DIFFERENTIAL ITEM FUNCTIONING AMONG ENGLISH LANGUAGE LEARNERS ON A LARGE-SCALE MATHEMATICS ASSESSMENT

Ruixue Liu
University of Kentucky, liuruixue2046@hotmail.com
Digital Object Identifier: https://doi.org/10.13023/etd.2019.097
STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student’s advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student’s thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Ruixue Liu, Student

Dr. Kelly D. Bradley, Major Professor

Dr. Margaret Bausch, Director of Graduate Studies
DIFFERENTIAL ITEM FUNCTIONING AMONG ENGLISH LANGUAGE LEARNERS ON A LARGE-SCALE MATHEMATICS ASSESSMENT

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the College of Education at the University of Kentucky

By
Ruixue Liu
Lexington, Kentucky

Director: Dr. Kelly D. Bradley, Professor of Educational Policy Studies & Evaluation
Lexington, Kentucky
2019

Copyright © Ruixue Liu 2019
ABSTRACT OF DISSERTATION

DIFFERENTIAL ITEM FUNCTIONING AMONG ENGLISH LANGUAGE LEARNERS ON A LARGE-SCALE MATHEMATICS ASSESSMENT

English language learner (ELL) is a term to describe students who are still acquiring English proficiency. In recent decades, ELLs are a very rapidly growing student group in United States. In school classrooms, ELLs are learning English and their academic subjects simultaneously. It is challenging for them to hear lectures, read textbooks, and complete tests in English despite of their inadequate English language proficiency (Ilich, 2013). As a result, the increasing number of ELLs in public schools has paralleled the increase in ELLs’ low mathematics performance (NCES, 2016).

Due to the popularization of international large-scale assessments in the recent decade, it is necessary to analyze their psychometric properties (e.g., reliability, validity) so that those results can provide with evidence-based implications for policymakers. Educational researchers need to assess the validity for subgroups within each country. The Programme for International Student Assessment (PISA), as one of the influential large-scale assessments, allows researchers to investigate academic achievement and group membership from a variety of different viewpoints.

The current study was to understand the nature and potential sources of the gaps in mathematics achievement between ELLs and non-ELLs. The nature of achievement gap was examined using three DIF methodologies including Mantel-Haenszel procedure, Rasch analysis, and Hierarchical Generalized Linear Model (HGLM) at the item level instead of total test level. Among the three methods, HGLM was utilized to examine the potential sources of DIF. This method can take into account of the nested structure of data where items are nested within students, and students nested within schools. At the student level, sources of DIF were investigated through students’ variations in mathematics self-efficacy, language proficiency, and student socioeconomic status. At the school level, school type and school educational resource were investigated as potential sources of DIF after controlling for the student variables. The U.S. sample from PISA 2012 was used, and 76 dichotomously coded items from PISA 2012 mathematics assessment were included to detect DIF effects.
Results revealed that ten common items are identified with DIF effects using MH procedure, Rasch analysis, and HGLM. These ten items are all in favor of non-ELs. The decreasing number of items showing DIF effects in HGLM after controlling for student-level variables revealed mathematics self-efficacy, language proficiency, and SES are potential sources of DIF between ELLs and non-ELs. In addition, the number of DIF items continued to decrease after controlling for both student and school-level variables. This finding proved that school type and school educational resources were also potential sources of DIF between ELLs and non-ELs.

Findings from this study can help educational researchers, administrators, and policymakers understand the nature of the gap at item level instead of the total test level so that United States can be competitive in middle school mathematics education. This study can also help guide item writers and test developers in the construction of more linguistically accessible assessments for students who are still learning English. The significance of this study lies in the empirical investigation of the gap between ELLs and non-ELs in mathematics achievement at an item level and from perspectives of both students and schools.

KEYWORDS: English Language Learners, Differential Item Functioning, PISA 2012, Mathematics Assessment

Ruixue Liu

04/16/2019
Date
DIFFERENTIAL ITEM FUNCTIONING AMONG ENGLISH LANGUAGE LEARNERS ON A LARGE-SCALE MATHEMATICS ASSESSMENT

By
Ruixue Liu

Kelly D. Bradley Ph.D.
Director of Dissertation

Margaret Bausch Ph.D.
Director of Graduate Studies

04/16/2019
Date
ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Dr. Kelly Bradley, my advisor, for her mentorship and guidance throughout my graduate studies, both intellectually and professionally. Her patience and contribution are critical in this endeavor. She is always supportive and very caring of students. Without her, the completion of this work would not be possible.

I am very grateful to my committee members, Dr. Cindy Jong, Dr. Willis Jones, and Dr. John Thelin for their insightful and constructive comments at every stage of the dissertation process, allowing me to complete this project in a timely manner. I appreciate the time taken by Dr. John Nash to attend my defense and give me invaluable suggestions to my draft.

I am very thankful to all my friends I met at EPE. Dr. Patrick Yang advised me when I entered the master’s program. Every conversation I had with him was very beneficial to me. Dr. Renee Setari and Dr. Anthony Setari gave many helpful suggestions on preparing for my qualifying exam and dissertation. Dr. Michael Peabody, as my internship mentor, provided with great professional support for my time at the American Board of Family Medicine.

I would also thank to my dear parents who always supported my schooling throughout my life. I also appreciate my sister for taking care of the whole family when I am far away from home.

Finally, my deepest gratitude goes to my husband and daughter, Meng and Qile. To Meng, I cannot thank you enough nor express with words how grateful I truly am. You made my life a wonderful adventure. I cannot wait for the next chapter of our lives to begin.
# TABLE OF CONTENTS

**ACKNOWLEDGMENTS** ........................................................................................................ iii

**LIST OF TABLES** ........................................................................................................... vi

**LIST OF FIGURES** ......................................................................................................... vii

**CHAPTER 1. INTRODUCTION** ......................................................................................... 1
  - Statement of Problem ..................................................................................................... 1
  - Purpose of Study .......................................................................................................... 2
  - Research Questions ..................................................................................................... 4

**CHAPTER 2. LITERATURE REVIEW** ............................................................................... 6
  - Cognitive Load Theory ................................................................................................ 6
    - Overview .................................................................................................................. 6
    - CLT in Educational Measurement ......................................................................... 8
  - ELL Students’ Performance on Mathematics ............................................................. 9
  - Factors Influencing ELL Students’ Mathematics Achievement ................................... 12
    - Primary Culture ...................................................................................................... 12
    - Parental Involvement ............................................................................................... 15
    - English Language Proficiency ............................................................................... 17
    - School Characteristics ............................................................................................ 21
  - Assessment Validity for ELLs ..................................................................................... 22
  - Differential Item Functioning ..................................................................................... 23
    - Description of DIF .................................................................................................. 23
    - DIF Detection Methods .......................................................................................... 25

**CHAPTER 3. METHOD** .................................................................................................. 33
  - Data Source ................................................................................................................ 33
    - Overview of PISA 2012 ........................................................................................ 33
    - PISA 2012 Assessment Design ............................................................................. 34
  - Sample ......................................................................................................................... 35
  - Measurement ............................................................................................................... 36
    - Mathematics Assessment ....................................................................................... 36
Assessment Items ........................................................................................................38
Student-Level Measures ..............................................................................................42
School-Level Measures ...............................................................................................44
Statistical Analyses .......................................................................................................46
Mantel-Haenszel Procedure ........................................................................................46
Rasch Model ..................................................................................................................47
Hierarchical Generalized Linear Model .....................................................................48
CHAPTER 4. RESULTS .................................................................................................54
Mantel-Haenszel Procedure .........................................................................................54
Rasch Model ..................................................................................................................55
Model Fit Statistics .......................................................................................................56
Item Fit Statistics ..........................................................................................................57
Person and Item Distributions ......................................................................................60
DIF between ELLs and Non-ELLs ................................................................................61
Hierarchical Generalized Linear Model .....................................................................63
Model 1 ..........................................................................................................................63
Model 2 ..........................................................................................................................64
Model 3 ..........................................................................................................................66
Consistency of Three DIF Detection Methods .............................................................68
CHAPTER 4. DISCUSSIONS .........................................................................................70
Summary of Findings .....................................................................................................71
Conclusions ....................................................................................................................75
Implications ....................................................................................................................77
Implications for Teachers and Educators .....................................................................77
Implications for Test Developers ................................................................................79
Implications for Policymakers ......................................................................................80
Limitations and Future Research ...............................................................................81
Reference .......................................................................................................................83
VITA .................................................................................................................................98
LIST OF TABLES

Table 1. 2 x 2 x K Contingency Table .................................................................26
Table 2. Descriptions of Mathematics Assessment Items in PISA 2012 .....................39
Table 3. Items of School Educational Resource Assessment .....................................46
Table 4. Summary Results from the MH Procedure to identify DIF Effects ................55
Table 5. Model Data Fit Statistics ........................................................................57
Table 6. Item Measure and Fit Statistics ..................................................................58
Table 7. Summary of Results from Rasch Analysis to Identify DIF Effects ................63
Table 8. Summary of HGLM Model 1 ....................................................................63
Table 9. Descriptive Statistics of Student-Level Variables ......................................64
Table 10. Summary of Results from HGLM Model 2 ..............................................65
Table 11. Descriptive Statistics of School-Level Variables ......................................66
Table 12. Summary of Results from HGLM Model 3 ..............................................67
Table 13. Summary of DIF Items Identified by Three Methods ..............................68
LIST OF FIGURES

Figure 1. Wright Map ........................................................................................................... 62
CHAPTER 1. INTRODUCTION

Statement of the Problem

English language learner (ELL) is a term to describe students who are still acquiring English proficiency. According to U.S. Department of Education, ELLs are defined as students “who are being served in appropriate programs of language assistance” (National Center for Education Statistics [NCES], 2016). In recent decades, ELLs are a rapidly growing student group in United States. The percentage of public school students in the United States identified as ELLs was higher in Fall 2014–15 (9.5%, or an estimated 4.8 million students) than in Fall 2000 (8.1%, or an estimated 3.8 million students) and Fall 2013 (9.2%, or an estimated 4.2 million students). In Fall 2015, the percentage of public school students who were ELLs ranged from 1.0% in West Virginia to 21.0% in California (NCES, 2018).

In school classrooms, ELLs are learning English and their academic subjects simultaneously. It is challenging for them to hear lectures, read textbooks, and complete tests in English due to their inadequate English language proficiency (Ilich, 2013). Based on the Cognitive Load Theory (CLT), bilingual learners or ELLs must face additional overall cognitive demands during problem solving while working in a non-primary language for them (Campbell, Adams and Davis, 2007). To this end, ELLs have been found to lag behind their non-ELL peers on large-scale, standardized assessments, particularly content areas that are high in language demand, such as mathematics, science, reading comprehension, writing, and social studies (Abedi, 2002; Abedi et al., 2005; Abedi, Hofstetter, Baker, & Lord, 2001; Abedi & Lord, 2001; Johnson & Monroe, 2004;
Mahoney, 2008; Martiniello, 2009; Ockey, 2007; Walker, Zhang, & Surber, 2008; Wolf & Leon, 2009).

The increasing number of ELLs in public schools has paralleled the increase in ELLs low mathematics performance (NCES, 2016). For instance, ELLs are among the lowest scoring groups in the National Assessment of Educational Progress (NAEP) mathematics assessment. Almost half of ELLs scored below Basic in the NAEP fourth-grade mathematics tests in 2005, 2007, and 2009 (46%, 44%, and 43%, respectively). By comparison, 18% of non-ELLs in 2005 and 16% in 2007 and 2009 scored below Basic in the same grade level (Martiniello, 2009). As a result of the performance gap, educational researchers have been concerned about the appropriateness of these assessments for students who are not yet proficient in English.

**Purpose of Study**

Due to the popularization of international large-scale assessments in the recent decade, it is necessary to analyze their psychometric properties (e.g., reliability, validity) so those results can provide evidence-based implications for policymakers. Educational researchers need to assess the validity for subgroups within each country. The Programme for International Student Assessment (PISA), one of the influential large-scale assessments, allows researchers to investigate academic achievement and group membership from a variety of different viewpoints (Organisation for Economic Co-Operation and Development [OECD], 2014).

When investigating the achievement gap between ELLs and non-ELLs on mathematics achievement, most of the existing studies relied on statistics such as means, variance, and effect sizes (e.g., Abedi, 2002, Beal, Adams, & Cohen, 2010; Fry, 2007).
These studies did not identify whether items in assessments can cause the gaps on the overall measures. To this end, item level analysis of mathematics assessments is supposed to be added to fill out the literature gap.

Differential item functioning (DIF) is a statistical approach to identify whether items on an assessment are of equal difficulty for examinees of distinct groups. In the past decade, some studies have been conducted to detect DIF in PISA assessment items. However, most of the existing studies explored possible DIF sources including gender (e.g., Qian, 2011; Huang, 2010; Le, 2009) and translation equivalence (e.g., Grisay, de Jong, Gebhardt, Berezner & Halleux-Monseur, 2007). Only one study was concerned with DIF on science assessment items between ELLs and non-ELLs (e.g., Shirley, 2014).

Above all, the current study was to understand the nature and potential sources of the gaps in mathematics achievement between ELLs and non-ELLs. The nature of achievement gap was examined using three DIF detection methods including Mantel-Haenszel procedure, Rasch analysis, and Hierarchical Generalized Linear Model (HGLM) at the item level instead of total test level. Among the three methods, HGLM was utilized to examine the potential sources of DIF. This method takes into account the nested structure of data where items are nested within students and students are nested within schools. At the student level, sources of DIF were investigated through students’ variations in mathematics self-efficacy, language proficiency, and student socioeconomic status (SES). At the school level, school type and school educational resource were investigated as potential sources of DIF after controlling for the student variables.

The U.S. sample of PISA 2012 was used, and 76 dichotomously coded items from PISA 2012 mathematics assessment were included to detect DIF effects. The U.S. sample
of PISA 2012 was selected to conduct the current study for two reasons. First, PISA is a large-scale assessment. The U.S. sample contains 4,978 students from 162 schools. Therefore, large sample sizes for both ELLs and non-ELLs can be provided. Second, in addition to assessment data, PISA not only consists of data from students on their family background and attitudes towards mathematics learning, but also includes data from school principals on the quality of school.

**Research Questions**

Specifically, the present study will mainly address three Research Questions.

1) Do items from PISA 2012 mathematics assessment exhibit DIF between ELLs and non-ELLs for the U.S. sample?

2) If DIF is detected, can English language proficiency and other student characteristics (e.g., student SES, mathematics self-efficacy) explain DIF? That is, after controlling for these three student variables, whether DIF between ELLs and non-ELLs changes was examined.

3) If DIF is detected, can school type and school educational resources contribute to DIF?

The first research question intends to find out the reasons behind the gap between ELLs and non-ELLs. The second and third research questions, incorporating a multilevel item analysis method, aim to identify the problem from multiple perspectives. Findings from this study can help educational researchers, administrators, and policymakers understand the nature of the gap at item level instead of the total test level so the United States can be competitive in middle school mathematics education. This study can help guide item writers and test developers in constructing more linguistically accessible
assessments for students who are still learning English. The significance of this study lies in the empirical investigation of the gap between ELLs and non-ELLs in mathematics achievement at an item level and from perspectives of both students and schools.
CHAPTER 2. LITERATURE REVIEW

As mentioned in Chapter 1, ELLs do not perform as well as non-ELLs on content assessments. However, it is less clear for the reasons for this differential performance. Meanwhile, educational researchers have been concerned about whether ELL students’ test scores can accurately reflect their school subject knowledge (Aguirre-Muñoz & Baker, 1999). This chapter begins with reviewing the Cognitive Load Theory (CLT) and its application in educational measurement. The review of ELL students’ performance on mathematics follows with the factors influencing ELL students’ mathematics achievement. Finally, this chapter introduces assessment validity of ELLs and several mainstream DIF detection methodologies.

Cognitive Load Theory

Overview

CLT, proposed by Swella (1988), deals with how psychological constructs are related to learning. Specifically, this theory mainly focuses on how cognitive construct is organized, what happens during the learning process, and how educators develop instructional materials to facilitate learning (Moreo & Park, 2010). It suggested that learning happens best when it is aligned with human cognitive architecture (Sweller, Van Merrienboer, & Pass, 1998).

CLT also suggested that people are equipped with a cognitive structure consisting of working memory and long-term memory (Sweller et al., 1998). When new information is acquired from people’s senses (e.g., visual, auditory), they are processed into people’s working memory (Thorne, 2005). In addition, working memory is regarded as conscious memory with limited capacity when it is used to hold information. Peoples’ performance
on complicated cognitive tasks relies on whether the amount of information presented to user equals or exceeds the availability of working memory. The probability of errors will increase as working memory capacity is exceeded (Kalyuga, Ayres, Chandler & Sweller, 2003; Paas, Renkl & Sweller, 2003).

In terms of long-term memory, people who keep practicing new information and become proficient with a particular topic can hold the information unconsciously for a very long time, which can then be retrieved automatically when dealing with a similar kind of task (Sweller, 1994). For instance, non-ELL students can converse in English more fluently and accurately than ELL students since this proficient use of English subconsciously brings their unconscious knowledge to deal with a new task. Non-ELL students are more easily to retrieve background information from their long-term memory to learn new knowledge (Sweller, 2010).

