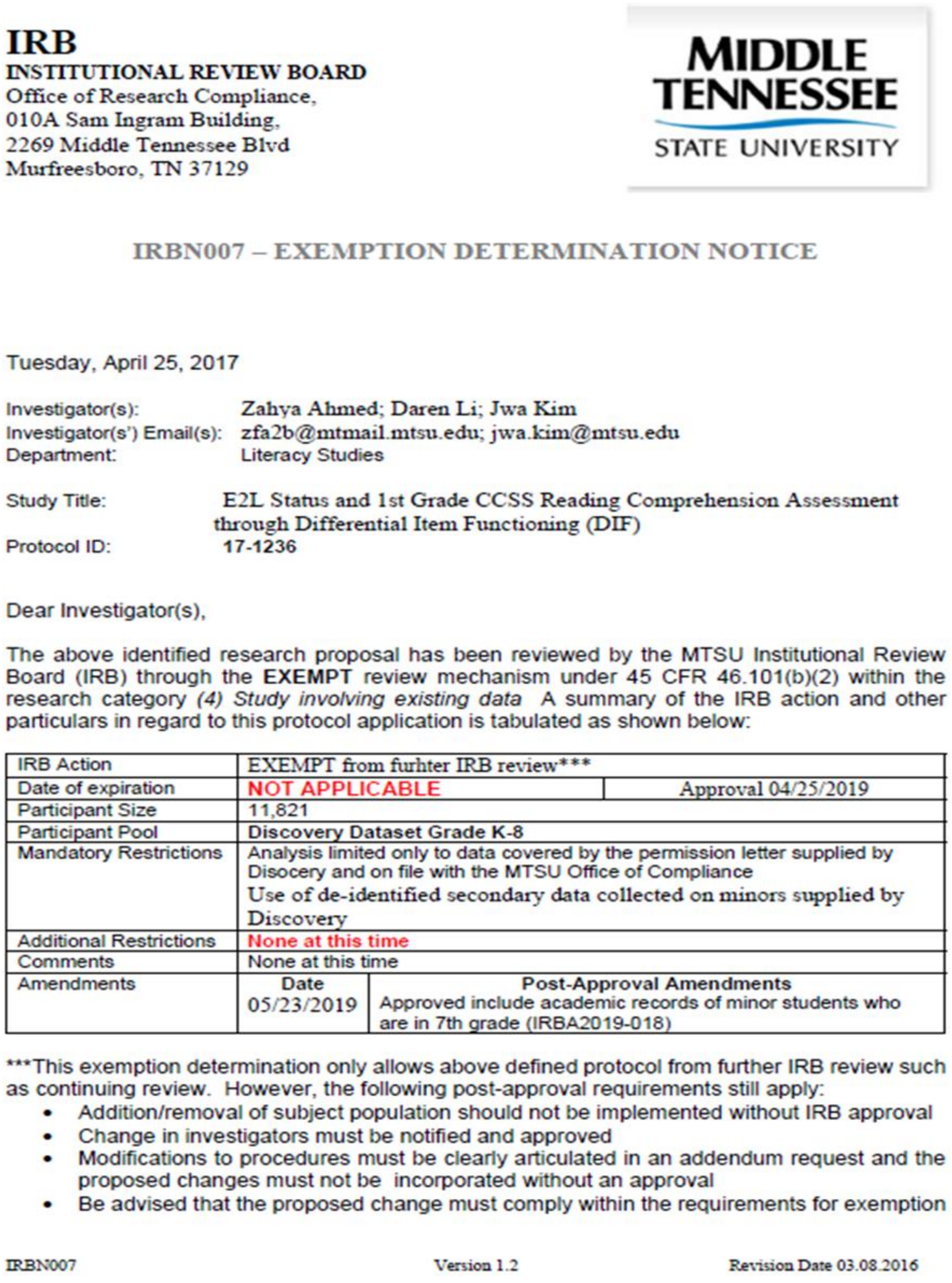
Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational. 26*(1). 55-66.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series, 2012*(1), 1-30.

