

6-2020

# DIFFERENTIAL ITEM FUNCTIONING (DIF) OF ENGLISH LEARNERS ON CONSTRUCTED RESPONSE (CR) AND MULTIPLE CHOICE (MC) ITEMS

Michael Nguyen

Follow this and additional works at: <https://scholarworks.lib.csusb.edu/etd>



Part of the [Educational Assessment, Evaluation, and Research Commons](#)

---

## Recommended Citation

Nguyen, Michael, "DIFFERENTIAL ITEM FUNCTIONING (DIF) OF ENGLISH LEARNERS ON CONSTRUCTED RESPONSE (CR) AND MULTIPLE CHOICE (MC) ITEMS" (2020). *Electronic Theses, Projects, and Dissertations*. 998.

<https://scholarworks.lib.csusb.edu/etd/998>

This Dissertation is brought to you for free and open access by the Office of Graduate Studies at CSUSB ScholarWorks. It has been accepted for inclusion in Electronic Theses, Projects, and Dissertations by an authorized administrator of CSUSB ScholarWorks. For more information, please contact [scholarworks@csusb.edu](mailto:scholarworks@csusb.edu).

DIFFERENTIAL ITEM FUNCTIONING (DIF) OF ENGLISH LEARNERS ON  
CONSTRUCTED RESPONSE (CR) AND MULTIPLE CHOICE (MC) ITEMS

---

A Dissertation  
Presented to the  
Faculty of  
California State University,  
San Bernardino

---

In Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Education  
in  
Educational Leadership

---

by  
Michael Quoc Nguyen  
June 2020

DIFFERENTIAL ITEM FUNCTIONING (DIF) OF ENGLISH LEARNERS ON  
CONSTRUCTED RESPONSE (CR) AND MULTIPLE CHOICE (MC) ITEMS

---

A Dissertation  
Presented to the  
Faculty of  
California State University,  
San Bernardino

---

by  
Michael Quoc Nguyen

June 2020

Approved by:

Joseph Jesunathadas, Committee Chair

Edward D'Souza, Committee Member

Ariel Rodriguez, Committee Member

© 2020 Michael Quoc Nguyen

## ABSTRACT

Research assessing Differential Item Functioning (DIF) in mathematics has mainly dealt with gender and content. The mathematics assessments used for those studies primarily focused on around Multiple Choice (MC) and Constructed Response (CR) item types. DIF research studies with English Learners (EL) focused on language complexity and accommodations. The mathematics items used in these studies also consisted of MC and CR items. The primary objective of this cross-sectional non-experimental quantitative study was to determine if DIF occurred between EL and non-EL students and at a more granular level, if DIF existed among students with different levels of English language proficiency as determined by the ELPAC on item types other than MC and CR.

In this study, the responses to ALEKS chapter tests for 8<sup>th</sup> grade students were analyzed. WinSteps software was used to transform the tests raw data into Rasch measures. DIF Pairwise-Rasch-Welch analysis was used to examine the responses of 463 students to determine if DIF was present between EL and non-EL students. The results showed that three Equation/Numeric items had DIF between EL and non-EL students. A t-test was used to examine the responses of 142 EL students to determine if DIF was present among students with different ELPAC levels. The analysis showed that DIF existed among students with different ELPAC levels on two Graphing (G) items and two Equation/Numeric (EQ) items. No commonality was found as to why DIF existed between EL and

non-EL students on the three EQ items. For ELPAC students, Graphing items were easier for ELPAC1 students while the Equation/Numeric item with language complexity was more favorable for ELPAC4 students than the other three EL students. It is recommended that teachers be made aware of potential DIF across test items and that they practice the routine usage of testing accommodations for EL students on assessments that are appropriate to their ELPAC level thus reducing the potential for DIF.

## ACKNOWLEDGEMENTS

I would like to thank the many different people and entity that have helped me through this journey. It was a difficult journey to walk alone; thankfully, I was not alone on this journey. I would like to thank Dr. Jesunathadas, my dissertation chair, for his encouragements and confidence on me. His kindness and supports helped pave the way for my completion. Thanks to Dr. Edward D'Souza for all of his excellent questions. A special thanks to Dr. Ariel Rodriguez for having kindly and calmly dealt with all of my questions, comments, and concerns. He was there at every step of the way. A special appreciation to Cohort 11 for being there for me, they helped me cope. Last, but not least, a big thank you to Bill Trac for editing my dissertation.