In mathematics learning, for example, if students must devote significant cognitive resources to text comprehension, fewer working memory resources will become available for mathematics problem solving, including identifying the appropriate mathematics operation formula, organizing the problem representation, conducting computations and checking progress towards the solution (Barbu, 2010). Meanwhile, the language of mathematics has been viewed as a unified system of meaning-making that incorporates the multiple semiotic (Martiniello, 2009). The need to allocate cognitive resources to comprehend a problem presented in a non-primary language would reduce the resources available for problem solving process and result in increasing the probability of errors (Barbu, 2010; Mestre, 1988).
CLT in Educational Measurement

Although CLT was developed for instructional purpose, some researchers have transferred its insights to educational measurement to measure the target construct more accurately. For example, Kettler et al. (2011) investigated whether tests consisting of modified items would have same level of reliability and whether modified items can reduce the achievement gap of students with disabilities. Three groups of eighth-grade students took original and modified version of reading and mathematics tests based on their disability status. Results showed that changes in reliability across groups for both reading and mathematics tests were minimal. The Rasch analysis revealed that mean item difficulties decreased more for students with disabilities. Meanwhile, findings suggested that shortening the question may be a highly effective modification to reduce the cognitive load.

Similarly, Gillmor, Poggio and Embretson (2015) modified 15 multiple-choice mathematics assessment items using research-based strategies to reduce cognitive load to test its effects on student mathematics performance. This experimental study revealed that three load-reducing item modifications are identified as particularly effective for reducing item difficulty, including signaling critical information, aesthetic item organization, and removing extraneous content.

Educational Testing Service (2009) provided guidelines for ELL students. The use of clear and accessible language is important to minimize construct-irrelevant variance. Some general guidelines are provided below:

“1. Use vocabulary that will be widely accessible to students. Avoid colloquial and idiomatic expressions, words with multiple meanings, and unduly challenging
words that are not part of the construct. 2. Keep sentence structures as simple as possible to express the intended meaning. For ELLs, a number of simple sentences are often more accessible than a single more complex sentence. 3. Avoid use of negatives and constructions utilizing not in the questions’ stems and options as they can cause confusion, especially for ELLs. 4. When a fictional context is necessary (e.g., for a mathematics word problem), use a simple context that will be familiar to as wide a range of students as possible. A school-based context will often be more accessible to ELLs than a home-based context.” (p.13)

**ELL Students’ Performance on Mathematics**

Abedi, Leon, and Mirocha (2000) compared students’ performance on state content assessments and level of language proficiency. They concluded that ELLs, particularly those with limited English proficiency, perform substantially lower than native English speakers and the gap between ELLs and non-ELLs increases as the language level increases. The assessments may not present a true picture of the content knowledge ELLs understand. Although these findings may not be surprising, the authors were able to provide statistical evidence across many states and school districts about how large the gap is between ELLs and their English-speaking peers.

Abedi and Lord (2001) investigated the importance of language factor in students’ mathematics performance in terms of word problems. Students were given released items from the NAEP mathematics assessment, along with parallel items that were modified to reduce their linguistic complexity. This study utilized mixed methodologies to investigate the Research Questions. In interviews, students typically preferred the revised items over the original counterparts. Tests in paper-and-pencil format containing original and revised
items were administered to 1,174 eighth grade students. ELL students were found to have lower scores on the mathematics test than non-ELLs. There were also differences in mathematics performance in terms of SES but not gender. Linguistic modification of test items led to significant differences in mathematics performance; scores on the linguistically modified version were slightly higher. Some student groups benefited more from the linguistic modification of items such as students in low-level and average mathematics classes, ELLs and low SES students.

Similarly, Abedi (2002) utilized existing data from several locations across the U.S. to examine the impact of students’ language background on the mathematics test performance. The analyses mainly focused on the comparison between the level of performance of ELL and non-ELL students. In addition, to develop an understanding about the role of other contributing factors in the assessment of ELL students, comparisons were also made between students with regard to other background variables, such as parent education and family income. Students mean normal-curve equivalent (NCE) scores on different subscales of standardized tests were compared across subgroups using analysis of variance (ANOVA) and t tests in a multiple-comparison framework. The results discovered that ELLs generally perform lower than non-ELL students on reading, science, and mathematics. Meanwhile, this study revealed a strong indication of the impact of English language proficiency on assessment.

Beal et al. (2010) focused on the relationship of English proficiency and mathematics performance among high school students. The sample included 47% ELL students. Data sources included state mathematics test scores, study-specific pre- and posttest scores, problem solving in an online mathematics tutorial, and responses to a self-
report assessment of mathematics self-concept. Results indicated that, although overall mathematics performance was poor, there were significant variations related to English proficiency, with the ELLs scoring less well than the students who spoke English as their primary language. In addition, the increase of mathematics test scores for the ELLs corresponded to English-reading proficiency in a nonlinear manner. ELLs’ English-reading proficiency predicted mathematics test scores, progress in the online mathematics tutorial, and mathematics self-concept.

Fry (2007) studied the achievement gap in mathematics and reading between ELL students and other student groups as measured by the NAEP, which examines fourth and eighth grade students in mathematics and reading and provides national level results. This achievement analysis is based on the 2005 NAEP and 35 state-administered. The report also used demographic data across the nation to analyze some of the characteristics of limited English-speaking students. According to Fry (2007), the 2005 assessment indicated that 46% of ELL students in Grade 4 achieved at the below basic level in mathematics. 73% of ELL test-takers in Grade 4 were below basic in reading. Among white Grade 4 test takers nationally, 11% were at the below basic level in mathematics and 25% were below basic in reading. In 2011 NAEP, ELL students’ national performance in eighth grade reading and mathematics on the NAEP has continued to lag far behind Whites, Blacks, and Hispanics. For instance, there were 72% of eighth grade ELL students scoring at the “below basic” level on the mathematics section of the NAEP (NCES, 2012).

Using the data from Early Childhood Longitudinal Survey, Kindergarten Class of 1998–1999, Chang (2008) found that ELL students across four ethnic groups, including Asian, Black, Hispanic, and White, performed significantly lower in mathematics than their
non-ELL peers. From kindergarten to fifth grade, this initial gap became wider for Hispanic and Asian ELL students. For White ELL students, the mathematics achievement gap became narrow over this time while failing to close. For Black ELL students, interpretations were deferred as a result of the low sample size of this group.

Above all, previous studies and statistics have been found to provide consistent evidence that there is an achievement gap between ELL students and non-ELL students. Evidence of this achievement gap has led educational researchers to concern the factors influencing ELL students’ performance in standardized tests. To this end, the following section includes previous studies on exploring the reasons underlying the performance of ELLs relative to non-ELLs on standardized tests. Since this dissertation mainly concentrates on the mathematics assessment, this section mainly reviewed factors influencing bilingual or ELL students’ performance in mathematics standardized tests.

**Factors Influencing ELL Students’ Mathematics Achievement**

This section reviewed previous studies on exploring the factors influencing ELL students’ performance on mathematics standardized tests. Four factors were identified by previous studies including primary culture, parental involvement, English language proficiency, and school characteristics.

**Primary Culture**

Primary culture has been found to influence mathematics achievement for ELLs or bilingual learners. Chen and Stevenson (1995) examined the motivation and mathematics achievement of Asian-American, White-American, and East Asian students. 304 Asian-American, 1,958 White-American, 1,475 Chinese (Taiwan), and 1,120 Japanese eleventh graders (mean age = 17.6 years) were selected to participate this study. Students were given
a curriculum-based mathematics test and a questionnaire. Analysis of variance was used to compare scores among distinct groups of students. The results discovered that Asian-American students obtained higher scores in mathematics than those of White-American students but lower scores than those of Chinese and Japanese students. In addition, factors associated with the achievement of Asian-American and East Asian students included having parents and peers who set high standards, believing that the road to success is through effort, having positive attitudes about academic achievement, studying diligently, and facing less interference with their schoolwork from jobs and informal peer interactions. Finally, Asian-American students were found not to report a greater frequency of maladjusted symptoms than White-American students.

Kao (1995) used the National Education Longitudinal Study of 1988 (NELS:88) to compare Asian and white eighth graders on reading and mathematics test scores and grades. Analysis of variance was used to compare scores among different groups of students. Results indicated that the difference between Asians and whites on reading and mathematics test scores can be explained by differences in family background. However, analyses by Asian subgroups revealed that Chinese, Korean, and Southeast Asian youth receive higher mathematics scores while Pacific Islanders earn considerably lower mathematics and reading scores than their white counterparts. Analyses of Asian subgroups show no statistical difference between ethnic groups.

Recently, Roberts and Bryant (2011) found that parental SES and educational resources do account for the average differences among ELLs. They used data from the Early Childhood Longitudinal Survey, Kindergarten Class of 1998 –1999, to “(a) estimate mathematics achievement trends through 5th grade in the population of students who are
English-language proficient by the end of kindergarten, (b) compare trends across primary language groups within this English-language proficient group, (c) evaluate the effect of low SES for English-language proficient students and within different primary language groups, and (d) estimate language-group trends in specific mathematics skill areas” (p.1).

The group of English-language proficient ELLs was disaggregated into native Spanish speakers and native speakers of Asian languages, the 2 most prevalent groups of ELLs in the United States. Multilevel latent variable growth modeling was used in this study. The findings suggested that SES may be more salient than primary culture when explaining the mathematics achievement of English-language proficient ELLs.

According to those three studies above, the success of Asian Americans in mathematics achievement mainly stems from cultural differences when compared to White Americans. A crucial cultural difference is that Asian parents invest more in educational resources than their white counterparts despite comparable family incomes. Hence, the availability of educational resources is partly driven by cultural values. When comparing the Spanish speakers and native speakers of Asian languages, SES will be a stronger predictor to explain the differences of mathematics achievement among ELLs or bilingual learners.

By comparison, Latino families may not become as involved in their children’s education as non-Latino parents. One reason is that Latino parents hold the belief that it is the school’s responsibility to deal with their child’s misbehaviors or academic concerns. In addition, Latino parents feel uncomfortable questioning teachers or school decisions for fear of being disrespectful (Sue & Sue, 2008). These cultural differences between home
and school can contribute to the difficulties that ELL families experience in navigating U.S. schools (Arias & Morillo-Campbell, 2008).

**Parental Involvement**

Although parent involvement is a key factor in the academic development of all children, it may be particularly important for families from diverse cultural and linguistic backgrounds (Harper & Pelletier, 2010). Hartsock (2004) investigated whether a relationship existed between parent involvement in homework and the mathematics achievement of ELL student and native speakers of English in third grade. 132 third grade students and selected parents participated in a program of homework in mathematics, which were classified into four groups: non-English proficient, limited English proficient, fully English proficient, and native speakers of English. A mixed methodology was applied in this study. In the quantitative part, Pearson Product-Moment Correlation Coefficient was computed. The qualitative piece utilized case study research methodology. In the quantitative results, parent involvement and mathematics achievement were positively related. The relationship varied across language proficiency groups and it was higher for non-English proficient and limited English proficient students. In the qualitative results, factors affecting the levels of support at home included the parents' perception of the need of the child, the level of English language proficiency of the child, the parents' perception of their instructional role in the education of their children, the students' predisposition to allow assistance in Spanish, and the parents' assessment of the quality of the homework assignment.

Harper et al. (2010) assessed parents’ communication, involvement and knowledge of their children’s abilities in reading and mathematics among parents who spoke English
as a first language (EL1) and those who were ELLs. 42 kindergarten-aged children, their parents and their teachers participated in this study. Analysis of variance and chi-square tests were used to analyze the data. This study found that that ELL parents communicated more frequently with the teacher than ELL parents. However, there were no language group differences in parents’ involvement in their children’s education (as rated by the teacher). Parents’ ratings of their children’s abilities in mathematics did predict their children’s mathematics scores. It is concluded that involvement of ELL students’ parents at home and their greater understanding of and emphasis on mathematics learning will result in more accurate knowledge of their children’s abilities.

Niehaus (2012) in his dissertation study investigated the relationships between school support, parental school involvement, and academic and social-emotional outcomes (reading and mathematics) for ELL students. Restricted-use data obtained from direct child assessments, children’s self-reports, and parent, teacher, and school administrator surveys from the Early Childhood Longitudinal Study-Kindergarten Cohort of 1998 were analyzed. Major findings included more parental involvement is linked to fewer social-emotional concerns for ELL students at school; higher levels of school support predict more parental involvement among ELL families; and fewer social-emotional problems are associated with higher achievement scores. In addition, contrary to expectations, results showed that ELL students had lower achievement and more social-emotional concerns when they attended schools that provided more support services.

However, one recent dissertation study conducted by Rodriguez (2016) failed to find the significant relationship between parental involvement and ELL students’ mathematics achievement. This study examined the relationship between parental
involvement of seventh grade middle school Latino students and students’ reading and mathematics achievement. A non-experimental correlational research methodology was used to obtain a better understanding about the types and intensity of Latino immigrant parental involvement and the relationship to their children’s reading and mathematics grades in middle school. The participants in the study included 134 Latino immigrant parents. Correlational and multiple regression analyses were used to test the research questions and examine the hypotheses. The results of the multiple regression analyses revealed that there are not any significant relationships between parental involvement and their children’s reading and mathematics first-quarter grades.

English Language Proficiency

Stanley (2005) aimed to examine if there are performance differences between those students labeled Limited English Proficient (LEP) and those labeled English Only on achievement tests in mathematics and whether the amount of language in the assessment items affect the difference in performance. A sample of 916 third grade students from a California school district were included in this study. A number of inferential statistical analyses were performed explore the research questions, including linear regression analyses, partial eta squared, ANOVA, and analyses of covariance (ANCOVA). The results discovered that there is a significant difference in mathematics achievement between LEP and non-LEP students. Besides, this achievement discrepancy decreases as the amount of language in the texts decreases.

Mando (2007) explored the relationship between language proficiency and academic achievement of eighth grade ELL student in a school district in Georgia. A total of 187 eighth grade ELL students were chosen for this study. Mathematics achievement
scores and language proficiency scores were analyzed to observe the extent that English language proficiency predicts academic achievement as measured by the Georgia Criterion-Referenced Competency Test (CRCT). Data was used from the Accessing Comprehension and Communication in English State-to-State for ELLs (ACCESS), and the CRCT. The results indicated that on the average, students with higher levels of English language proficiency have higher levels of mathematics achievement.

A group of factors were examined by Prasad (2007) to see their influences on language arts and mathematics achievement of ELL students referred for special education evaluation. Factors that were incorporated into the study include gender, number of years in English speaking schools, Spanish language skills, English language skills, nonverbal cognitive ability, Spanish academic skills, and English academic skills. The data were collected by examining past special education testing and year-end competency-based testing in language arts and mathematics. A total number of 182 students from 59 schools were included in this study. In addition to descriptive statistics, Hierarchical Linear Models (HLM) were used to evaluate the predictive effects of initial special education testing on students’ language arts and mathematics achievement on basic competency skills tests. HLM was also used to investigate school effects on language arts and mathematics achievement. Findings indicated that a broad measure of English language skills is the best predictor of achievement in both language arts and mathematics. Exploratory analyses revealed that in addition to English language skills and IQ, academic performance on certain subtests of reading, written language, and mathematics is a good predictor of achievement on competency-based measures of language arts and mathematics for ELL students.
Mosqueda (2010) investigated the impact of English proficiency and tracking on the mathematics achievement of Latino English learners. A nationally representative sub-sample of 2,234 native and non-native English-speaking Latino 10th graders from Education Longitudinal Study of 2002 (ELS:2002) dataset were included in this study. This study applied generalized least squares (GLS) regression analysis to fit multilevel models that describe the mathematics achievement of native and non-native English-speaking Latino students as a function of their English proficiency, their placement in a general or academic track in school, and whether they are provided with native language support. The findings revealed that Latino non-native English speakers with low levels of English proficiency perform at much lower levels than their native-English speaking peers—about one standard deviation lower. In addition, school level factors (e.g., teacher qualification) also played a strong role in mediating their achievement.

Grant, Cook and Phakiti (2011) investigated whether there are meaningful relationships across academic English language proficiency, as measured by ACCESS of ELLs. 613 ELL students in Grade 3 to 5 and 560 in Grade 6 to 8. Structural equation modeling was selected to investigate relationships between English language proficiency and mathematics achievement. The findings suggested that success in mathematics is influenced by English language proficiency in both productive (writing and speaking) and receptive (listening and reading) skills, with receptive skills being more closely associated with success on mathematics content tests. Receptive skills in both general and technical areas directly influence mathematics achievement.

Denfield (2013) investigated the predictive power of English proficiency on mathematics scores, controlling for gender, SES, and grade level among ELLs at the south
Florida elementary school. Multiple linear regression was applied to analyze archival data for 177 ELLs in Grades 3 to 5. English proficiency emerged as a statistically significant predictor of mathematics scores. Results revealed that mathematics scores increased simultaneously with English proficiency but inversely with grade level. Grade level moderated the effect of English proficiency on mathematics scores, but gender and SES had no significant moderating effect. Factors other than English proficiency might impact ELLs' decreasing mathematics achievement and warrant further research.

Chen and Chalhoub-Deville (2016) mentioned that previous studies on the relationship between language proficiency and mathematics achievement show conflicts supporting either an increasing or a decreasing longitudinal relationship. They aimed to detect more information on the long-term relationship between language proficiency and mathematics achievement. SES, gender, and ethnicity background were taken into consideration at the same time. A longitudinal data was analyzed using quartile regression to overcome several limitations of previous studies. Results confirmed a persistent relationship between language proficiency and mathematics achievement. More importantly, it revealed that the strength of the relationship between those two differed for students with various abilities both within and across grades. For example, the relationship between language proficiency and mathematics achievement is increasing until the 75th percentile of conditional mathematics ability at Grade 1 but decreasing at all other grades. Besides, the variance of mathematics score distribution is larger at high language proficiency and smaller at low language proficiency at Grade 1 but the opposite pattern was found for later grades.
Henry, Nistor and Baltes (2016) investigated the predictive power of English proficiency and mathematics achievement. Data from the Florida Comprehensive Assessment Test for Grade 3-5 ELL students were analyzed using multiple linear regression. Gender, SES, and grade level were controlled. Findings revealed that English proficiency is a statistically significant predictor of mathematics achievement. Mathematics achievement increases simultaneously with English proficiency but inversely with grade level. Besides, the influence of English proficiency on mathematics achievement was found to be moderated by grade level.