APPENDICES

APPENDIX A

IRB APPROVAL



IRB
INSTITUTIONAL REVIEW BOARD
Office of Research Compliance,
010A Sam Ingram Building,
2269 Middle Tennessee Blvd
Murfreesboro, TN 37129

MIDDLE TENNESSEE
STATE UNIVERSITY

IRBN007 – EXEMPTION DETERMINATION NOTICE

Tuesday, April 25, 2017

Investigator(s): Zahya Ahmed; Daren Li; Jwa Kim
Investigator(s') Email(s): zfa2b@mtmail.mtsu.edu; jwa.kim@mtsu.edu
Department: Literacy Studies

Study Title: E2L Status and 1st Grade CCSS Reading Comprehension Assessment through Differential Item Functioning (DIF)
Protocol ID: **17-1236**

Dear Investigator(s),

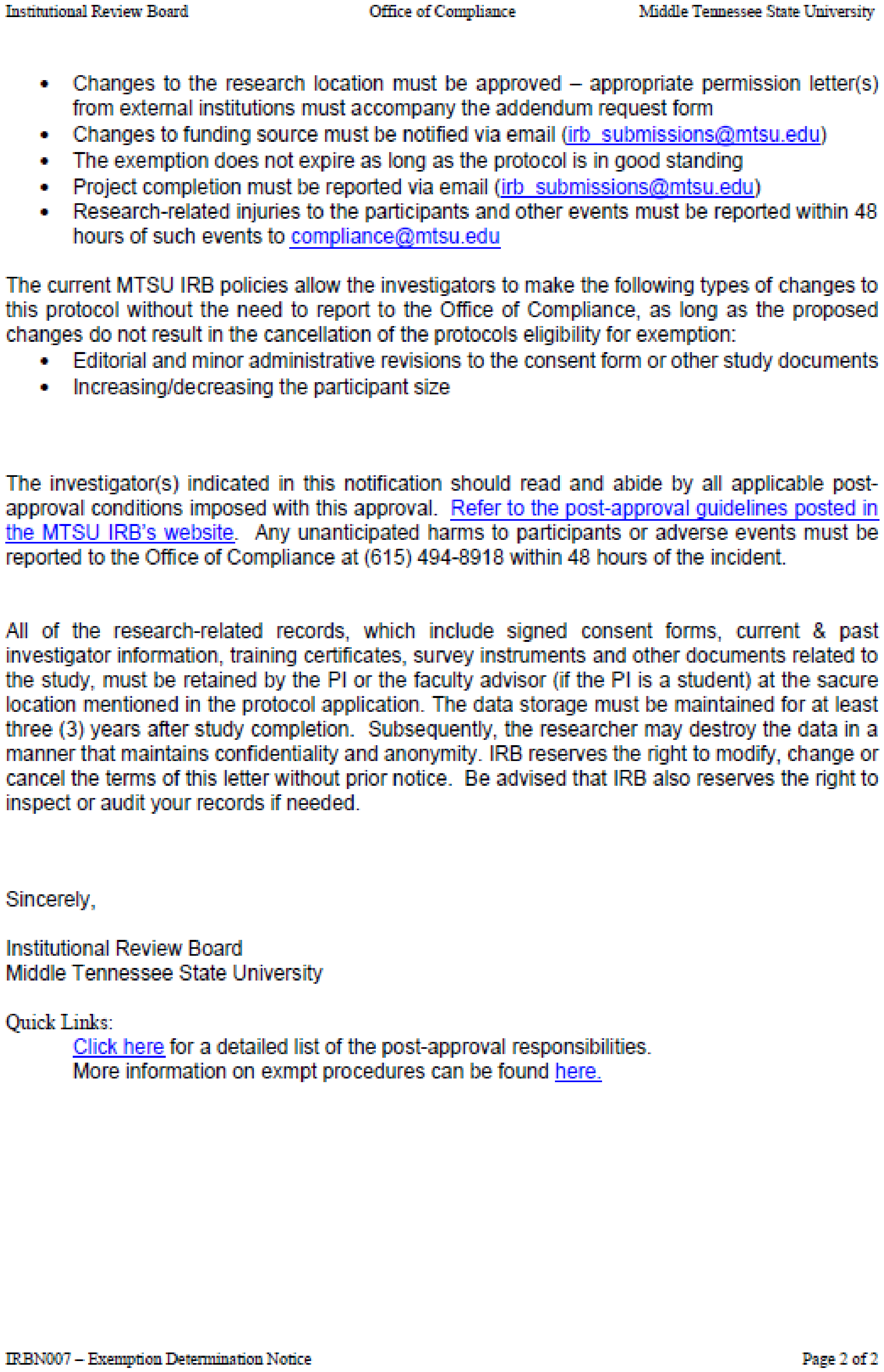
The above identified research proposal has been reviewed by the MTSU Institutional Review Board (IRB) through the **EXEMPT** review mechanism under 45 CFR 46.101(b)(2) within the research category *(4) Study involving existing data* A summary of the IRB action and other particulars in regard to this protocol application is tabulated as shown below:

IRB Action	EXEMPT from furhter IRB review***	
Date of expiration	NOT APPLICABLE	Approval 04/25/2019
Participant Size	11,821	
Participant Pool	**Discovery Dataset Grade K-8**	
Mandatory Restrictions	Analysis limited only to data covered by the permission letter supplied by Disocery and on file with the MTSU Office of Compliance Use of de-identified secondary data collected on minors supplied by Discovery	
Additional Restrictions	None at this time	
Comments	None at this time	
Amendments	**Date** 05/23/2019	**Post-Approval Amendments** Approved include academic records of minor students who are in 7th grade (IRBA2019-018)

***This exemption determination only allows above defined protocol from further IRB review such as continuing review. However, the following post-approval requirements still apply:

- Addition/removal of subject population should not be implemented without IRB approval
- Change in investigators must be notified and approved
- Modifications to procedures must be clearly articulated in an addendum request and the proposed changes must not be incorporated without an approval
- Be advised that the proposed change must comply within the requirements for exemption

IRBN007 Version 1.2 Revision Date 03.08.2016



- Changes to the research location must be approved – appropriate permission letter(s) from external institutions must accompany the addendum request form
- Changes to funding source must be notified via email (irb_submissions@mtsu.edu)
- The exemption does not expire as long as the protocol is in good standing
- Project completion must be reported via email (irb_submissions@mtsu.edu)
- Research-related injuries to the participants and other events must be reported within 48 hours of such events to compliance@mtsu.edu

The current MTSU IRB policies allow the investigators to make the following types of changes to this protocol without the need to report to the Office of Compliance, as long as the proposed changes do not result in the cancellation of the protocols eligibility for exemption:

- Editorial and minor administrative revisions to the consent form or other study documents
- Increasing/decreasing the participant size

The investigator(s) indicated in this notification should read and abide by all applicable post-approval conditions imposed with this approval. Refer to the post-approval guidelines posted in the MTSU IRB's website. Any unanticipated harms to participants or adverse events must be reported to the Office of Compliance at (615) 494-8918 within 48 hours of the incident.

All of the research-related records, which include signed consent forms, current & past investigator information, training certificates, survey instruments and other documents related to the study, must be retained by the PI or the faculty advisor (if the PI is a student) at the sacure location mentioned in the protocol application. The data storage must be maintained for at least three (3) years after study completion. Subsequently, the researcher may destroy the data in a manner that maintains confidentiality and anonymity. IRB reserves the right to modify, change or cancel the terms of this letter without prior notice. Be advised that IRB also reserves the right to inspect or audit your records if needed.

Sincerely,

Institutional Review Board
Middle Tennessee State University

Quick Links:

Click here for a detailed list of the post-approval responsibilities.
More information on exmpt procedures can be found here.