## DEDICATION

I would like to dedicate this accomplishment to my mom. She has started the inception process decades before I enrolled into the program. She made sure that I received all of the home support that I need. Her encouragements were invaluable. She would often say, "While I am still alive, I will cook and help support you with whatever I can." I would not be able to finish this without her. To my mom for all of her love, dedication, and support.



## TABLE OF CONTENTS

ABSTRACT .....	iii
ACKNOWLEDGEMENTS.....	v
LIST OF TABLES .....	ix
LIST OF FIGURES .....	x
CHAPTER ONE: INTRODUCTION .....	1
Background.....	1
Problem Statement .....	4
Purpose Statement .....	5
Research Questions .....	5
Significance of the Study .....	6
Theoretical Underpinnings .....	6
Assumptions .....	7
Delimitations .....	7
Definitions of Key Terms.....	8
Summary .....	9
CHAPTER TWO: LITERATURE REVIEW.....	10
Background.....	10
Format Familiarity .....	12
Item Type .....	14
Language Complexity .....	17
Testing Accommodations.....	18
Constructed Response Versus Multiple Choice Item Format.....	26

Differential Item Functioning (DIF) .....	28
CHAPTER THREE: RESEARCH DESIGN AND METHODOLOGY .....	31
Research Design .....	31
Research Setting .....	33
Research Sample .....	34
Research Data .....	36
Partial Credits.....	37
Instrumentation .....	38
Item Categorization .....	41
Data Collection .....	42
Data Analysis.....	43
Summary .....	48
CHAPTER FOUR: RESULTS.....	50
Introduction .....	50
Sample Demographics and Data Consolidation .....	51
Sample Demographics.....	51
Participants Data Consolidation .....	52
Test Data Consolidation.....	53
Results of the Study.....	55
Research Question One.....	55
Research Question Two.....	68
Summary .....	81
CHAPTER FIVE: CONCLUSIONS AND RECOMMENDATIONS.....	83
Introduction .....	83

Limitations of Study .....	83
Characteristics of Assessments.....	84
Interpretation of Results.....	86
Research Question One.....	86
Research Question Two.....	90
Recommendations for Educational Leaders .....	96
Recommendations for Future Research .....	97
Conclusion.....	99
APPENDIX A: INSTITUTIONAL REVIEW BOARD APPROVAL LETTER .....	100
APPENDIX B: CHAPTER 3 TEST.....	103
APPENDIX C: CHAPTER 5 TEST.....	112
APPENDIX D: MEASURE ORDER OF 463 STUDENTS AND 29 ITEMS.....	120
APPENDIX E: MEASURE ORDER OF 142 STUDENTS AND 29 ITEMS .....	122
APPENDIX F: SUMMARY OF DIF ANALYSIS BY ELPAC LEVEL .....	124
APPENDIX G: COMMAND CODE FOR ANALYSIS OF 463 STUDENTS AND 29 ITEMS .....	130
APPENDIX H: COMMAND CODE FOR ANALYSIS OF 142 STUDENTS AND 29 ITEMS .....	132
REFERENCES.....	134

## LIST OF TABLES

Table 1. Frequency Table for Demographic Information .....	52
Table 2. Valid Data Points After Removing Entry without Test Scores.....	53
Table 3. Item Breakdown by Type for Chapter 3 .....	54
Table 4. Item Breakdown by Type for Chapter 5 .....	55
Table 5. Summary Statistics for 441 Measured (Non-Extreme) Students .....	58
Table 6. Summary Statistics for 463 Measured (Extreme and Non-Extreme) Students .....	60
Table 7. Summary Statistics for 29 Measured (Non-Extreme) Items.....	61
Table 8. Summary of Z-fit Statistics of Items Falling Above and Below 2 Zstd ...	62
Table 9. Item Dimensionality Summary of 463 Students and 29 Items .....	64
Table 10. Summary of DIF Analysis by EL Status (Rasch-Welch Analysis) .....	66
Table 11. Items Meeting or the Criteria to Reject the Null Hypothesis.....	68
Table 12. Summary Statistics for 140 Measured (Non-Extreme) Students .....	72
Table 13. Summary Statistics for 142 Measured (Extreme and Non-Extreme) Students .....	73
Table 14. Summary Statistics for 29 Measured (Non-Extreme) Items.....	74
Table 15. Summary of Z-fit Statistics of Items Falling Above and Below 2 Zstd .	75
Table 16. Item Dimensionality Summary of 142 Students and 29 Items .....	77
Table 17. Summary of Items Meeting the Criteria to Reject the Null Hypothesis	79
Table 18. Correlational Analysis of Item Measures for Different ELPAC Level ..	81