**School Characteristics**

In addition to those individual and family characteristics, some studies have examined how school characteristics are influencing the mathematics performance of ELL students.

Computer use at school has been considered as an effective approach to develop ELL students’ mathematics achievement (Freeman & Crawford, 2008; Ganesh & Middleton, 2006; Kim & Chang, 2010). For instance, using data from the Early Childhood Longitudinal Study-Kindergarten Cohort of 1998, Kim et al. (2010) conducted both cross-sectional and longitudinal analyses to investigate the direct and longitudinal effects of computer use in classrooms. It has been found that computer use for mathematics is associated with a reduced gap in the mathematics achievement between ELL and non-ELL students. Specifically, when Hispanic and Asian students frequently used computers for mathematics, they showed high mathematics performances when compared with their non-ELL peers.
Finally, ELL student performance can be related the availability of school educational resources. Han and Bridglall (2009) utilized growth curve modeling to identify the association between the school educational resources available to ELLs and their academic trajectories from kindergarten through fifth grade. Results showed that ELL students either in a high or low ELL school close the initial achievement gap in mathematics from kindergarten to fifth grade. This finding highlighted the importance of school educational resources that are tailored to ELL students and families.

Assessment Validity for ELLs

Validity, as one of the most important attributes of an assessment, refers to how well the assessment tool actually measures the underlying outcome of interest. Validity is not a property of the tool itself, but rather of the interpretation or specific purpose of the assessment tool with particular settings and learners (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 1999.). For ELLs, as well as for all populations, it is critical to consider whether the test scores can really reflect the skill or proficiency that an assessment is intended measure. Although students may have different English proficiency, the meaning of their scores on content assessments should be comparable (Educational Testing Service [ETS], 2009).

Construct related evidence for validity of an assessment refers to the degree of association between the test score and what ability it is meant to describe or predict. Threats to the validity of the test score interpretation can occurs from either (a) construct-irrelevant variance (measuring something other than construct of interest) or (b) construct under-representation (incomplete measurement of the construct). Specifically, construct-
irrelevant variance occurs when one or more constructs are being measured which lowers or raises scores for subgroups of students. Potential sources of construct-irrelevant variance may include linguistic demands of items, item format, response mode, and rater’s attention to irrelevant features of responses. Construct underrepresentation occurs when target construct is not fully captured. Then the generalizability of score inferences to the larger domain will be jeopardized. For example, if ELLs are not measured in their home language, their access to the target construct will be limited. Test scores cannot represent their proficiency on the target construct (Messick, 1989).

In addition, according to Standard for Educational and Psychological Testing, “the linguistic or reading demands of the test should be kept to the minimum necessary for the valid of assessment for the intended construct” (AERA, APA, & NCME, 1999, p.82). Since ELLs test takers have not yet acquired sufficient mastery of English language, high language demand is still evident on mathematics assessment (Loughran, 2014). For instance, ELLs who need to solve a word problem may fail to understand the task due to limited English proficiency. In this case, English language proficiency also becomes the purpose of assessment in addition to mathematical ability. Consequently, the lack of English proficiency used to comprehend mathematics assessment items, which results in an increasing cognitive load and contributes to measurement error of ELL students’ mathematics content knowledge (ETS, 2009).

Differential Item Functioning

Description of DIF

According to Holland and Wainer (1993), DIF analysis is a statistical technique to identify whether items on an assessment are of equal difficulty for examinees of distinct
groups. DIF is present if an item on a test functions differently for different groups of interest (e.g., ELLs vs. non-ELLs), given the ability level. In the DIF analysis, examinees are matched based on their underlying ability (e.g., total score of an assessment), and differences in item performance between groups of examinees at the same level of ability are then determined. An item displays DIF if matched examinees have significantly different probabilities in answering an item correctly while non-DIF is identified if matched examinees have a similar probability of getting an item correctly. The presence of DIF items suggested multidimensionality in items. In addition to the measurement of an intended ability/dimension, persons require at least a secondary ability/dimension to answer the items (Qian, 2011; Roussos & Stout, 2004). Usually the groups are called focal group and reference group. The focal group (e.g., ELLs) is usually the group of interest while reference group (e.g., non-ELLs) is the comparison group.

There are two types of DIF: uniform and non-uniform (Mellenbergh, 1982). Uniform DIF occurs when one group constantly performs better than the other group across all score levels of the attribute. For example, non-ELL students systematically perform better than matched ELL students on test items. Non-uniform DIF is present if the probability of giving a certain response to the item in the two groups is not the same for all levels of the attribute. Non-uniform DIF represents an interaction between the proficiency and performance differences across groups. For example, when high proficiency non-ELL students outperform high-proficiency ELL students, then the pattern change to low-proficiency ELL students outperform low-proficiency non-ELL students (Husin, 2014). According to (Loughran, 2014), DIF analyses can yield several results which reveal that
an item: 1) fails to show evidence of DIF or shows negligible DIF, 2) shows evidence of uniform DIF, or 3) shows evidence of non-uniform DIF.

In testing organizations, DIF analysis is a standard statistical procedure while it does not guarantee the content of test items. Therefore, panels of experts are recruited to review item for bias called judgmental methods (Camilli & Shepard, 1994). Results from DIF analysis and judgmental methods can be crosschecked to improve test equity.

**DIF Detection Methods**

There are numerous statistical methods for detecting DIF. The following methods are frequently used: Mantel-Haenszel procedure (Holland & Thayer, 1988), SIBTEST (Shealy & Sout, 1993), Item Response Theory methods (Camille & Shepard, 1994), logistic regression (Swaminathan & Rogers, 1990), and multilevel DIF analysis (Kamata, 2001). In this section, three methods applied in current study were discussed in particularly.

**Mantel-Haenszel (MH) Procedure**

The MH statistic was applied by Holland et al. (1988) in determining DIF. The MH DIF procedure compares dichotomous item performance between two groups after matching respondents on overall scores. The null hypothesis states that the population odds of getting an item correct is the same in the reference and focal groups. Respondents in the focal and reference groups were matched on total test scores by dividing respondents in both groups into defined strata on those scores. The total scores were generated by summing item scores across all items. Estimates of the odds ratio for a given item can be calculated based on a $2 \times 2 \times K$ contingency table with $k$ representing the $k$-th group, ($k = 1,2,\ldots,K$). The following shows the $2 \times 2$ contingency table for the $k$-th group of an item.
The $A_k$, $B_k$, $C_k$, and $D_k$ denote the numbers of respondents in the cells. $T_k$ represents the number of respondents in the k-the stratum.

The cells $A_k$ and $C_k$ represent the total number of respondents who answered the item correctly in the reference and focal groups, respectively, within the matched subgroup $k$. $B_k$ and $D_k$ denote the total number of respondents who answered the item incorrectly in the reference and focal groups, respectively, within subgroup $k$. $N_{rk}$ and $N_{fk}$ are the total number of respondents in the reference and focal groups, respectively, within $k$-th group. $M_{1k}$ and $M_{0k}$ denote the number of respondents who answered the item correct and incorrect, respectively, within $k$-th group.

Table 1. $2 \times 2 \times K$ Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>1 (Right)</th>
<th>0 (Wrong)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reference Group</strong></td>
<td>$A_k$</td>
<td>$B_k$</td>
<td>$N_{rk}$</td>
</tr>
<tr>
<td><strong>Focal Group</strong></td>
<td>$C_k$</td>
<td>$D_k$</td>
<td>$N_{fk}$</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>$M_{1k}$</td>
<td>$M_{0k}$</td>
<td>$T_k$</td>
</tr>
</tbody>
</table>

The MH chi-square statistic is used for testing the null hypothesis of whether the population odds of getting an item correct is the same in the reference and focal groups. The statistic is given by

$$MH_{CHISQ} = \frac{\left(1 \sum^n_{k=0} A_k - \sum^n_{k=0} E(A_k) - \frac{1}{2}\right)^2}{\sum^n_{k=0} Var(A_k)}$$

In this equation, $E(A_k)$ and $Var(A_k)$ follow:

$$E(A_k) = \frac{N_{fk} \cdot M_{1k}}{T_k}$$

$$Var(A_k) = \frac{N_{rk} \cdot N_{fk} \cdot M_{1k} \cdot M_{0k}}{[T_k^2 (T_k - 1)]}$$
The common odds-ratios formula is:

\[ \text{OR}_{\text{MH}} = \frac{\sum_{k=0}^{n} \frac{A_k + D_k}{T_k}}{\sum_{k=0}^{n} \frac{B_k + C_k}{T_k}} \]

The scale for \( \text{OR}_{\text{MH}} \) is from 0 to \( \infty \), with \( \text{OR}_{\text{MH}} = 1 \) denoting the case of no DIF. For convenience, \( \text{OR}_{\text{MH}} \) is converted into a symmetrical scale \( \Delta \text{OR}_{\text{MH}} \) given as

\[ \Delta \text{OR}_{\text{MH}} = -2.35 \ln (\text{OR}_{\text{MH}}) \]

\( \Delta \text{OR}_{\text{MH}} \) is applied as a measure of the degree of DIF referred to as DIF effect size.

Educational Testing Service classified the DIF effect size as follows (Dorans & Holland, 1993) in order to aid in interpretation in applications: Class A denotes negligible magnitudes of DIF, when \( |\Delta \text{OR}_{\text{MH}}| < 1.00 \); Class B denotes moderate magnitudes of DIF, when \( 1.00 \leq |\Delta \text{OR}_{\text{MH}}| < 1.50 \), and Class C denotes large magnitudes of DIF, when \( |\Delta \text{OR}_{\text{MH}}| \geq 1.50 \).

The validity for MH DIF detection method has been established by numerous studies (Qian, 2011). It is the most widely used procedure to detect DIF in practice, since it is not only easy to understand and compute, it can provide both a significance test and estimate of the magnitude of DIF as well. (Millsap, 2011). The major criticism of the MH procedure is the adequacy of using the total score as a substitute for the latent trait (Millsap, 2011). Besides, requirement for sample sizes is a technical challenge to detecting items with DIF since DIF statistics become less stable as sample sizes decrease.

**Rasch Model**

The Rasch model (Rasch, 1960) can produce a comprehensive and informative picture of the construct under measurement as well as the respondents on that measure. The Rasch model for dichotomous data is similar to one-parameter Item Response Theory
(IRT). However, Rasch model has several unique features. First, the Rasch model enables to parameterize each individual for estimating item difficulty while one-parameter IRT parameterize the entire sample by mean score and standard deviation. Second, the Rasch model can provide diagnostic fit statistics to examine the performance of each individual and function of each item. One-parameter IRT can only provide a “global” fit to accept or reject the model. Third, data is not required to be approximately normal in the Rasch model. Fourth, it is robust for missing values in the responses (Linacre, 2005).

Since PISA employs Rasch model to estimate student ability, item difficulty, and create the overall PISA literacy scale (OECD, 2012), Rasch model was selected to detect DIF in current study. The Rasch model follows mathematically from the requirement of invariance of comparisons among persons and items (Andrich & Luo, 2003). The Rasch model follows the following form:

$$P_{ij} (Y_{ij}=1 | \theta_j, b_i) = \frac{1}{\left[1 + e^{-(\theta_j - b_i)}\right]}$$

where \(p_{ij}\) is the probability of person \(j\) answering correctly to item \(i\). \(\theta_j\) is the person trait, or ability. \(b_i\) is the item parameter indicating difficulty of the item.

The Rasch model provides a theoretically useful way to detect DIF which can be modeled using estimated item parameters and ability. According to Wen (2014), the understanding of DIF detection can be clarified through the assumptions of IRT. First, the unidimensionality assumption corresponds to the multidimensionality perspective on why DIF occurs. Second, the local independence assumption implies that any pair of items is independent. Third, the item and sample invariance assumption states the item should not differ across samples, which supports the reason for detecting DIF.
In the Rasch model, item location depends on group membership. If this is the case, it suggests the latent variable is being defined differently across groups. If these differences are substantial, then group comparisons are problematic since the latent variables are not being defined in the same way and are therefore not comparable (Meredith, 1993). In Winsteps (Linacre, 2017), DIF detection using Rasch model is conducted by a subtraction of the item location parameters (item difficulties) for two groups, \( d_1 \) and \( d_2 \). They are converted to standard normal variates using a pooled standard error. DIF is detected if the difference of item location parameters is statistically different.

\[
t = \frac{d_1 - d_2}{\sqrt{\text{var}(d_1) + \text{var}(d_2)}}
\]

**Hierarchical Generalized Linear Model (HGLM)**

In behavioral and social sciences, data commonly have a nested structure (Wen, 2014). For example, repeated observations are nested within persons (e.g., responses nested within examinees), and persons are nested within organizational units such as classrooms, schools, and communities, and so on. Students within a particular hierarchy share some common characteristics and experiences as a result of being in the same environment (e.g., demographic, environmental, and instructional). The nature of this hierarchical structure undermines the statistical assumption that students are independent from each another and thus causes aggregation bias (Raudenbush & Bryk, 2002).

Multilevel models, as extensions of standard multiple regression, have been designed to handle interdependencies among the data points. Multilevel models have been termed as hierarchical linear model (Raudenbush & Bryk, 1986), random coefficient models (Longford 1993), and linear mixed model (Littell, Milliken, Stroup & Wolfinger, 1996). One assumption for multilevel model is that the dependent variable should be
continuous and normally distributed. However, this assumption is violated if the dependent variable is a dichotomous variable (Hox, 2010). To this end, a variant of multilevel models named hierarchical generalized linear model (HGLM) has been proposed when the dependent variable is neither normal nor continuous, and the relationship between the the dependent variable and predictor variable is not linear.

Kamata (2001) proposed to use HGLM to detect DIF effects. Dichotomous Rasch model has been shown to be a special case of HGLM (Kamata, 2001; Raudenbush, Johnson, & Sampson, 2003). The model is shown below with three-level model notation. The first level of the model is the item level, the second level the student level, and the third level is the school level. The three-level models have a nested structure where items are nested within students, and students are nested within schools.

The Level-1 model is an item-level model. For a student, the response on the item can be formulized as:

$$\text{Log} \left( \frac{p_{ijk}}{1-p_{ijk}} \right) = \beta_{0jk} + \sum_{q=1}^{k-1} (\beta_{qjk} * X_{qijk})$$

Where $X_{qijk}$ is the $q$-th ($q = 1, 2, \ldots, m-1$) is the dummy variable that indicates the item $i$ for student $j$ in school $k$. Its value is 1 when $q = i$ and 0 when $q \neq i$. $\beta_{0jk}$ is the effect of the reference item and $\beta_{qjk}$ is the difference between the $q$-th item and the reference item. The probability of student $j$ in school $k$ getting an item $i$ correct is noted as $p_{ijk}$.

Level-2 models are the student-level models. Each item effect coefficient in equation (8) is further modeled across schools as given by the following equations:

$$\beta_{0jk} = \gamma_{00k} + \mu_{0jk}$$

$$\beta_{qjk} = \gamma_{q0k}$$
The random effect $\mu_{0jk}$ represents the variance of students' ability in school $k$, with $\mu_{0jk}$ assumed to be randomly distributed. The notation of the random student ability in school $k$ is given by $\mu_{0jk} \sim N(\gamma_{00k}, \tau)$. The variance of students’ ability is denoted by $\tau$ and is assumed identical across schools. The parameter $\gamma_{00k}$ is the effect of reference item in school $k$, and $\gamma_{q0k}$ is the effect of the $i$-th ($i = 1, 2, \ldots, m-1$) item in school $k$.

The overall item effect can be further modeled at the additional school level. For school $k$ we have

$$
\gamma_{00k} = \pi_{00k} + \theta_{00k}
$$

$$
\gamma_{q0k} = \pi_{q00}
$$

Where $\theta_{00k} \sim N(0, \tau_0)$. At the school level, $\pi_{00k}$ and $\pi_{q00}$ are both fixed item effects. $\theta_{00k}$ is a random effect with variance $\tau_0$.

According to Kamata, Chaimongkol, Genc, and Bilir (2005), the three-level Rasch model allows the coefficient corresponding to the person-level DIF to be random across higher level clusters (schools in their study). To this end, the item-level model remains the same, and the student-level model becomes

$$
\beta_{0jk} = \gamma_{00k} + \mu_{0jk}
$$

$$
\beta_{qjk} = \gamma_{q0k} \quad \text{If no DIF}
$$

$$
\beta_{qjk} = \gamma_{q0k} + \gamma_{q1k} * G_{qjk} \quad \text{Otherwise}
$$

where $G_{qjk}$ is the group membership at the student level and $\gamma_{q1k}$ is the effect of DIF. Then the level-3 model becomes

$$
\gamma_{00k} = \pi_{00k} + \theta_{00k}
$$

$$
\gamma_{q0k} = \pi_{q00}
$$

$$
\gamma_{q1k} = \pi_{q10} + \theta_{q1k}
$$
where $\theta_{q1k}$ is the random effect of DIF across schools. If the variance of $\theta_{q1k}$ is larger than 0, the DIF effect varies across schools, indicating the effect of group membership at the student level is different from school to school.

Kamata (2001) mentioned that the three-level HGLM would be useful when the variation of the effect of a student characteristic variable across groups and the identification of a group-characteristic variable that explains such variation are of interest. There are several advantages of using HGLM to detect DIF in the large-scale assessments. First, since the dependency of the data due to the nested data structure can be considered, DIF and item difficulty parameters can be modeled randomly across schools. Then student and school variables can be examined simultaneously as source of DIF. Second, additional student level variables can be added as covariates to reduce student variations when identifying DIF. Third, various sources of DIF unique to each DIF item can be modeled simultaneously. Fourth, DIF detection using HGLM does not require two separate groups (focus and reference groups). This is especially beneficial if the source of the hypothesized DIF is a continuous variable (Adams, Wilson, & Wu, 1997; Qian, 2011)
CHAPTER 3. METHOD

Data Source

Overview of PISA 2012

The primary database used in this research is constructed from the PISA 2012. PISA is the most comprehensive and rigorous international assessment on 15-year-old student performance in mathematics, reading, and science. The first PISA study took place in 2000, and it takes every three years to collect data on the student, family and institutional factors that can help to explain differences in performance. For each assessment, one of mathematics, reading, and science is chosen as the principal domain and given greater emphasis. The remaining two minor domains are assessed less thoroughly. Mathematics was selected as the principal domain in 2003 and 2012; reading was selected as the principal domain in 2000 and 2009; and science was selected as the principal domain in 2006 (Organization for Economic Co-operation and Development [OECD], 2014).

PISA aims to investigate how well students are prepared to meet the challenges of the future and how well students are prepared for life in a larger context, rather than how well they master a particular curriculum. PISA also collects information from students using Student Questionnaires on various aspects of their home, family and school background, and from schools using School Questionnaires about various aspects of organization and educational provision in schools. In PISA 2012, 11 countries also administered a Parent Questionnaire to the parents of the students participating in PISA (OECD, 2014).

PISA 2012 examined and compared the performance of schools and education systems in all 34 OECD member countries and 31 partner countries. It employed a two-
stage stratified sample design for assessments. The first-stage sampling units included individual schools having 15-year-old students. Schools were selected systematically from a national list of all PISA-eligible schools, with probabilities that were proportional to a measure size. The second-stage sampling units were students within sampled schools. A Target Cluster Size (TCS) was set for each country. This value was typically 35 students who were selected with equal probability (OECD, 2014). Approximately 510,000 students between the ages of 15 years 3 months and 16 years 2 months participated in the assessment were selected to take a standardized test representing about 28 million 15-year-olds globally (OECD, 2013).

**PISA 2012 Assessment Design**

The three domains of mathematics, reading, and science were assessed in PISA 2012. Their construct definitions were defined by OECD (2013) as follows:

- **Mathematical literacy**: An individual’s capacity to formulate, employ, and interpret mathematics in a variety of contexts. It includes reasoning mathematically and using mathematical concepts, procedures, facts and tools to describe, explain and predict phenomena. It assists individuals to recognize the role that mathematics plays in the world and to make the well-founded judgments and decisions needed by constructive, engaged and reflective citizens.

- **Reading literacy**: An individual’s capacity to understand, use, reflect on and engage with written texts, in order to achieve one’s goals, to develop one’s knowledge and potential, and to participate in society.

- **Scientific literacy**: An individual’s scientific knowledge and use of that knowledge to identify questions, to acquire new knowledge, to explain scientific
phenomena, and to draw evidence-based conclusions about science-related issues, understanding of the characteristic features of science as a form of human knowledge and enquiry, awareness of how science and technology shape our material, intellectual, and cultural environments, and willingness to engage in science-related issues, and with the ideas of science, as a reflective citizen. (p. 17)

PISA 2012 utilized paper-based instruments which included a total of 270 minutes material. The material was arranged into nine clusters of items with each cluster representing 30 minutes of testing time. The clusters were assigned into test booklets based on a rotated test design, with each form consisting of four clusters of materials from the mathematics, reading, and science domains. Each student completed one form, representing a total of 120 minutes testing time. PISA 2012 assessment contained 13 booklets with anchor (common) items between the booklets. Each student might not take the same items but a sufficient number of students took each booklet so that appropriate estimates for all items could be made (OECD, 2013). The anchor items were used to link item parameters using IRT scaling procedures before equating procedures (OECD, 2010). The application of IRT with anchor items enabled students to be scored on the same scale even if they responded to different sets of booklets (Shivraj, 2014).

Sample

According to National Center for Education Statistics ([NCES], 2017), PISA 2012 U.S. sample was stratified into eight explicit groups based on control of school (public or private) and regions including Northeast, Central, West, Southeast. Within each stratum, five categorical stratification variables were included: grade range of the school; type of location relative to populous areas (city, suburb, town, rural); minority status; gender; and
Within each school, 50 students aged 15 were randomly sampled. Each age-eligible student had an equal probability of being selected. Sampled students were born between July 1, 1996, and June 30, 1997. Sampled students mainly came from three grades—grades 9, 10, and 11. Finally, the PISA 2012 U.S. sample contains 4,978 students from 162 schools.

The targeted population for the focal group in this study is ELL students. Groups were identified using information collected from Student Questionnaire that was administered with the test. Basically, these two groups were designed to differ only in their relationships with English (as a first or second language). Home Language (ST25Q01) was used to form the groups. Home Language has the following binary categories: (1) language at home is same as the language of the test and (2) language at home is another language. Students failing to answer this question were excluded from current study. Finally, 670 students were identified as ELL students while 4,196 students were identified as non-ELL students.

Measurement

Mathematics Assessment

Content Category

A total of four content areas in mathematics were assessed in PISA 2012 including space and shape, quantity, change and relationships, and uncertainty and data, which relate to curricular strands such as numbers, geometry, algebra, and probability and data analysis. Specifically, space and shape represents phenomena that are encountered everywhere in our visual and physical world such as patterns, properties of objects, positions and orientations, representations of objects, decoding and encoding of visual information,
navigation and dynamic interaction with real shapes as well as with representations; quantity encompasses the quantification of attributes of objects, relationships in the world, and judging interpretations and arguments based on quantity; change and relationship involves various temporary and permanent relationship among objects and circumstances, where changes occur within systems of interrelated objects; uncertainty and data cover the mathematical analysis of many problem situations, theory of probability and statistics, and techniques of data representation (OECD, 2013).

**Context Category**

PISA 2012 aimed to assess mathematics literacy that is engaged in solving a problem set in a context. The proper use of mathematical strategies is often dependent on the context where a problem arises. PISA 2012 mathematics assessment covers a variety of contexts. Specifically, four context categories have been defined and used to classify assessment items developed for PISA mathematics assessment: personal, occupational, societal, and scientific.

Problems in personal context category are mainly about activities of one’s self, family or peer group including (but are not limited to) shopping, games, personal health, sports, and travel; problems in the occupational context category relate to the world of work involving (but are not limited to) measuring, costing and ordering materials for building, payroll/accounting, quality control, scheduling/inventory, design/architecture and job-related decision making; problems in the societal context category center on one’s local, national, or global community including (but are not limited to) voting systems, public transport, government, public policies, demographics, advertising, national statistics and economics; problems in the scientific focus on the application of mathematics to the natural
world, science, and technology including (but are not limited to) weather, ecology, medicine, genetics, measurement, and space science (OECD, 2013).

Since all of these content areas are critical for constructive, engaged, and reflective citizens, PISA 2012 aimed to provide as balance a distribution of score points as possible. The approximate distribution of score points in mathematics assessment is 25% for each content category and context category (OECD, 2013).

Assessment Items

PISA 2012 assessment used a matrix-sampling design with spiraling booklet administration. Each student randomly received a booklet containing a subset of the total assessment items. 76 dichotomously coded items from PISA 2012 mathematics assessment were analyzed. These items are either selected response multiple choice and closed-constructed response. Multiple-choice items asked students to select or produce simple responses that can be directly compared with a single correct answer. Closed-constructed response items required students to construct a numeric response within specific limited constraint. These items were scored as correct or incorrect and coded dichotomously with 1 and 0. Brief descriptions of each item and counts of ELL and non-ELL students can be found in Table 2.

For PISA 2012, IRT was used to estimate student proficiencies for mathematics. The reporting scale for mathematics was the linear transformation of the natural logit metrics that result from the IRT scaling. Then the mean and standard deviation of the PISA 2012 mathematics score were 500 and 100 respectively (OECD, 2013). In the U.S. sample, the average score of ELL students is 456.60 while the average score of non-ELL students
is 486.91. An independent sample t-test revealed that the average mathematics score for non-ELL students were significantly higher than ELL students.

Table 2. Descriptions of Mathematics Assessment Items in PISA 2012.

<table>
<thead>
<tr>
<th>Item Number</th>
<th>PISA Variable Name</th>
<th>Description</th>
<th>ELL</th>
<th>Non-ELL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PM00FQ01</td>
<td>MATH - P2012 Apartment Purchase Q1</td>
<td>224</td>
<td>1386</td>
</tr>
<tr>
<td>2</td>
<td>PM00GQ01</td>
<td>MATH - P2012 An Advertising Column Q1</td>
<td>202</td>
<td>1291</td>
</tr>
<tr>
<td>3</td>
<td>PM00KQ02</td>
<td>MATH - P2012 Wheelchair Basketball Q2</td>
<td>174</td>
<td>1184</td>
</tr>
<tr>
<td>4</td>
<td>PM033Q01</td>
<td>MATH - P2000 A View with a Room Q1</td>
<td>187</td>
<td>1248</td>
</tr>
<tr>
<td>5</td>
<td>PM034Q01T</td>
<td>MATH - P2000 Bricks Q1</td>
<td>184</td>
<td>1239</td>
</tr>
<tr>
<td>6</td>
<td>PM155Q01</td>
<td>MATH - P2000 Pop Pyramids Q1</td>
<td>187</td>
<td>1246</td>
</tr>
<tr>
<td>7</td>
<td>PM155Q04T</td>
<td>MATH - P2000 Pop Pyramids Q4</td>
<td>187</td>
<td>1245</td>
</tr>
<tr>
<td>8</td>
<td>PM192Q01T</td>
<td>MATH - P2000 Containers Q1</td>
<td>195</td>
<td>1222</td>
</tr>
<tr>
<td>9</td>
<td>PM273Q01T</td>
<td>MATH - P2000 Pipelines Q1</td>
<td>207</td>
<td>1318</td>
</tr>
<tr>
<td>10</td>
<td>PM305Q01</td>
<td>MATH - P2000 Map Q1</td>
<td>195</td>
<td>1223</td>
</tr>
<tr>
<td>11</td>
<td>PM406Q01</td>
<td>MATH - P2003 Running Tracks Q1</td>
<td>195</td>
<td>1222</td>
</tr>
<tr>
<td>12</td>
<td>PM406Q02</td>
<td>MATH - P2003 Running Tracks Q2</td>
<td>194</td>
<td>1220</td>
</tr>
<tr>
<td>13</td>
<td>PM408Q01T</td>
<td>MATH - P2003 Lotteries Q1</td>
<td>207</td>
<td>1317</td>
</tr>
<tr>
<td>14</td>
<td>PM411Q01</td>
<td>MATH - P2003 Diving Q1</td>
<td>186</td>
<td>1242</td>
</tr>
<tr>
<td>15</td>
<td>PM411Q02</td>
<td>MATH - P2003 Diving Q2</td>
<td>186</td>
<td>1241</td>
</tr>
<tr>
<td>16</td>
<td>PM420Q01T</td>
<td>MATH - P2003 Transport Q1</td>
<td>207</td>
<td>1315</td>
</tr>
<tr>
<td>17</td>
<td>PM423Q01</td>
<td>MATH - P2003 Tossing Coins Q1</td>
<td>195</td>
<td>1222</td>
</tr>
<tr>
<td>18</td>
<td>PM442Q02</td>
<td>MATH - P2003 Braille Q2</td>
<td>185</td>
<td>1240</td>
</tr>
<tr>
<td>19</td>
<td>PM446Q01</td>
<td>MATH - P2003 The Thermometer Cricket Q1</td>
<td>207</td>
<td>1315</td>
</tr>
<tr>
<td>20</td>
<td>PM446Q02</td>
<td>MATH - P2003 The Thermometer Cricket Q2</td>
<td>206</td>
<td>1314</td>
</tr>
<tr>
<td>21</td>
<td>PM462Q01D</td>
<td>MATH - P2003 The Third Side Q1</td>
<td>184</td>
<td>1240</td>
</tr>
</tbody>
</table>
Table 2. (continued)

<table>
<thead>
<tr>
<th></th>
<th>Course Code</th>
<th>Course Title</th>
<th>Q1</th>
<th>Q2</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>PM464Q01T</td>
<td>MATH - P2003 The Fence Q1</td>
<td>203</td>
<td>1309</td>
</tr>
<tr>
<td>23</td>
<td>PM474Q01</td>
<td>MATH - P2003 Running Time Q1</td>
<td>187</td>
<td>1247</td>
</tr>
<tr>
<td>24</td>
<td>PM496Q01T</td>
<td>MATH - P2003 Cash Withdrawal Q1</td>
<td>195</td>
<td>1223</td>
</tr>
<tr>
<td>25</td>
<td>PM496Q02</td>
<td>MATH - P2003 Cash Withdrawal Q2</td>
<td>195</td>
<td>1222</td>
</tr>
<tr>
<td>26</td>
<td>PM559Q01</td>
<td>MATH - P2003 Telephone Rates Q1</td>
<td>205</td>
<td>1312</td>
</tr>
<tr>
<td>27</td>
<td>PM564Q01</td>
<td>MATH - P2003 Chair Lift Q1</td>
<td>193</td>
<td>1213</td>
</tr>
<tr>
<td>28</td>
<td>PM564Q02</td>
<td>MATH - P2003 Chair Lift Q2</td>
<td>193</td>
<td>1213</td>
</tr>
<tr>
<td>29</td>
<td>PM571Q01</td>
<td>MATH - P2003 Stop the Car Q1</td>
<td>193</td>
<td>1214</td>
</tr>
<tr>
<td>30</td>
<td>PM603Q01T</td>
<td>MATH - P2003 Number Check Q1</td>
<td>193</td>
<td>1219</td>
</tr>
<tr>
<td>31</td>
<td>PM800Q01</td>
<td>MATH - P2003 Computer Game Q1</td>
<td>202</td>
<td>1308</td>
</tr>
<tr>
<td>32</td>
<td>PM803Q01T</td>
<td>MATH - P2003 Labels Q1</td>
<td>185</td>
<td>1241</td>
</tr>
<tr>
<td>33</td>
<td>PM828Q01</td>
<td>MATH - P2003 Carbon Dioxide Q1</td>
<td>205</td>
<td>1311</td>
</tr>
<tr>
<td>34</td>
<td>PM828Q02</td>
<td>MATH - P2003 Carbon Dioxide Q2</td>
<td>205</td>
<td>1310</td>
</tr>
<tr>
<td>35</td>
<td>PM828Q03</td>
<td>MATH - P2003 Carbon Dioxide Q3</td>
<td>203</td>
<td>1309</td>
</tr>
<tr>
<td>36</td>
<td>PM903Q03</td>
<td>MATH - P2012 Drip Rate Q3</td>
<td>224</td>
<td>1394</td>
</tr>
<tr>
<td>37</td>
<td>PM905Q01T</td>
<td>MATH - P2012 Tennis Balls Q1</td>
<td>233</td>
<td>1432</td>
</tr>
<tr>
<td>38</td>
<td>PM906Q01</td>
<td>MATH - P2012 Crazy Ants Q1</td>
<td>179</td>
<td>1195</td>
</tr>
<tr>
<td>39</td>
<td>PM909Q01</td>
<td>MATH - P2012 Speeding Fines Q1</td>
<td>204</td>
<td>1298</td>
</tr>
<tr>
<td>40</td>
<td>PM909Q02</td>
<td>MATH - P2012 Speeding Fines Q2</td>
<td>204</td>
<td>1298</td>
</tr>
<tr>
<td>41</td>
<td>PM909Q03</td>
<td>MATH - P2012 Speeding Fines Q3</td>
<td>203</td>
<td>1298</td>
</tr>
<tr>
<td>42</td>
<td>PM915Q01</td>
<td>MATH - P2012 Carbon Tax Q1</td>
<td>180</td>
<td>1200</td>
</tr>
<tr>
<td>43</td>
<td>PM915Q02</td>
<td>MATH - P2012 Carbon Tax Q2</td>
<td>180</td>
<td>1198</td>
</tr>
<tr>
<td>44</td>
<td>PM918Q01</td>
<td>MATH - P2012 Charts Q1</td>
<td>227</td>
<td>1402</td>
</tr>
<tr>
<td>45</td>
<td>PM918Q02</td>
<td>MATH - P2012 Charts Q2</td>
<td>227</td>
<td>1401</td>
</tr>
<tr>
<td>46</td>
<td>PM918Q05</td>
<td>MATH - P2012 Charts Q5</td>
<td>227</td>
<td>1401</td>
</tr>
<tr>
<td>47</td>
<td>PM919Q01</td>
<td>MATH - P2012 Zs Fan Merchandise Q1</td>
<td>232</td>
<td>1431</td>
</tr>
</tbody>
</table>
| PM919Q02 | MATH - P2012 Zs Fan Merchandise Q2 | 230 1430  
| PM923Q01 | MATH - P2012 Sailing Ships Q1 | 227 1401  
| PM923Q03 | MATH - P2012 Sailing Ships Q3 | 227 1397  
| PM923Q04 | MATH - P2012 Sailing Ships Q4 | 227 1397  
| PM924Q02 | MATH - P2012 Sauce Q2 | 222 1383  
| PM943Q01 | MATH - P2012 Arches Q1 | 229 1427  
| PM943Q02 | MATH - P2012 Arches Q2 | 228 1424  
| PM949Q01T | MATH - P2012 Roof Truss Design Q1 | 203 1297  
| PM949Q02T | MATH - P2012 Roof Truss Design Q2 | 203 1295  
| PM953Q02 | MATH - P2012 Flu Test Q2 | 228 1421  
| PM953Q03 | MATH - P2012 Flu Test Q3 | 226 1419  
| PM954Q01 | MATH - P2012 Medicine Doses Q1 | 229 1430  
| PM954Q02 | MATH - P2012 Medicine Doses Q2 | 229 1430  
| PM954Q04 | MATH - P2012 Medicine Doses Q4 | 229 1429  
| PM955Q01 | MATH - P2012 Migration Q1 | 202 1291  
| PM955Q02 | MATH - P2012 Migration Q2 | 202 1290  
| PM982Q01 | MATH - P2012 Employment Data Q1 | 182 1203  
| PM982Q02 | MATH - P2012 Employment Data Q2 | 182 1201  
| PM982Q03T | MATH - P2012 Employment Data Q3 | 182 1201  
| PM982Q04 | MATH - P2012 Employment Data Q4 | 181 1201  
| PM992Q01 | MATH - P2012 Spacers Q1 | 180 1201  
| PM992Q02 | MATH - P2012 Spacers Q2 | 180 1200  
| PM992Q03 | MATH - P2012 Spacers Q3 | 180 1200  
| PM995Q01 | MATH - P2012 Revolving Door Q1 | 224 1393  
| PM995Q02 | MATH - P2012 Revolving Door Q2 | 224 1390  
| PM995Q03 | MATH - P2012 Revolving Door Q3 | 224 1389  
| PM998Q02 | MATH - P2012 Bike Rental Q2 | 201 1288  

41
Student-Level Measures

The second Research Question aimed to investigate whether mathematics self-efficacy, English language proficiency, and student SES can explain DIF. These three variables were selected since they were found to be significant predictors to influence mathematics performance for ELLs (Aikens et al., 2008; Guglielmi, 2012; Stanley, 2005). In addition, these variables served as control variables to reduce student variations. The measures of mathematics self-efficacy, language proficiency, and student SES were introduced in this section.