## LIST OF FIGURES

Figure 1. Summary of ELPAC overall reporting levels.....	9
Figure 2. Item format conditions (Moon et al., 2018). .....	25
Figure 3. District longitudinal math CAASPP results of percentage of students getting a score of proficient or advanced.....	35
Figure 4. 8 <sup>th</sup> grade longitudinal math CAASPP results compared with county and state.....	36
Figure 5. A description of the CAASPP claims that provide a summary about what students are able to do.....	40
Figure 6. A variable map of 463 students and 29 items.....	57
Figure 7. A bubble map of the items z-fit statistics for the 29 items.....	62
Figure 8. A graph of the DIF measures of EL and non-EL students with trend lines. ....	65
Figure 9. A variable map of 142 students and 29 items.....	70
Figure 10. A bubble map of the items z-fit statistics for the 29 items.....	75
Figure 11. A graph of the DIF size for student with different ELPAC level.....	78
Figure 10. Item 306 – the 6 <sup>th</sup> question in chapter 3 test.....	87
Figure 11. Item 509 – the 9 <sup>th</sup> question in chapter 5 test.....	88
Figure 12. Item 515 – the 15 <sup>th</sup> question in chapter 5 test.....	89
Figure 13. Item 308 – the 8 <sup>th</sup> question in chapter 3 test.....	92
Figure 14. Item 313 – the 13 <sup>th</sup> question in chapter 3 test.....	93
Figure 15. Item 501 – the 1 <sup>st</sup> question in chapter 5 test.....	94
Figure 16. Item 502 – the 2 <sup>nd</sup> question in chapter 5 test. ....	95

# CHAPTER ONE

## INTRODUCTION

### Background

Standardized testing has become entrenched and institutionalized in the educational system of the United States and California (Haertel & Calfee, 1983). States and districts are strongly incentivized to demonstrate growth by continuously achieving average test scores that are higher than the previous year. In their fervor to demonstrate overall growth and improvement, however, states and districts have allowed certain subgroups to get ahead while others fall behind, creating an achievement gap in our educational system (Students affected by achievement gaps, n.d.). A recent report of the National Assessment of Educational Progress (NAEP) 8th grade math test from the Nation's Report Card showed that African American and Hispanic students significantly trailed behind their white counterparts by 33 and 24 points and their Asian counterparts by 50 and 41 points respectively (NAEP mathematics: National student group scores and score gaps, n.d.).

Changes in standardized testing may have inadvertently widened the achievement gap. Prior to the adoption of the Common Core State Standards in Mathematics (CCSSM) in 2013, Standardized state testing in California used multiple-choice (MC) as its only item type (Alcocer, n.d.). The multiple-choice test in arithmetic, which emphasized a single correct answer, tends to be less difficult than created responses test (Kastner & Stangla, 2011). Single-answer

multiple choice tests fail to test depth and complexity, as well as analysis, statistical inference, mathematical problem solving, and mathematical communications.

As educators recognized the need for greater depth and complexity into the school curriculum, especially mathematics education, a new movement toward national standards emerged in 2010 (Akkus, 2016). At that time, a national curriculum for mathematics and English language took shape in the form of the Common Core State Standard Initiative (Khaliqi, 2016). This curriculum was adopted by the California Department of Education (CDE) and the framework for CCSSM education came about in 2013 (California Department of Education, 2017).