Mathematics Self-Efficacy

According to Bandura (1997), self-efficacy or perceived ability refers to the confidence an individual has in his or her ability to successfully perform a specific task. Mathematics self-efficacy was included as one of the covariates at student-level since previous studies indicated that mathematics self-efficacy and mathematics achievement were positively related. Students with high mathematics self-efficacy are associated with high mathematics achievement (e.g., Ayotola & Adedeji, 2009; Liu & Koirala, 2009). Conversely, students with low self-efficacy are less likely to regulate their achievement behaviors or be motivated to engage in learning (Klassen & Usher, 2010; Schunk & Pajares, 2009).

In PISA 2012, eight items were used to measure mathematics self-efficacy. These items ask students how confident do they feel about having to do the following tasks: (1) Using a train timetable to work out how long it would take to get from one place to another;
(2) calculating how much cheaper a TV would be after a 30% discount; (3) calculating how many square meters of tiles you need to cover a floor; (4) understanding graphs presented in newspapers; (5) solving an equation like $3x+5= 17$; (6) finding the actual distance between two places on a map; (7) solving an equation like $2(x+3) = (x + 3) (x - 3)$; and (8) calculating the petrol consumption rate of a car. Mathematics self-efficacy were measured in a four-point Likert-type scale (1= Very confident; 2= Confident; 3= Not very confident; and 4= Not at all confident). As all items were inverted for scaling so that the higher score corresponds to higher level of confidence. These items were scaled using IRT scaling methodology (OECD, 2013).

**Language Proficiency**

When using HGLM to detect DIF between ELL and non-ELL students, language proficiency was used as one of the covariates. However, this information is not available in PISA 2012. Reading literacy, a proxy for language proficiency, was used to represent language proficiency since understanding written text is the first form of language proficiency relevant to cognitive functions (Chen, 2010).

According OECD (2013), reading literacy is referred as students’ capacity to understand, use, and reflect on and engage with written texts, in order to achieve one’s goals, develop one’s knowledge and potential, and participate in society. Reading literacy includes a wide range of cognitive competencies, from basic decoding, to knowledge of words, grammar and larger linguistic and textual structures and features, to knowledge about the world.

PISA 2012 assessed reading literacy based on students’ performance on three broad aspect categories including ability to access and retrieve; integrate and interpret; and reflect
and evaluate. These aspects were evaluated on printed and electronic texts which were defined as description, narration, exposition, argumentation, instruction, and transaction. In addition, IRT was used to estimate average scores for reading literacy. IRT identifies patterns of response and uses statistical models to predict the probability of answering an item correctly as a function of the students’ proficiency in answering other questions. Using this method, the performance of a sample of students in reading literacy can be summarized on a simple scale or series of scales, even when students are administered different items (OECD, 2014).

**Socioeconomic Status**

This study utilized the PISA index of economic, social and cultural status (ESCS) to represent student SES. Variables comprising ESCS included home possessions (HOMEPOS), number of books at home (HISEI), and the highest parental education expressed as years of schooling (PARED). The ESCS scores were obtained as component scores for the first principal component with zero being the score of an average OECD student and one being the standard deviation across equally weighted OECD countries. ESCS scores were calculated using the following formula:

\[ \text{ESCS} = \frac{\beta_1 \cdot \text{HOMEPOS} + \beta_2 \cdot \text{HISEI} + \beta_3 \cdot \text{PARED}}{\varepsilon} \]

where \( \beta_1, \beta_2, \) and \( \beta_3 \) are the OECD factor loadings and \( \varepsilon \) is the eigenvalue of the first principal component (OECD, 2014).

**School-Level Measures**

The third Research Question aimed to investigate whether school type and school educational resources can contribute to DIF after controlling for student-level variables. These two variables were selected since they were found to be significant predictors to
influence mathematics performance for ELLs (Freeman et al., 2008; Han et al., 2009; Kim et al., 2010). These variables were used to explain the variation of DIF across schools. The measures of school type and school educational resource were introduced in this section.

**School Type**

In PISA 2012, schools were categorized into public and private according to whether a private entity or a public agency has the ultimate power to make decisions concerning its affairs. The school type (SCHLTYPE) in PISA 2012 has three categories, based on two questions: (1) government-independent private schools controlled by a non-government organization or with a governing board not selected by a government agency which receive less than 50% of their core funding from government agencies, (2) government-dependent private schools controlled by a non-government organization or with a governing board not selected by a government agency which receive more than 50% of their core funding from government agencies, (3) public schools controlled and managed by a public education authority or agency. The first two categories were combined, then the dummy variable of school type was created (0=public, 1=private).

**School Educational Resources**

The PISA 2012 school questionnaire contained 13 items about school educational resources, measuring principals’ perceptions of potential factors hindering instruction at schools (e.g., a lack of qualified science teachers; shortage or inadequacy of science laboratory equipment; shortage or inadequacy of computer software for instruction; Shortage or inadequacy of audio-visual resources). A four-point Likert-type scale was used (1= not at all, 2= very little, 3= to some extent, 4= a lot). As all items were inverted for scaling, higher values on this index indicate more school educational resources (OECD,
The detailed items can be found in Table 3. Responses to the 13 items measuring school educational resources were summed up and rescaled to a Z score to form the predictor of school educational resources.

Table 3. *Items of School Educational Resource Assessment*

<table>
<thead>
<tr>
<th>Question</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is your school’s capacity to provide instruction hindered by any of the following?</td>
<td>A lack of qualified science teachers</td>
</tr>
<tr>
<td></td>
<td>A lack of qualified mathematics teachers</td>
</tr>
<tr>
<td></td>
<td>A lack of qualified (test language) teachers</td>
</tr>
<tr>
<td></td>
<td>A lack of teachers of other subjects</td>
</tr>
<tr>
<td></td>
<td>A lack of laboratory technicians</td>
</tr>
<tr>
<td></td>
<td>A lack of other support personnel</td>
</tr>
<tr>
<td></td>
<td>Shortage or inadequacy of science laboratory equipment</td>
</tr>
<tr>
<td></td>
<td>Shortage or inadequacy of instructional materials</td>
</tr>
<tr>
<td></td>
<td>Shortage or inadequacy of computers for instruction</td>
</tr>
<tr>
<td></td>
<td>Lack or inadequacy of Internet connectivity</td>
</tr>
<tr>
<td></td>
<td>Shortage or inadequacy of computer software for instruction</td>
</tr>
<tr>
<td></td>
<td>Shortage or inadequacy of library materials</td>
</tr>
<tr>
<td></td>
<td>Shortage or inadequacy of audio-visual resources</td>
</tr>
</tbody>
</table>

**Statistical Analyses**

**Mantel-Haenszel Procedure**

The MH procedure used contingency tables to examine if item responses and group membership are independent. The PROC FREQ under software of SAS 9.4 was used to conduct MH procedure (Zhang, 2015). The PISA item dataset will be read in SAS 9.4 in a matrix format whose rows correspond to the students and columns correspond to the items.
The total scores generated by summing item scores across all items were used to match students. The DIF procedures in SAS 9.4 can provide key statistics including MH chi-square, common log-odds ratio and estimated standard error. The MH chi-square statistic is distributed as chi-square with one degree of freedom. Critical values of this statistic are 3.84 at the 0.05 significance level. The MH odds ratio is asymptotically normally distributed. ETS guidelines were used to classify items displaying DIF effects (See Chapter 2).

**Rasch Model**

The Rasch analysis was completed in the Winsteps measurement software, Version 3.9.1. The dichotomous Rasch model was utilized. Prior to the DIF detection, Winsteps can provide some critical statistics to display whether data fits the model. First of all, Winsteps reported both person reliability, person separation, item reliability and item separation. Person reliability is analogous to Cronbach’s alpha reliability while item reliability has no traditional equivalent. Low item reliability indicates a narrow range of item measures or a small sample. Person separation is used to classify people, and item separation is used to verify the item hierarchy (Linacre, 2017).

In addition, Winsteps can produce two different statistics for assessing the model fit, mean square fit statistics (MNSQ) and a standardized transformation of the mean-square to approximate a t-statistic (ZSTD). The MNSQ indicates the size of the randomness. The expected value is 1.0. Values smaller than 1.0 indicate observations are too predictable, and values greater than 1.0 indicate unpredictability. In terms of ZSTD scores, infit ZSTD scores are sensitive to irregular inlying patterns and outfit ZSTD scores are sensitive to unexpected rare extremes. While there is not a specific rule defining the
cutoff, the commonly accepted interpretation is that INFIT and OUTFIT values greater than +2 or less than –2 indicate less compatibility with the model than expected (Linacre, 2017).

After investigating the fit statistics, item difficulty measures for both groups were calculated to examine whether the property of invariance was met. Winsteps outputs for DIF is equivalent to construct a “ruler” based on the persons, and measuring the items on it, first for the one person-group, then for the other person-group. The equivalent procedure is: “(a) The joint run of all person-group classifications, producing anchor values for person abilities and rating (or partial credit) scale structure. (b) The classification A run with person abilities and rating (or partial credit) scale structure anchored at their joint values to produce person-group classification A item difficulties. (c) The classification B run with person abilities and rating (or partial credit) scale structure anchored at their joint values to produce person-group classification B item difficulties. (d) Pairwise item difficulty difference t-tests between the two sets of item difficulties (for person-group classifications A and B)” (Linacre, 2017, p. 574). The DIF contrast is the difference between the DIF sizes and is a log-odds estimate.

Hierarchical Generalized Linear Model (HGLM)

The current study also used HGLM on DIF detection based on Kamata and Binici’s (2003) study. This study discussed that HGLM is equivalent to the Rasch model and showed how the two-level HGLM can be extended to a three-level latent regression model. Specifically, three models were created to answer the three Research Questions. Model 1 (DIF Identification Model) examined each item for DIF between ELLs and non-ELLs. Model 2 (DIF Estimation Model Controlling for Student-Level Variables) further
examined whether student-level variables (mathematics self-efficacy, language proficiency and SES) can explain DIF. Finally, Model 3 (Random Effects DIF Model) included student-level variables and school-level (school type and school educational resource) variables to explain DIF.

**Model 1.** Model 1 including Level-1 and Level-2 models was applied to the 76 items to detect DIF effects. Level-1 model is specified as given by Equation 15 where the log odds of the probability of answering each item correctly versus incorrectly is a linear function of person ability and item difficulty.

\[
\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \sum_{q=1}^{76} (\beta_{qj} \times X_{qij})
\]

$X_{qij}$ is the $q$-th ($q = 1, 2, \ldots, 76$) dummy coded variable that indicates the item $i$ for student $j$. Its value is 1 when $q = i$ and 0 when $q \neq i$. $\beta_{0j}$ is the effect of the reference item and $\beta_{qj}$ is the difference between the $q$-th item and the reference item. The probability of student $j$ getting an item $i$ correct is noted as $p_{ij}$.

Level-2 model was then created by adding the group membership (ELL status) and modeling regression coefficients, $\beta_{qj} (q = 1, 2, \ldots, 76)$ in Equation 15 as given by Equation 16.

\[
\beta_{0j} = \tau_{00} + \mu_{0j}
\]

\[
\beta_{1j} = \tau_{10} + \mu_{11} \times ELL\, Status
\]

\[
\beta_{2j} = \tau_{20} + \mu_{21} \times ELL\, Status
\]

\[\vdots\]

\[
\beta_{76j} = \tau_{760} + \mu_{761} \times ELL\, Status
\]
In Equation 16, coefficients $\tau_{00}$ to $\tau_{760}$ are the DIF coefficients associated with items 1 through 76. ELL students were coded as 1 and non-ELL students were coded as 0. A significant DIF coefficient indicates the existence of DIF for the item under investigation. The exponential term of the DIF coefficient is the odds of answering the corresponding item correctly by the reference versus the focal group.

**Model 2.** Model 2 further examined whether DIF effects decreases or disappears after controlling for student-level variables. As shown in Equation 17, Model 2 was created by adding student-level variables (language proficiency, SES, and mathematics self-efficacy) to Level-2 of Model 1. Level-1 of Model 2 is the same with Level-1 of Model 1 as shown as Equation 15. Level-2 of Model 2 was specified as follows.

\[
\beta_{0j} = \tau_{00} + \mu_{0j} \\
\beta_{1j} = \tau_{10} + \tau_{11} \times ELL\, Status + \tau_{12} \times Language + \tau_{13} \times SES + \tau_{14} \times Selfefficacy \\
\beta_{2j} = \tau_{20} + \tau_{21} \times ELL\, Status + \tau_{22} \times Language + \tau_{23} \times SES + \tau_{24} \times Selfefficacy \\
\vdots \\
\beta_{76j} = \tau_{760} + \tau_{761} \times ELL\, Status + \tau_{762} \times Language + \tau_{763} \times SES + \tau_{764} \times Selfefficacy
\]

In Equation 17, Coefficients $\tau_{11}$ to $\tau_{761}$ are the estimates of DIF after controlling for student-level variables. The exponential term of the DIF coefficients is the odds of answering the corresponding item correctly by the reference versus the focus group. Coefficients $\tau_{12}$ to $\tau_{762}$ are the log odds of answering the corresponding item correctly with one unit of standard deviation (SD) increase in language proficiency. Similarly, $\tau_{13}$
to $\tau_{763}$ and $\tau_{14}$ to $\tau_{764}$ indicate the log odds of answering the corresponding item correctly with one unit of SD increase in SES and mathematics self-efficacy.

**Model 3.** Model 3 examined whether school type and school educational resource contribute to the DIF. Specifically, Model 3 is a three-level DIF identification model including student-level and school-level variables. It investigated whether school type and school educational resources were significant predictors of DIF variations among 162 schools between ELL and non-ELL students.

DIF items those were detected by Model 1 were included in the analysis of Model 3. The Level-1, Level-2, and Level 3 of Model 3 were specified as shown in Equation 18, 19, and 20.

**Level-1:**

$$\text{Log} \left( \frac{p_{ijk}}{1-p_{ijk}} \right) = \beta_{0j} + \sum_{q=1}^{n} (\beta_{qjk} \cdot X_{qijk})$$

**Level-2:**

$$\beta_{0jk} = \tau_{00k} + \mu_{0jk}$$

$$\beta_{1jk} = \tau_{10k} + \tau_{11k} \cdot \text{ELL Status} + \tau_{12k} \cdot \text{Language} + \tau_{13k} \cdot \text{SES} + \tau_{14k} \cdot \text{Self-efficacy}$$

$$\ldots$$

$$\beta_{njk} = \tau_{n0k} + \tau_{n1k} \cdot \text{ELL Status} + \tau_{n1k} \cdot \text{Language} + \tau_{n1k} \cdot \text{SES} + \tau_{n1k} \cdot \text{Self-efficacy}$$

**Level-3:**

$$\tau_{00k} = \pi_{000} + \varepsilon_{00k}$$

$$\tau_{10k} = \pi_{100}$$
\[ \tau_{11k} = \pi_{110} + \pi_{111} \times Schooltype + \pi_{112} \times Resources \]

\[ \tau_{12k} = \pi_{120} \]

\[ \tau_{13k} = \pi_{130} \]

\[ \tau_{14k} = \pi_{140} \]

\[ \tau_{20k} = \pi_{200} \]

\[ \tau_{21k} = \pi_{210} + \pi_{211} \times Schooltype + \pi_{212} \times Resources \]

\[ \tau_{22k} = \pi_{220} \]

\[ \tau_{23k} = \pi_{230} \]

\[ \tau_{24k} = \pi_{240} \]

\[
\vdots
\]

\[ \tau_{100k} = \pi_{1000} \]

\[ \tau_{nk} = \pi_{n10} + \pi_{n11} \times Schooltype + \pi_{n12} \times Resources \]

\[ \tau_{n2k} = \pi_{n20} \]

\[ \tau_{n3k} = \pi_{n30} \]

\[ \tau_{n4k} = \pi_{n40} \]

The subscripts \( n \) and \( k \) indicate \( n \)-th DIF item and \( k \)-th school respectively at Level-3. At Level-2, coefficients of \( \tau_{11k} \) to \( \tau_{n1k} \) are random DIF coefficients that vary from school to school. At Level-3, coefficients \( \pi_{111} \) to \( \pi_{n11} \) indicate how much DIF increases when a school is from public to private, and coefficients \( \pi_{112} \) to \( \pi_{n12} \) denote how much DIF increases when school educational resources increase by one unit of SD. \( \pi_{120} \) to \( \pi_{n20} \) are the fixed regression coefficients for language proficiency. \( \pi_{130} \) to \( \pi_{n30} \) are the fixed regression coefficients for SES. \( \pi_{130} \) to \( \pi_{n30} \) are the fixed regression coefficients for mathematics self-efficacy.
The application of HGLM to detect DIF was conducted with PROC GLIMMIX under the software SAS 9.4. This procedure can fit models to outcome variables that generate a linear model with explanatory variables that account for variations at each level, utilizing variables specified at each level. PROC GLIMMIX can not only estimate model coefficients at each level, but it also predicts the random effects associated with each sampling unit at every level.

In current analysis, gender and race were not controlled for two reasons. First, PISA data was organized in wide format data. It is necessary to run PROC GLIMMIX using long format data. The transformation from wide format to long format data resulted in 378,328 observation (4,978 * 76) in the dataset to be analyzed. The inclusions of two additional covariates will add complexity to the model. Second, ELL status could be highly correlated with race since most ELL students could be in the minority group. The inclusion of race could result in multicollinearity in the model (Yu, Jiang, & Land, 2015).

In addition, sampling weights were not included. According to Linacre (2017), unweighted data is preferable for calibrating the items since each observation is modeled to contribute one unit of independent statistical information. The effect of weighting is to distort the distribution of independent statistical information in the data.
CHAPTER 4. RESULTS

This chapter is divided into four sections. The first three sections consist of the results of the relevant statistical analyses including Mantel-Haenszel procedure, Rasch analysis, and Hierarchical Generalized Linear Model (HGLM). In the fourth section, results of three DIF detection methods are compared and discussed in more detail.

Mantel-Haenszel (MH) Procedure

The MH procedure was the first approach in this study to examine DIF effects between ELL and non-ELL students. The PROC FREQ under SAS 9.4 was used to conduct MH procedure. Table 4 displays the summary results from the MH procedure. As mentioned in Chapter 2, $\Delta \text{OR}_{\text{MH}}$ is applied as a measure of the degree of DIF referred to as DIF effect size. Educational Testing Service classified the DIF effect size into three categories in practice: Class A denotes negligible magnitudes of DIF when $|\Delta \text{OR}_{\text{MH}}| < 1.00$; Class B denotes moderate magnitudes of DIF when $1.00 \leq |\Delta \text{OR}_{\text{MH}}| < 1.50$; and Class C denotes large magnitudes of DIF when $|\Delta \text{OR}_{\text{MH}}| \geq 1.50$ (Dorans & Holland, 1993).

Table 4 shows the results of DIF effects using MH procedure. Among the 76 items, 60 items with negligible DIF were categorized into Class A. Six items with moderate values of $\Delta \text{OR}_{\text{MH}}$ were categorized into Class B. Ten items with large values of $\Delta \text{OR}_{\text{MH}}$ were categorized into Class C. In Class B, all six items were in favor of non-ELL students. For example, Item 8 with the odds ratio of .55 indicated that ELL students are 45% less likely to answer this item correctly. In Class C, all ten items were in favor of non-ELL students. For example, Item 16 with the odds ratio of .77 indicated that ELL students are 23% less likely to answer this item correctly.
### Table 4. Summary results from the MH Procedure to identify DIF Effects

<table>
<thead>
<tr>
<th>Item Number</th>
<th>PISA Variable Name</th>
<th>MH Chi-Square</th>
<th>Odds-Ratio</th>
<th>DIF Effect Size</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>PM192Q01T</td>
<td>12.76 **</td>
<td>.55</td>
<td>1.42</td>
<td>B</td>
</tr>
<tr>
<td>16</td>
<td>PM420Q01T</td>
<td>21.23 **</td>
<td>.50</td>
<td>1.63</td>
<td>C</td>
</tr>
<tr>
<td>40</td>
<td>PM909Q01</td>
<td>14.84 **</td>
<td>.39</td>
<td>2.24</td>
<td>C</td>
</tr>
<tr>
<td>41</td>
<td>PM909Q02</td>
<td>2.02 **</td>
<td>.51</td>
<td>1.59</td>
<td>C</td>
</tr>
<tr>
<td>42</td>
<td>PM909Q03</td>
<td>31.67 **</td>
<td>.34</td>
<td>2.57</td>
<td>C</td>
</tr>
<tr>
<td>43</td>
<td>PM915Q01</td>
<td>13.10 **</td>
<td>.55</td>
<td>1.42</td>
<td>B</td>
</tr>
<tr>
<td>46</td>
<td>PM918Q02</td>
<td>23.31 **</td>
<td>.47</td>
<td>1.80</td>
<td>C</td>
</tr>
<tr>
<td>47</td>
<td>PM918Q05</td>
<td>7.78 **</td>
<td>.64</td>
<td>1.05</td>
<td>B</td>
</tr>
<tr>
<td>49</td>
<td>PM919Q02</td>
<td>1.10 **</td>
<td>.62</td>
<td>1.12</td>
<td>B</td>
</tr>
<tr>
<td>56</td>
<td>PM949Q01T</td>
<td>14.22 **</td>
<td>.56</td>
<td>1.35</td>
<td>B</td>
</tr>
<tr>
<td>57</td>
<td>PM949Q02T</td>
<td>14.00 **</td>
<td>.50</td>
<td>1.61</td>
<td>C</td>
</tr>
<tr>
<td>61</td>
<td>PM954Q02</td>
<td>16.74 **</td>
<td>.50</td>
<td>1.59</td>
<td>C</td>
</tr>
<tr>
<td>63</td>
<td>PM955Q01</td>
<td>16.90 **</td>
<td>.51</td>
<td>1.57</td>
<td>C</td>
</tr>
<tr>
<td>64</td>
<td>PM955Q02</td>
<td>11.75 **</td>
<td>.48</td>
<td>1.73</td>
<td>C</td>
</tr>
<tr>
<td>68</td>
<td>PM982Q04</td>
<td>9.52 **</td>
<td>.60</td>
<td>1.20</td>
<td>B</td>
</tr>
<tr>
<td>73</td>
<td>PM995Q02</td>
<td>5.11 *</td>
<td>.18</td>
<td>4.05</td>
<td>C</td>
</tr>
</tbody>
</table>

Note: *p≤.05 **, p≤.01

### Rasch Model

Winsteps measurement software was used to conduct the Rasch analysis. Before delving into the DIF detection, it is necessary to examine the psychometric properties of PISA 2012 mathematics assessment. Specifically, overall model data fit statistics, item fit statistics, and person and item distributions were examined. These statistics revealed how well the assessment functions for the whole sample.
Model Data Fit Statistics

The model data fit statistics can be found in Table 5. First, Winsteps produced reliability and separation statistics for both persons and items. The values of person reliability and item reliability range from zero to one and can be interpreted as the Cronbach’s alpha (Linacre, 2012). Person separation was used to classify people. Low person separation (< 2.00) with low person reliability (< .80) implies that the assessment may not be sensitive enough to distinguish between high and low performers. Item separation is used to verify the item hierarchy. Low item separation (< 3.00) with low item reliability (< .90) implies that person sample is not large enough to confirm the item difficulty hierarchy of the assessment. Person reliability and separation were .78 and 1.89 respectively. Also, the item reliability and separation were 1.00 and 21.41 respectively. PISA 2012 assessment used a matrix-sampling design and each student only completed a subset of total assessment items. As a result, the missing data can result in the low person reliability and separation statistics. However, person and item measures were not biased (Linacre, 2012).

In addition to reliability and separation statistics, mean square fit statistics (MNSQ) and a standardized transformation of the mean-square to approximate a t-statistic (ZSTD) were used to assess the model data fit. The expected value of MNSQ is 1.0. Values less than 1.0 imply observations are too predictable while values greater than 1.0 indicate unpredictability (Linacre, 2017). In general, MNSQ near 1.0 indicate little distortion of the measurement system, regardless of the ZSTD (Linacre, 2002). ZSTD is reported as z-scores to test the hypothesis “Do the data fit the model perfectly?”. The expected value of ZSTD is zero. Less than zero indicates too predictable while more than zero indicates lack
of predictability. Generally, the ZSTD within the range of -1.9 to 1.9 indicate the instrument indicates a reasonable predictability (Linacre, 2002). Table 5 showed that infit MNSQ, outfit MNSQ, infit ZSTD and outfit ZSTD could meet this requirement.

Table 5. Model Data Fit Statistics

<table>
<thead>
<tr>
<th>Measure</th>
<th>Infit MNSQ</th>
<th>Outfit MNSQ</th>
<th>Infit ZSTD</th>
<th>Outfit ZSTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Reliability = .78; Separation = 1.89)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M$</td>
<td>-.33</td>
<td>1.00</td>
<td>.99</td>
<td>.00</td>
</tr>
<tr>
<td>$SD$</td>
<td>1.49</td>
<td>.29</td>
<td>.73</td>
<td>.90</td>
</tr>
<tr>
<td>Item</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Reliability = 1.00; Separation = 21.41)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M$</td>
<td>.00</td>
<td>.99</td>
<td>1.01</td>
<td>-.20</td>
</tr>
<tr>
<td>$SD$</td>
<td>1.67</td>
<td>.12</td>
<td>.32</td>
<td>3.80</td>
</tr>
</tbody>
</table>

Item Fit Statistics

Two item fit statistics including point measure correlation and infit MNSQ are the most commonly used indices to examine how well the response data meets the expectation of the Rasch model (Wilson, 2005). Linacre (2009) suggests the point measure correlation should be investigated before checking with other item fit statistics. The point measure correlation is a point-biserial correlation between responses and person raw scores or measures. This is an important diagnostic indicator of data miscoding or item miskeying. Negative correlations indicate that the responses to the item contradict the latent variable defined by the consensus of the items. The items with negative correlations may need to
be omitted or rescored in the opposite direction. According to Table 6, all the point measure correlations were found to be positive, indicating the coding scheme was correct. In addition, Wilson (2005) suggested that infit MNSQ between .75 and 1.33 indicate reasonably good model data fit. Table 6 shows that the infit MNSQ ranging from .78 to 1.31 met the requirement. Thus, the following DIF analyses included all the 76 items in the PISA 2012 mathematics assessment.

Table 6. *Item Measure and Fit Statistics*

<table>
<thead>
<tr>
<th>Item Number</th>
<th>Measure</th>
<th>Point Measure Correlation</th>
<th>Infit MNSQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.39</td>
<td>.51</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>3.36</td>
<td>.40</td>
<td>.91</td>
</tr>
<tr>
<td>3</td>
<td>3.08</td>
<td>.36</td>
<td>1.01</td>
</tr>
<tr>
<td>4</td>
<td>-1.74</td>
<td>.35</td>
<td>1.12</td>
</tr>
<tr>
<td>5</td>
<td>.56</td>
<td>.53</td>
<td>.97</td>
</tr>
<tr>
<td>6</td>
<td>-1.49</td>
<td>.51</td>
<td>.91</td>
</tr>
<tr>
<td>7</td>
<td>-.65</td>
<td>.40</td>
<td>1.15</td>
</tr>
<tr>
<td>8</td>
<td>.30</td>
<td>.54</td>
<td>.94</td>
</tr>
<tr>
<td>9</td>
<td>-.35</td>
<td>.43</td>
<td>1.1</td>
</tr>
<tr>
<td>10</td>
<td>-.21</td>
<td>.31</td>
<td>1.3</td>
</tr>
<tr>
<td>11</td>
<td>1.72</td>
<td>.54</td>
<td>.87</td>
</tr>
<tr>
<td>12</td>
<td>2.59</td>
<td>.55</td>
<td>.78</td>
</tr>
<tr>
<td>13</td>
<td>.55</td>
<td>.50</td>
<td>1.00</td>
</tr>
<tr>
<td>14</td>
<td>-.23</td>
<td>.57</td>
<td>.90</td>
</tr>
<tr>
<td>15</td>
<td>-.29</td>
<td>.46</td>
<td>1.07</td>
</tr>
<tr>
<td>16</td>
<td>-.83</td>
<td>.52</td>
<td>.94</td>
</tr>
<tr>
<td>17</td>
<td>-1.85</td>
<td>.33</td>
<td>1.10</td>
</tr>
<tr>
<td>18</td>
<td>.50</td>
<td>.63</td>
<td>.81</td>
</tr>
<tr>
<td>19</td>
<td>-1.80</td>
<td>.50</td>
<td>.89</td>
</tr>
</tbody>
</table>
Table 6 (continued)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>3.36</td>
<td>.37</td>
<td>.93</td>
</tr>
<tr>
<td>21</td>
<td>-1.20</td>
<td>.48</td>
<td>.99</td>
</tr>
<tr>
<td>22</td>
<td>2.04</td>
<td>.56</td>
<td>.80</td>
</tr>
<tr>
<td>23</td>
<td>-1.50</td>
<td>.33</td>
<td>1.16</td>
</tr>
<tr>
<td>24</td>
<td>-.47</td>
<td>.53</td>
<td>.94</td>
</tr>
<tr>
<td>25</td>
<td>-1.05</td>
<td>.46</td>
<td>.99</td>
</tr>
<tr>
<td>26</td>
<td>-.74</td>
<td>.48</td>
<td>1.03</td>
</tr>
<tr>
<td>27</td>
<td>.03</td>
<td>.39</td>
<td>1.18</td>
</tr>
<tr>
<td>28</td>
<td>.13</td>
<td>.43</td>
<td>1.13</td>
</tr>
<tr>
<td>29</td>
<td>-.14</td>
<td>.52</td>
<td>.96</td>
</tr>
<tr>
<td>30</td>
<td>.12</td>
<td>.44</td>
<td>1.09</td>
</tr>
<tr>
<td>31</td>
<td>-2.11</td>
<td>.26</td>
<td>1.16</td>
</tr>
<tr>
<td>32</td>
<td>1.09</td>
<td>.59</td>
<td>.85</td>
</tr>
<tr>
<td>33</td>
<td>.81</td>
<td>.56</td>
<td>.91</td>
</tr>
<tr>
<td>34</td>
<td>-.87</td>
<td>.39</td>
<td>1.14</td>
</tr>
<tr>
<td>35</td>
<td>1.39</td>
<td>.40</td>
<td>1.12</td>
</tr>
<tr>
<td>36</td>
<td>.91</td>
<td>.58</td>
<td>.88</td>
</tr>
<tr>
<td>37</td>
<td>-2.13</td>
<td>.45</td>
<td>.98</td>
</tr>
<tr>
<td>38</td>
<td>.29</td>
<td>.63</td>
<td>.81</td>
</tr>
<tr>
<td>39</td>
<td>-.36</td>
<td>.51</td>
<td>.97</td>
</tr>
<tr>
<td>40</td>
<td>-3.50</td>
<td>.30</td>
<td>.94</td>
</tr>
<tr>
<td>41</td>
<td>-.69</td>
<td>.49</td>
<td>.99</td>
</tr>
<tr>
<td>42</td>
<td>.89</td>
<td>.62</td>
<td>.81</td>
</tr>
<tr>
<td>43</td>
<td>.00</td>
<td>.46</td>
<td>1.05</td>
</tr>
<tr>
<td>44</td>
<td>-.48</td>
<td>.55</td>
<td>.92</td>
</tr>
<tr>
<td>45</td>
<td>-3.43</td>
<td>.27</td>
<td>1.00</td>
</tr>
<tr>
<td>46</td>
<td>-1.95</td>
<td>.39</td>
<td>1.03</td>
</tr>
<tr>
<td>47</td>
<td>-1.90</td>
<td>.41</td>
<td>1.02</td>
</tr>
<tr>
<td>48</td>
<td>-2.19</td>
<td>.45</td>
<td>.95</td>
</tr>
<tr>
<td>49</td>
<td>.23</td>
<td>.45</td>
<td>1.13</td>
</tr>
</tbody>
</table>
Table 6 (continued)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>-.30</td>
<td>.53</td>
<td>.97</td>
</tr>
<tr>
<td>51</td>
<td>.07</td>
<td>.40</td>
<td>1.19</td>
</tr>
<tr>
<td>52</td>
<td>2.48</td>
<td>.47</td>
<td>.94</td>
</tr>
<tr>
<td>53</td>
<td>-.36</td>
<td>.54</td>
<td>.96</td>
</tr>
<tr>
<td>54</td>
<td>-.44</td>
<td>.39</td>
<td>1.24</td>
</tr>
<tr>
<td>55</td>
<td>4.54</td>
<td>.36</td>
<td>.88</td>
</tr>
<tr>
<td>56</td>
<td>-.82</td>
<td>.57</td>
<td>.87</td>
</tr>
<tr>
<td>57</td>
<td>.93</td>
<td>.55</td>
<td>.90</td>
</tr>
<tr>
<td>58</td>
<td>-.28</td>
<td>.56</td>
<td>.92</td>
</tr>
<tr>
<td>59</td>
<td>-.03</td>
<td>.62</td>
<td>.83</td>
</tr>
<tr>
<td>60</td>
<td>-1.55</td>
<td>.55</td>
<td>.86</td>
</tr>
<tr>
<td>61</td>
<td>.72</td>
<td>.53</td>
<td>.99</td>
</tr>
<tr>
<td>62</td>
<td>1.33</td>
<td>.59</td>
<td>.86</td>
</tr>
<tr>
<td>63</td>
<td>-1.71</td>
<td>.38</td>
<td>1.01</td>
</tr>
<tr>
<td>64</td>
<td>1.57</td>
<td>.49</td>
<td>.99</td>
</tr>
<tr>
<td>65</td>
<td>-2.82</td>
<td>.29</td>
<td>1.06</td>
</tr>
<tr>
<td>66</td>
<td>.67</td>
<td>.36</td>
<td>1.23</td>
</tr>
<tr>
<td>67</td>
<td>-.89</td>
<td>.43</td>
<td>1.07</td>
</tr>
<tr>
<td>68</td>
<td>.05</td>
<td>.51</td>
<td>.99</td>
</tr>
<tr>
<td>69</td>
<td>-1.87</td>
<td>.43</td>
<td>1.00</td>
</tr>
<tr>
<td>70</td>
<td>2.05</td>
<td>.49</td>
<td>.91</td>
</tr>
<tr>
<td>71</td>
<td>3.38</td>
<td>.45</td>
<td>.83</td>
</tr>
<tr>
<td>72</td>
<td>-.13</td>
<td>.62</td>
<td>.82</td>
</tr>
<tr>
<td>73</td>
<td>4.68</td>
<td>.30</td>
<td>.95</td>
</tr>
<tr>
<td>74</td>
<td>.04</td>
<td>.57</td>
<td>.91</td>
</tr>
<tr>
<td>75</td>
<td>-2.20</td>
<td>.37</td>
<td>1.00</td>
</tr>
<tr>
<td>76</td>
<td>.71</td>
<td>.26</td>
<td>1.31</td>
</tr>
</tbody>
</table>
Person and Item Distributions

In Rasch analysis, item difficulty and person ability measures were calibrated to be on the same logic metric. The Wright map (See Figure 1) can provide distribution for both item difficulty and person ability measures on a single line of logit scale to facilitate the graphical representation of the relationships. In this map, the ability measures were shown on the left side and the item difficulty locations were shown on the right. Person ability and item difficulty increase as one moves towards the top of the figure. Ideally, item difficulty distribution should cover the span of student ability distribution so persons at different proficient levels can be accurately measured. In Figure 1, mathematics items cover the person ability distribution quite well. It indicates there are enough items providing accurate ability estimates across the whole range of students.