One positive aspect of the adoption of the Common Core standards is that the mathematics domains in CCSSM are closely aligned with the domains tested in the Trends in International Mathematics and Science Study (TIMSS), a series of international assessments of the mathematics and science knowledge of students around the world; however, the CCSSM still lacks rigor in key areas of algebraic knowledge and problem solving (Khaliqi, 2016). Despite those shortcomings, the commonalities between the CCSSM and TIMSS standards could possibly account for the steady increase of mathematics achievement scores of fourth and eighth graders in the U.S. on the TIMSS. In conjunction with the CCSSM, the Standard of Mathematical Practices (SMPs) was also released to assist teachers in teaching the CCSSM:

1. Make sense of problems and persevere in solving them
2. Reason abstractly and quantitatively
3. Construct viable arguments and critique the reasoning of others
4. Model with mathematics
5. Use appropriate tools strategically
6. Attend to precision
7. Look for and make use of structure
8. Look for and express regularity in repeated reasoning

(Inside Mathematics, 2018)

The emphasis on standardized tests shifted from multiple choice tests in which a single answer was emphasized to applications and communication of knowledge. The eight SMPs now serve as an integral part of mathematics education in every math classroom and is the guiding principle to comprehensive and effective instructions (California Department of Education, 2015). These standards place more emphasis on the process than the correct answer. The CCSSM “attempts to balance procedure and understanding to draw students away from reliance on procedural algorithms to a more flexible problem solving knowledge base” (Khaliqi, 2016, p. 201). Students are asked to analyze their work, draw statistical inferences, critique the thinking of others, and communicate their learning (Common Core State Standards Initiative, 2010).

The new assessment that came about as a result of the CCSSM is the California Assessment of Student Performance and Progress (CAASPP). Not



only does it test students on key standards, the CAASPP grades students on the different claims that are presented in the chart below. The format of the CAASPP is also different; no longer is it only composed of multiple choice questions in which a single answer is the preferred method of answering, the CAASPP has questions with single answer, questions with multiple answers, questions with short answers, and a Performance Task (PT) is presented at the end to serve as a culminating activity. The assessment is rigorous and requires students to possess higher order thinking skills. The items on the CAASPP address depth and complexity as well as data analysis, statistical inference, mathematical problem solving, and mathematical communications which were lacking in prior state testing format.

#### Problem Statement

Abedi and Levine (2013) stated that in order for students to do well on the new CAASPP tests, they need to master the content as well as be proficient in all domains of English. Furthermore, the reading comprehension requirement of the 9 different item formats on the CAASPP contribute to inherent testing bias due to its linguistics and language complexity (Abedi & Lord, 2001; Rhodes et al., 2015), which in turn, unequally impacts students with lower English proficiency (Abedi, 2004; Johnson & Monroe, 2004). Researches around the issue of differential item functioning (DIF) on mathematics assessments with respect to English Learners (EL) have mainly focused on only two item format: multiple choice (MC) and constructed response (CR). There is a lack of research on DIF of EL

students on the seven other types of item format (multiple choice with multiple correct responses (MSMC), matching tables, drag and drop, hot spot, table fill in, graphing, equation/numeric) that are present on the CAASPP.

### Purpose Statement

The purpose of this cross-sectional non-experimental quantitative study is to examine the effect of different item format types on the 8th grade ALEKS chapter tests and its effect on students' scores with a special focus on EL students. Students are exposed to the different formats through the use of common formative assessment (CFA) since the beginning of the school year. The focal group for this study will be EL students and the reference group will be all other non-EL students.

### Research Questions

Question 1: Does Differential Item Functioning (DIF) exist between English Learners (EL) and native English speakers on item formats other than MC and CR items?

Question 2: If DIF exists between these groups on item formats other than MC and CR, does DIF exist among students with different levels of English proficiency (ELPAC level)?

### Significance of the Study

Currently, studies involving DIF of EL students on mathematics assessments only deal with MC and CR item formats. There is a lack of research on DIF of EL students with regards to other item formats. This research will provide a better understanding of EL students and their performance on mathematics assessment items on item formats that are not MC and CR. The results would allow educators to better understand how item formats affect EL students and find appropriate accommodation in order to provide EL students with equitable access to assessments.