DIF between ELLs and non-ELLs

DIF can be examined within the Rasch model by comparing item difficulties between groups. Table 7 reports the difficulty estimates for both groups and their difficulty contrast. The t statistics was calculated using Equation 7 (See Chapter 2). Then the corresponding p value was obtained using the specified degrees of freedom. If the difficulty measures are significantly different between ELL and non-ELL students for the same item, this item was considered to have a DIF issue. Difficulty contrast can be considered as the effect size in logits. A positive difficulty contrast indicates the item is more difficult for non-ELL students, and a negative difficulty contrast implies the item is more difficult for ELL students. According to Table 7, 11 items have been found to have DIF effects. Among the 11 items, ten items were more difficult for ELL students and one item was more difficult for non-ELL students (Item 51). According to de Ayala (2009), items with
difficulty contrast above .30 are considered as being noticeable. Thus these 11 items were found to display practically significant DIF effects.

Figure 1. Wright Map
### Table 7. Summary of Results from Rasch Analysis to Identify DIF Effects

<table>
<thead>
<tr>
<th>Item Number</th>
<th>Difficulty Measures</th>
<th>Difficulty Contrast</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-ELLS</td>
<td>ELLs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>.25</td>
<td>.66</td>
<td>-.40 *</td>
<td>-1.98</td>
</tr>
<tr>
<td>16</td>
<td>-.88</td>
<td>-.48</td>
<td>-.40 *</td>
<td>-2.28</td>
</tr>
<tr>
<td>40</td>
<td>-3.62</td>
<td>-3.08</td>
<td>-.54 *</td>
<td>-2.23</td>
</tr>
<tr>
<td>42</td>
<td>.82</td>
<td>1.56</td>
<td>-.74 **</td>
<td>-3.08</td>
</tr>
<tr>
<td>43</td>
<td>-.06</td>
<td>.42</td>
<td>-.48 *</td>
<td>-2.45</td>
</tr>
<tr>
<td>46</td>
<td>-2.04</td>
<td>-1.50</td>
<td>-.53 **</td>
<td>-2.98</td>
</tr>
<tr>
<td>49</td>
<td>.19</td>
<td>.49</td>
<td>-.30 *</td>
<td>-1.67</td>
</tr>
<tr>
<td>51</td>
<td>.14</td>
<td>-.41</td>
<td>.54 **</td>
<td>3.16</td>
</tr>
<tr>
<td>61</td>
<td>.65</td>
<td>1.21</td>
<td>-.56 *</td>
<td>-2.76</td>
</tr>
<tr>
<td>68</td>
<td>.01</td>
<td>.36</td>
<td>-.35 *</td>
<td>-1.80</td>
</tr>
<tr>
<td>73</td>
<td>4.58</td>
<td>5.96</td>
<td>-1.38 *</td>
<td>-1.74</td>
</tr>
</tbody>
</table>

Note: *p≤.05 **, p≤.01

### Hierarchical Generalized Linear Model (HGLM)

The PROC GLIMMIX under SAS 9.4 was used to conduct HGLM by generating three models. Model 1 (DIF Identification Model) examined each item for DIF between ELL and non-ELL students. Model 2 (DIF Estimation Model Controlling for Student-Level Variables) further examined whether student-level variables (language proficiency, SES, and mathematics self-efficacy) can explain DIF. Finally, Model 3 (Random Effects DIF Model) included student-level variables and school-level variables to explain DIF.

**Model 1**

Table 8 summarizes the results of items with DIF effects using HGLM. Estimates are the DIF coefficients in Model 1. Estimates were exponentiated to obtain the DIF odds ratios. Then odds ratios were transformed to DIF effect size (Δ OR_{MH}). Similar with MH
procedure, effect sizes were categorized into three classes. Among the 76 items, 66 items with negligible DIF were categorized into Class A. Five items with moderate values of $\Delta OR_{MH}$ were categorized into Class B. Six items with large values of $\Delta OR_{MH}$ were categorized into Class C. In Classes B and C, all the items were in favor of non-ELLs. For example, Item 8 with the DIF odds ratio of .62 indicated that ELL students are 38% less likely to answer this item correctly.

Table 8. Summary of HGLM Model 1

<table>
<thead>
<tr>
<th>Item Number</th>
<th>Estimates</th>
<th>Odds Ratio</th>
<th>DIF Effect size</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>-.47 *</td>
<td>.62</td>
<td>1.12</td>
<td>B</td>
</tr>
<tr>
<td>16</td>
<td>-.52 **</td>
<td>.60</td>
<td>1.21</td>
<td>B</td>
</tr>
<tr>
<td>40</td>
<td>-.70 **</td>
<td>.50</td>
<td>1.65</td>
<td>C</td>
</tr>
<tr>
<td>42</td>
<td>-.87 **</td>
<td>.42</td>
<td>2.04</td>
<td>C</td>
</tr>
<tr>
<td>43</td>
<td>-.56 **</td>
<td>.57</td>
<td>1.32</td>
<td>B</td>
</tr>
<tr>
<td>46</td>
<td>-.64 **</td>
<td>.53</td>
<td>1.51</td>
<td>C</td>
</tr>
<tr>
<td>49</td>
<td>-.40 **</td>
<td>.65</td>
<td>1.01</td>
<td>B</td>
</tr>
<tr>
<td>61</td>
<td>-.64 **</td>
<td>.53</td>
<td>1.50</td>
<td>C</td>
</tr>
<tr>
<td>68</td>
<td>-.45 *</td>
<td>.64</td>
<td>1.05</td>
<td>B</td>
</tr>
<tr>
<td>73</td>
<td>-1.60 *</td>
<td>.20</td>
<td>3.77</td>
<td>C</td>
</tr>
</tbody>
</table>

Note: * $p \leq .05$ ** $p \leq .01$

Model 2

In Model 2, student-level variables including mathematics self-efficacy, language proficiency, and SES were included to identify whether they are the sources of DIF between ELLs and non-ELLs. If the number of items showing DIF effects and their effect sizes decrease after controlling for the student-level variables, these three variables are the
sources of DIF at the student-level. Descriptive statistics of student-level variables can be found in the Table 9.

Table 9. *Descriptive Statistics of Student-Level Variables*

<table>
<thead>
<tr>
<th>Variable</th>
<th>ELLS Mean</th>
<th>ELLS SD</th>
<th>Non-ELLs Mean</th>
<th>Non-ELLs SD</th>
<th>Combined Mean</th>
<th>Combined SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematics Self-Efficacy</td>
<td>.06</td>
<td>.99</td>
<td>.16</td>
<td>.99</td>
<td>.14</td>
<td>1.00</td>
</tr>
<tr>
<td>Language Proficiency</td>
<td>477.70</td>
<td>91.92</td>
<td>503.30</td>
<td>89.75</td>
<td>497.47</td>
<td>91.38</td>
</tr>
<tr>
<td>SES</td>
<td>-.59</td>
<td>1.03</td>
<td>.32</td>
<td>.90</td>
<td>.19</td>
<td>.97</td>
</tr>
</tbody>
</table>

Table 10 displays the results from Model 2 to identify DIF effects controlling for student-level variables. Estimates are the DIF coefficients in Model 2. Odds ratios of student-level variables are the exponential terms for the regression coefficients of mathematics self-efficacy, language proficiency, and SES, which indicate the odds of getting each item correct associated one standard deviation (SD) increase in those three variables.

Mathematics self-efficacy was a significant predictor on six of ten DIF items. For example, Item 8 with the odds ratio of 1.58 indicated that students with one SD increase of mathematics self-efficacy were 1.58 times more likely to answer this item correctly. Language proficiency was a significant predictor for all ten items even its effect was minimal. For example, Item 73 with the odds ratio of 1.02 indicated that students with one SD increase of language proficiency were 1.02 times more likely to answer this item correctly. SES was a significant predictor on four of ten DIF items. For example, Item 61 with the odds ratio of 1.18 indicated that students with one SD increase of SES were 1.18
times more likely to answer this item correctly. After controlling for student-level variables, seven items still displayed DIF effects while the remaining three items no longer showed DIF effects. Besides, all the DIF effect sizes decreased after controlling for student-level variables.

Table 10. Summary of Results from HGLM Model 2

<table>
<thead>
<tr>
<th>Item Number</th>
<th>DIF Estimates</th>
<th>DIF Odds Ratio</th>
<th>DIF Effect Size</th>
<th>Mathematics Self-Efficacy</th>
<th>Language Proficiency</th>
<th>SES</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>-.46 *</td>
<td>.63</td>
<td>1.08</td>
<td>1.58 **</td>
<td>1.01 **</td>
<td>.86</td>
</tr>
<tr>
<td>16</td>
<td>-.36</td>
<td>.70</td>
<td>.85</td>
<td>1.06</td>
<td>1.01 **</td>
<td>1.09</td>
</tr>
<tr>
<td>40</td>
<td>-.67 *</td>
<td>.51</td>
<td>1.58</td>
<td>1.02</td>
<td>1.01 **</td>
<td>.92</td>
</tr>
<tr>
<td>42</td>
<td>-.57 *</td>
<td>.56</td>
<td>1.34</td>
<td>1.03 **</td>
<td>1.01 **</td>
<td>1.21</td>
</tr>
<tr>
<td>43</td>
<td>-.46 *</td>
<td>.63</td>
<td>1.09</td>
<td>1.02 **</td>
<td>1.01 **</td>
<td>1.08 *</td>
</tr>
<tr>
<td>46</td>
<td>-.59 **</td>
<td>.55</td>
<td>1.39</td>
<td>1.00</td>
<td>1.01 **</td>
<td>.99</td>
</tr>
<tr>
<td>49</td>
<td>-.42 *</td>
<td>.66</td>
<td>.99</td>
<td>1.02</td>
<td>1.01 **</td>
<td>.91</td>
</tr>
<tr>
<td>61</td>
<td>-.50 *</td>
<td>.61</td>
<td>1.18</td>
<td>1.02 **</td>
<td>1.01 **</td>
<td>1.18 *</td>
</tr>
<tr>
<td>68</td>
<td>-.28</td>
<td>.75</td>
<td>.66</td>
<td>1.18 **</td>
<td>1.01 **</td>
<td>1.14</td>
</tr>
<tr>
<td>73</td>
<td>-1.46</td>
<td>.23</td>
<td>3.43</td>
<td>1.18 **</td>
<td>1.02 **</td>
<td>1.11</td>
</tr>
</tbody>
</table>

Note: *p≤.05, **p≤.01

Model 3

The three-level model was implemented for the ten items displaying DIF effects. In this model, DIF effects were modeled to vary across 157 schools after controlling for the three student-level variables. Items with significant DIF variations across the schools were identified. Descriptive statistics of school-level variables can be found in the Table 11.
Table 11. Descriptive Statistics of School-Level Variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>School Type (Public = 0, Private = 1)</td>
<td>1.09</td>
<td>.28</td>
</tr>
<tr>
<td>School Educational Resource</td>
<td>.36</td>
<td>1.08</td>
</tr>
</tbody>
</table>

Table 12 displays the results from Model 3 to identify DIF effects controlling for student and school-level variables. Only odds ratios of school-level variables were displayed. Three out of ten items were found to show significant DIF effects as both student and school-level variables were controlled (Item 8, 40, and 46). Nevertheless, school educational resources were not a significant predictor for these ten items. School type was found to be a significant predictor for Item 40 and 46. For Item 40, students in private schools are 1.17 times more likely to answer this item correctly. For Item 46, students in private schools are 1.64 times more likely to answer this item correctly.

Table 12. Summary of Results from HGLM Model 3

<table>
<thead>
<tr>
<th>Item Number</th>
<th>DIF Estimates</th>
<th>DIF Odds Ratio</th>
<th>DIF Effect Size</th>
<th>School Type</th>
<th>School Educational Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>-.66 *</td>
<td>.52</td>
<td>1.56</td>
<td>1.06</td>
<td>.98</td>
</tr>
<tr>
<td>16</td>
<td>-.45</td>
<td>.64</td>
<td>1.06</td>
<td>.70</td>
<td>.94</td>
</tr>
<tr>
<td>40</td>
<td>-.60 *</td>
<td>.44</td>
<td>1.41</td>
<td>1.17 *</td>
<td>.98</td>
</tr>
<tr>
<td>42</td>
<td>-.56</td>
<td>.57</td>
<td>1.31</td>
<td>.76</td>
<td>1.07</td>
</tr>
<tr>
<td>43</td>
<td>-.46</td>
<td>.63</td>
<td>1.09</td>
<td>1.08</td>
<td>.97</td>
</tr>
<tr>
<td>46</td>
<td>-.78 *</td>
<td>.46</td>
<td>1.84</td>
<td>1.64 *</td>
<td>.94</td>
</tr>
<tr>
<td>49</td>
<td>-.17</td>
<td>.84</td>
<td>.40</td>
<td>.97</td>
<td>.97</td>
</tr>
</tbody>
</table>
Table 12 (Continued)

<table>
<thead>
<tr>
<th></th>
<th>Mantel-Haenszel Procedure</th>
<th>Rasch Analysis</th>
<th>HGLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>61</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>68</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>73</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note: *p≤.05 **, p≤.01

Consistency of Three DIF Detection Methods

Table 13 summarizes the DIF items identified by MH procedure, Rasch analysis, and HGLM. 17 items were identified with DIF effects in one of those three methods. Among those 17 items, ten items (Items 8, 16, 40, 43, 46, 49, 61, 68, and 73) were identified with DIF effects by all the three methods. Rasch analysis and HGLM approaches showed consistent results except Item 51 which was found to be in favor of ELLs. Besides, MH approach discovered six items with DIF effects that were not identified by the other two methods.

Table 13. Summary of DIF Items Identified by Three Methods

<table>
<thead>
<tr>
<th>Item Number</th>
<th>Mantel-Haenszel Procedure</th>
<th>Rasch Analysis</th>
<th>HGLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>16</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>40</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>41</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>43</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>46</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>47</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>51</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 13 (Continued)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>57</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>63</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>73</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
CHAPTER 5. DISCUSSIONS

ELLs have been considered as one of the fastest growing groups among the school-aged population in the United States. ELLs have historically left behind their English proficient peers in all content areas, especially in the subjects that are high in language demand such as mathematics (Abedi, 2002; Abedi, 2008). For example, the mathematics achievement gap between ELLs and non-ELLs has been identified from the results of National Assessment of Educational Progress (NCES, 2016; Martiniello, 2009). Most of the existing studies on the investigation of achievement gap between ELLs and non-ELLs relied on statistics such as means, variance, and effect sizes (e.g., Abedi, 2002, Beal, Adams, & Cohen, 2010; Fry, 2007). These studies failed to identify whether items in assessments can cause the gaps on the overall measures while they are useful to demonstrate the achievement gap. As a result, item level analysis of mathematics assessments is supposed to be added to fill out the literature gap.

Using the U.S. sample of PISA 2012, the purpose of current study was to understand the nature and potential sources of the gaps in mathematics achievement between ELLs and non-ELLs. The nature of achievement gap was examined using three DIF detection methodologies including MH procedure, Rasch analysis, and HGLM. HGLM can incorporate student and school-level variables to examine the potential sources of DIF. At the student-level, sources of DIF were investigated through the students’ variations in mathematics self-efficacy, SES, and language proficiency. At the school-level, school type and school educational resources were investigated as potential sources of DIF after controlling for the student variables.
Results of large-scale international assessments such as the PISA, can be considered as a benchmark of quality for a national education system. It is necessary to evaluate its validity of assessment for all subgroups (e.g., ELLs). The current investigation was a construct validity study addressing whether construct-irrelevant language factors may be present in PISA 2012 mathematics assessment for ELLs.