### Theoretical Underpinnings

This research is an extension of two theoretical framework: format familiarity affects test takers' scores (Baghaei & Aryadoust, 2015), and linguistics and language complexity in math problems is an inherent contributor to testing bias which unequally impact students with lower English proficiency (Abedi, 2004; Abedi & Lord, 2001; Johnson & Monroe, 2004; Rhodes et al., 2015). EL students with low mathematics scores isn't always an indicator of low mathematics skills but could be caused by familiarity of the language present on the test items as well as the comprehension of the problem due to the language complexity of the items.

## Assumptions

Baghaei and Aryadoust (2015) have determined that format familiarity affects test takers' scores. That is to say students who aren't familiar with the format of the test items have more construct-irrelevant variance which leads to construct validity (Rhodes et al., 2017). The students in this study have been exposed to the different item formats on the ALEKS math assessments for more than half a year. This research assumes that the students are familiar with the item formats, which is a contributor to construct validity (Rhodes et al., 2017) and test takers' scores (Baghaei & Aryadoust, 2015), and any DIF on the test items is correlated to linguistic complexities.

## Delimitations

District assessments are the product of collaboration from teachers within the LEAs. This means that it is difficult to ask LEAs to change their math CFAs without prior approval and inputs from their teachers. Furthermore, different districts adopt different books from different publishers which would make it difficult to align assessments. Additionally, each school site in the district is afforded the freedom to modify the CFA to the needs of the site teacher team. As such, choosing one grade level at one school site is the best way to control variance in item quantities and item types. This research is limited to teachers and students from grade 8 at one middle school and recognizes the limitation of sample size in this research.

## Definitions of Key Terms

**Common Formative Assessments (CFA):** math unit assessments that all teachers in the same grade have to administer.

**California Assessment of Student Performance and Progress (CAASPP):** the end of the year test use by the state of California in which all students take.

**Standards of Mathematical Practice (SMP):** a set of standards that were introduced along with the CCSSM to promote good mathematical habits.

**Local Control and Accountability Plan (LCAP):** LEAs yearly plan with goals and actions to address the needs and priorities of the district.

**Differential Item Functioning (DIF):** The performance of a group (better or worse) on an item compared to the expected overall ability of the group to the overall difficulty of the item.

**English Language Proficiency Assessments for California (ELPAC):** is an assessment used by the state of California to see how well English learners (ELs) are progressing annually toward English language proficiency (ELP). The ELPAC has four levels.

**Assessment and LEarning in Knowledge Spaces (ALEKS):** is an online artificial intelligent assessment and learning system.





ELPAC Levels		What Students Can Typically Do at Each Level
LEVEL <b>4</b>		Students at this level have <b>well developed</b> English skills. <ul style="list-style-type: none"> <li>• They can usually use English to learn new things in school and to interact in social situations.</li> <li>• They may occasionally need help using English.</li> </ul>
LEVEL <b>3</b>		Students at this level have <b>moderately developed</b> English skills. <ul style="list-style-type: none"> <li>• They can sometimes use English to learn new things in school and to interact in social situations.</li> <li>• They may need help using English to communicate on less-familiar school topics and in less-familiar social situations.</li> </ul>
LEVEL <b>2</b>		Students at this level have <b>somewhat developed</b> English skills. <ul style="list-style-type: none"> <li>• They usually need help using English to learn new things at school and to interact in social situations.</li> <li>• They can often use English for simple communication.</li> </ul>
LEVEL <b>1</b>		Students at this level are at a <b>beginning stage</b> of developing English skills. <ul style="list-style-type: none"> <li>• They usually need substantial help using English to learn new things at school and to interact in social situations.</li> <li>• They may know some English words and phrases.</li> </ul>

Figure 1. Summary of ELPAC overall reporting levels (*Parent/Guardian resources, n.d.*).

### Summary

Mathematics education is an important stepping stone into upward mobility in society today; yet overall math achievement scores have been low for many districts, especially Okuno’s district for EL students. Low mathematics scores of EL students have been linked to language complexity rather than low mathematics skills. Studies involving DIF of EL students have only used MC and CR test item format and there is a lack of research in DIF of EL students on other item formats.

In the next chapter, a brief summary of the assessment system in the United States will be presented. It will be followed by research on format familiarity when it comes to MC and CR items. The literature review would then provide a review of accommodations that have been thus far researched to assist EL students on mathematics assessments because EL students experience DIF on mathematics items due to language complexity.