Above all, this chapter began with summarizing findings corresponding to each research question and making conclusions. Then implications of the work were presented. Finally, limitations and future research were discussed.

**Summary of Findings**

1) Do items from PISA 2012 mathematics assessment exhibit DIF between ELLs and non-ELLs?

The first research question in this study asked whether 76 dichotomous items from PISA 2012 mathematics assessment exhibit DIF between ELLs and non-ELLs. Three DIF detection methods including MH procedure, Rasch analysis, and HGLM were used. Ten items (Items 8, 16, 40, 43, 46, 49, 61, 68, and 73) were identified with DIF effects by all the three methods. Rasch analysis and HGLM approaches showed consistent results except Item 55 which was found to be in favor of ELLs. In addition, the MH approach discovered six items with DIF effects which were not identified by the other two methods.

As mentioned earlier, some previous studies were conducted to detect DIF in PISA assessment while most of them investigated DIF for gender and translation equivalence (e.g., Huang, 2010; Le, 2009; Qian, 2011). Only one study was concerned with DIF on science assessment items between ELLs and non-ELLs (e.g., Shirley, 2014). Despite the limited studies on large-scale international assessments, several researchers examined the
DIF effects between ELLs and non-ELLs on some districtwide or statewide mathematics tests. For example, Loughran (2014) examined mathematics items for evidence of DIF against ELLs using logistic regression in a Midwestern statewide test. Results revealed that more DIF items against ELLs were found in the test for Grade 8 than Grade 4 since linguistic demand is higher for Grade 8 test. Besides, Martiniello (2008) detected the DIF effects between ELLs and non-ELLs on the Massachusetts Comprehensive Assessment System (MCAS) Grade 4 mathematics test. Six out of ten items were found to display DIF effects, which were in favor of non-ELLs. Similarly, Wolf and Leon (2009) examined 542 items from 11 assessments at Grades 4, 5, 7, and 8 from three different states. They found mathematics items with low difficulty estimates but high ratings of language complexity are more likely to show DIF effects against ELLs. Those three studies mentioned above agreed that DIF is the detection of an item that performs differently between two groups. Further analysis should be done to explore the potential sources of DIF effects between ELLs and non-ELLs. Therefore, the second and third research questions led this study to examine the potential sources of DIF effects.

2) If DIF is detected, can English language proficiency and other student characteristics (e.g., student SES, mathematics self-efficacy) explain DIF? That is, after controlling for these three student variables, whether DIF between ELLs and non-ELLs changes was examined.

HGLM Model 2 was used to examine whether DIF effects decreases or disappears after controlling for student-level variables. Among the ten DIF items that were identified by HGLM, seven items still displayed DIF effects after controlling for student-level variables. The rest of three items no longer showed DIF effects. These results suggest that
mathematics self-efficacy, language proficiency, and SES are potential sources of DIF between ELLs and non-ELLs. Moreover, mathematics self-efficacy was a significant predictor on six of ten DIF items. Language proficiency was a significant predictor for all ten items even its effect was minimal. SES was a significant predictor on four of ten DIF items.

Mathematics self-efficacy has been confirmed to influence mathematics achievement (Ayotola et al., 2009, Liu et al., 2009; Klassen et al., 2009; Nicolaidou & Philippou, 2003; Shunk et al., 2009). The current study added more evidence that mathematics self-efficacy can be tied up with home language (e.g., Niehaus & Adelson, 2013; Guglielmi, 2012). Niehaus et al. (2013) found ELLs with distinct linguistic backgrounds display different relationships between self-efficacy and mathematics achievement. Similarly, Guglielmi (2012) discovered the relationships between self-efficacy and mathematics achievement are moderated by home language.

Furthermore, the finding that students with high language proficiency are more likely to answer the questions correctly was consistent with previous studies (e.g., Mando, 2007; Stanley, 2005; Danfield, 2013). For instance, Stanley (2005) found that there is a significant difference in mathematics achievement between ELLs and non-ELLs and the achievement gap appears to be narrowing as the amount of language in the texts decreases. Danfield (2013) also discovered that language proficiency is a statistically significant predictor of mathematics scores.

Student SES has been found to influence academic achievement in previous studies (e.g., Aikens & Barbarin, 2008; Jordan & Levine, 2009; Robert & Bryant, 2011), which is consistent findings in current study. ELLs with high SES are usually able to gain access to
extensive literature and more active parental involvement which contributes to their higher achievement levels while ELLs with low SES might not be able to afford to the requisite resources to create positive literacy environment (Aikens et al., 2008).

3) If DIF is detected, can school type and school educational resources contribute to DIF?

HGLM Model 3 was used to examine whether DIF effects decreases or disappears after controlling for both student and school-level variables. Results show that three items still displayed DIF effects after controlling for both student and school-level variables. The rest of seven items no longer displayed DIF effects. Besides, school type is a significant predictor for two items, while school educational resources were not a significant predictor for these ten items.

Over the past decades, researchers have investigated the relationship between school characteristics and students’ mathematics achievement in a large amount of studies (Coleman et al., 1966; Greenwald, Hedges & Laine, 1996; Mullis, Martin, Foy & Arora, 2012). Among those studies that examined the relationship between school type and mathematics achievement, Lubienski and Lubienski (2006) used national data from NAEP and found students in private schools achieve better than students in public schools. However, when controlling for student characteristics, the average differences in adjusted school mean mathematics scores were not significantly different from zero. In terms of school educational resources, it has proven difficult to determine its relationship with mathematics achievement (Sala, 2014). Similarly, evidence in this study was not found to support a significant relationship between school educational resources and mathematics achievement. However, Vandiver (2011) discovered that quality and adequacy of
educational facilities were statistically significantly correlated with student performance. It is suggested more studies should be conducted to identify the important school characteristics of successful schools.

**Conclusions**

Examinations of DIF among language groups are a practical concern due to the increasing language diversity and the prevalence of testing. This study investigated the gap between ELLs and non-ELLs in mathematics achievement at an item level and from perspectives of both students and schools. Results revealed that ten common items are identified with DIF effects using MH procedure, Rasch analysis, and HGLM. These ten items are all in favor of non-ELLs. These findings provided evidence supporting the claim that language ability has negative impact on the mathematics performance for ELLs (Abedi, 2003; Loughran, 2014; Martiniello, 2009).

Item 55, identified by Rasch analysis, was found to be in favor of ELLs. This item may be related to ELLs’ prior educational experiences in their native languages. Although students are classified into ELLs as a result of their lack of English language proficiency, they may have been able to transfer key skills needed for Item 55 from their native languages to English.

When identifying the achievement gap between ELLs and non-ELLs, it is imperative to note that there are many possible reasons for the score differences. For instance, ELLs are more likely from low SES groups and may not have the equal chance to learn the content knowledge of mathematics. The unequal opportunities to learn result in true test score differences (Abedi et al., 2001). Inclusions of covariates in HGLM can solve this issue (Kamata, 2001). Finally, three items show strong evidence of DIF between
ELLs and non-ELLs, even after controlling for student (e.g., mathematics self-efficacy, language proficiency, SES) and school (e.g., school type, school educational resources) level variables. Among the three items, two items (Item 40 and 46) with large DIF effect sizes (above 1.5) were categorized into Class C. According to ETS guidelines, items from Class C should not be used unless they are judged to be essential to meet test specifications (Zwick, 2012). Thus, it is suggested that PISA test developers should examine the language demand for these two items. Modifications or replacements should be made to reduce the DIF against ELLs.

The decreasing number of items showing DIF effects in HGLM Model 2 revealed that mathematics self-efficacy, language proficiency, and SES are potential sources of DIF between ELLs and non-ELLs. In addition, the number of DIF items continued to decrease after controlling for both student and school-level variables. This finding implied that DIF effects between ELLs and non-ELLs can vary in different schools. School type and school educational resources were also potential sources of DIF effects.

In addition to the identification of DIF items, Rasch analysis also provided the information of psychometric properties on PISA 2012 mathematics assessment. The item separation and reliability statistics indicated that the student sample was large enough to confirm the item difficulty hierarchy (construct validity) of the instrument. Person separation and reliability showed that the instrument was sensitive enough to distinguish between high and low performers. Meanwhile, item fit statistics met the requirement of criteria range. The item-person map reveals that there are enough items providing accurate ability estimates across the whole range of students.
Since it is difficult to estimate the amount of error in the data from empirical studies, applying more than one DIF detection approach was suggested to increase the confidence in the results (Hambleton & Jones, 1994; Hidalgo & LÓPez-Pina, 2004; Ilich, 2013). Rasch analysis and HGLM methods in this study were generally in agreement. Rasch analysis only identified one more item that was in favor of ELLs. However, MH procedure discovered six items that were not identified by Rasch analysis and HGLM. This disagreement was mainly resulted from the different mechanism of DIF detection methods. MH procedure used raw scores to match students from different groups for DIF detection, but raw scores cannot represent students’ true ability levels properly when tests have DIF items or the impact is large (Jin, Chen & Wang, 2018). Besides, MH procedure failed to make any assumptions about the classical test theory decomposition of scores. By comparison, Rasch analysis and HGLM can be classified into the parametric and the latent matching category. They are closely linked to a test theory that decomposes an observed score into a systematic true score and a stochastic error score (Kim, 2003; Potenza & Dorans, 1995).

**Implications**

**Implications for Teachers and Educators**

The current study can be used to inform mathematics teachers and educators how best to respond to the instructional needs of their ELL students. Among the three student-level variables, mathematics self-efficacy was a significant predictor on six of ten DIF items. Self-efficacy is impacted by past experiences, social environment, and individual factors (Bandura, 1997). For ELLs, they may struggle to develop self-efficacy through experiences if they have unsuccessful experiences due to their low English proficiency
(Barajas-Lopez, 2014). In addition, they may receive discouraging messages about their capabilities in both mathematics and language proficiency (Kanno & Kangas, 2014).

Therefore, it is necessary for mathematics teachers and educators to develop this psychological belief for all students. Mathematics self-efficacy could be increased by using the right instructional strategies such as helping students to set learning goals, providing timely and explicit feedbacks, encouraging students to study harder, and using high achieving students as models (Liu et al., 2009). As a result of the language barrier and potentially negative perceptions of their academic ability from others, ELLs need additional support from mathematics teachers to enhance mathematics self-efficacy (Briscoe, 2014; Menken & Kleyn, 2010). Since the major sources of self-efficacy include mastery experience, vicarious experience, social persuasion, and psychological responses, it is helpful for ELLs to build self-efficacy by providing more successful experiences with mathematics, modeling, and verbal affirmations (Bandura, 1994). In addition, since learning involves the development of students’ identities in communities of practice, providing more opportunities for ELLs to share solutions and engage in discussions have been found to be effective to develop mathematics self-efficacy for ELLs (Barajas-Lopez & Aguirre, 2015; Takeuchi, 2016).

The finding in current study aligned with previous studies that proved language proficiency is a determinant factor to influence mathematics achievement (Abedi, 2002, Abedi et al., 2001; Haag, et al., 2013; Loughran, 2014). Prediger et al. (2018) found students with low language proficiency do not only encounter reading obstacles in the test situation, but also experience long-term accumulation of deficits for overcoming processual and conceptual obstacles in mathematics learning. Therefore, it is suggested
mathematics instruction should not isolate the word level from the discourse level. Teachers should provide opportunities for the discourse practices of explaining meanings of mathematical concepts and operations (Setati, 2005). The lexical support of meaning-related vocabulary offered in structured phrases rather than isolated words are important to the technical vocabulary (Moschkovich, 2013; Prediger & Wessel, 2013).

**Implications for Test Developers**

The current study also provided some implications for test developers in terms of assessment development and administration. Reliability and validity of testing should be the primary considerations to make sure students are competing in an equitable manner (Lane et al., 2015; Qian, 2011). Equity and comparability of test scores are essential for the validity of score interpretations for subgroups for one occasion and across occasions. A critical assumption in testing is the test score is measuring the same construct with the same precision for all subgroups of students. If that assumption is satisfied, comparisons of the score among different subgroups are appropriate and meaningful (Lane & Leventhal, 2015). A test with high language demands will result in measurement error of ELL students’ mathematics content knowledge.

For assessment development, it is suggested testing companies should be more aware of linguistic diversity within student population to make academic assessments more accessible for ELL students. Better data should be gathered to help test developers understand ELL population (Sireci & Faulkner-Bond, 2015). Moreover, quantitative and qualitative control procedure should be included to facilitate validity for subgroups of students. Quantitative process should include item analysis to evaluate statistical qualities such as item difficulty and discrimination. DIF detection comparing ELLs and non-ELLs
should be standard procedure for the evaluation of mathematics assessment as gender and ethnicity DIF studies are currently doing. The qualitative process can incorporate the sensitivity review, which is an independent review of tests and items by experts trained to consider the unique characteristics of important subgroups (Sireci et al., 2015; Sireci & Mullane, 1994). ELL students can also be interviewed and asked to explain why the pilot items confused them (Ilich, 2013).

**Implications for Policymakers**

At present, accommodations can be considered to promote validity of score interpretations for ELL students (Abedi, 2002; Abedi, Hofstetter & Lord, 2004; Sireci et al., 2015). However, there are no existing standards that can guide the use of testing accommodations for ELLs and state policies on accommodations permitted for ELLs vary widely from state to state. It is imperative for policymakers to cooperate with researchers to find the most appropriate method of testing accommodations. At present, accommodations are classified into two categories including direct and indirect linguistic support. Direct linguistic support involves translation and implication of assessment content. Indirect linguistic support involves extended testing time and bilingual dictionaries (Pennock-Roman, 2011). Recently, some progress has been made to develop systems for making decisions on testing accommodations for ELLs, but additional work is required before any of these systems are fully ready for use by administrators or teachers (Bailey & Carol, 2015).

**Limitations and Future Research**

The first limitation of this study was the classification of ELL students in the sample. In this study, students who were born outside United States and whose primary
language spoken at home was not English were classified as ELL students. While reading literacy was used to represent language proficiency and was controlled as a covariate in HGLM, levels of English proficiency for ELL students were unknown. Some students who were categorized into ELL group may have transitioned out of English as a second language classes. Some ELL students may have already spent years in an English classroom and have reached the same proficiency level as their peers. This variability within the ELL group may limit the results. It is recommended that large-scale assessments can collect samples of data from students who are at varying levels of English language proficiency. Also, it is recommended OECD can include an assessment for home language literacy or survey items to screen status of language learners (e.g., time in the United States, time in ELL program) in the future circle of assessment.

The second limitation in this study was the unbalanced number of students in the focal and reference groups. The ratio of focal group to reference group was 1.5 to 10. While this ratio was acceptable for DIF detections using MH detection, Rasch analysis and HGLM, future studies should attempt to balance focal and reference groups to increase the statistical power and provide more accurate estimates (Jodoin & Gierl, 2001; Paek & Guo, 2011). Testing companies can consider including oversampled ELL students to have balanced samples.

PISA only released a small portion of mathematics items so that it is impossible to review the detailed content of each item. Further comprehensive content analysis on the DIF items should be conducted when the PISA 2012 mathematics assessment items are released in the future. First, vocabulary and terminologies of DIF items should be reviewed by mathematics educators and assessment experts to see whether cultural bias exists.
Second, content reviews can be made to rate the level of linguistic complexity by experts in the areas of literacy, linguistics, and bilingual education. Whether linguistic complexity can predict the magnitude DIF effects between ELL and non-ELL students can be investigated.

In terms of the estimation procedure in HGLM, PROC GLIMMIX under SAS 9.4 employed maximum likelihood estimation. According to Raudenbush and Bryk (2002), the estimation of group means put more weight on the grand mean when group size is unbalanced. It will result in the unstable estimation of random coefficients. Future studies may also explore additional estimation procedures in the application in the multilevel DIF analysis.

In conclusion, future studies should continue with how to solve the threats of valid assessment for ELL students. Additional progress should be made to improve test development and administration processes so that educational assessments are as fair as possible for ELL students.
References


Klassen, R., & Usher, E. (2010), Self-efficacy in educational settings: Recent research and emerging directions, in T.C. Urdan and S.A. Karabenick (eds.). *The Decade Ahead: Theoretical Perspectives on Motivation and Achievement* (pp. 1 – 33), London, United Kingdom: Emerald.


Shirley, T. (2014). *Language learners and the impact of reading ability and income and education inequality on science and math achievement: Discriminate function analysis (DFA) and differential item functioning (DIF) studies of the PISA across 45 countries.* ProQuest Dissertations and Theses database. (UMI No. 3646388).


VITA

EDUCATION

University of Kentucky
Ph.D., Quantitative and Psychometric Methods  Lexington, KY  August 2014 – Present

University of Kentucky
Master of Science, Educational Policy Studies and Evaluation  Lexington, KY  August 2012 – May 2014

Southwest University
Bachelor of Art, English Education  Chongqing, China  September 2007 – June 2011

RESEARCH AND WORK EXPERIENCES

UK Robinson Scholar Program
Research Assistant  August 2016 – May 2017

UK Evaluation Center
Research Assistant  August 2014 – June 2016

American Board of Family Medicine
Intern of Psychometrics Services  August 2015 – December 2015

MANUSCRIPTS


CONFERENCE PRESENTATIONS


Poster was presented to the annual convention of the American Educational Research Association, San Antonio, TX.


Liu, R., Seteri, R., Stanek, M., & Bradley, K. (2015, October) *A Discussion of Sampling Methodology of Kentucky Statewide Victimization Survey*. Paper was accepted by the annual meeting of Mid-Western Educational Research Association, Chicago, IL.


**HONORS AND SCHOLARSHIPS**

| Outstanding Student Paper by AERA Survey Research in Education SIG | April 2015 |
| UK Frank G. and Elizabeth D. Dickey Graduate Fellowship | April 2014 |
| UK International Student Tuition Scholarship | April 2013 |