## CHAPTER TWO

### LITERATURE REVIEW

#### Background

Alcocer (n.d.) detailed a history of standardized testing in the United States on the National Education Association (NEA) website (<http://www.nea.org//home/66139.htm>). Alcocer (n.d.) noted that the articulation of formal assessment of student achievement started in 1838. Prior to this date, assessments had been done through oral examinations. When schools moved from educating the elite to educating the masses, formal written testing began to be more widely used for assessment. New testing instruments surfaced thereafter to assess students on a wide range of areas, from mental ability to college preparedness. The idea of a common college entrance exam was proposed by Harvard President Charles William Eliot in 1890 and the first set of examinations was administered in 1900. The NEA endorsed standardized testing in 1914 and the College Board started to develop comprehensive examinations in 1916, including performance type assessments (e.g., essay questions). These standardized assessments were classified by the U.S. Bureau of Education as tools used to classify students (Alcocer, n.d.). The first SAT tests were administered in 1926 and statewide testing, started by the University of Iowa in 1929, became widely available in other states by the late 1930s.

By 1930, multiple-choice (MC) item format was the most common assessment format in schools. Efficiency was the driving factor that made MC

item format popular in its conception. With the development of an automatic test scanner in 1936, scanning multiple choice assessments in large quantities was done with ease. Yet even as its popularity grew, some began to criticize this item format for assessment was criticized for encouraging students to memorize and guess.

By 1958, Iowa introduced a system for assessment scoring and reporting to the school systems. Following suit, the Elementary and Secondary Education Act (ESEA) of 1965 established national precedents for using norm-referenced testing to evaluate schools and programs. The apex of the MC item format in assessment took shape when, in 2001, the No Child Left Behind expanded the use of state-mandated standardized testing requiring students to be tested each year. The results of the standardized tests were used to gauge school performance and determine school funding. It wasn't until the introduction of the Common Core State Standards that standardized testing shifted from multiple choice format to multi-formatted item assessments. Currently, there are nine item formats that are used to assess students' academic performance administered by the Smarter Balance Consortium (SBAC) (*2013 Mathematics Framework Chapters - Curriculum Frameworks (CA Dept of Education)*, n.d.).

Educators and practitioners in the classroom always desire to improve students' outcomes on those assessments. As such, there exists a large body of literature around different variables affecting students' achievement on these assessments, especially mathematics achievement. After the proliferation of



norm-referenced test as a mean to evaluate schools propagated by the Elementary and Secondary Education Act of 1965 (Alcocer, n.d.), many studies surfaced to analyze variables affecting students' achievement. Some of the researches on high mathematics achievement revolved around a common theme that students should possess high mathematics self-efficacy (Betz & Hackett, 1983), have low mathematics anxiety (Richardson & Suinn, 1972), and assume a growth mathematical mindset (Rattan et al., 2012) with few focusing on the idea of how format familiarity affects students' scores (Baghaei & Aryadoust, 2015) and differential item functioning between groups of students on different item format.

### Format Familiarity

With the changes to the mathematics framework introduced by the Common Core State Standards, standardized test format migrated from strictly Multiple Choice (MC) to nine different item types (*Smarter Balanced Question Types*, 2018). As such, analyzing how format familiarity and differential item functioning affects students' performance would potentially reveal structural improvements that can be addressed to provide all students a fair chance at doing well on the new CAASPP assessment in mathematics.

While there exist a large body of research on how increasing mathematics self-efficacy, decreasing mathematics anxiety, and possessing a growth mathematical mindset positively influence math achievement scores, there is a deficit in research studies that examine the association between students'

familiarity with assessment format and math achievement. In education, the concept of format familiarity is familiar to psychometricians; the concept format familiarity is not often explored by educators as a variable to increase students' achievement in mathematics. The premise of this research builds on findings by Baghaei and Aryadoust (2015) who applied Rasch measurement theory to find that test format familiarity affected test takers' scores. Baghaei and Aryadoust (2015) examined the English listening comprehension scores of 209 international students from Singapore, Malaysia, and the Philippines. The assessment consisted of 40 binary items based on four audio stimuli: map labeling, multiple-choice, table completion, and sentence completion. Baghaei and Aryadoust (2015) found that "the test formats that were familiar to examinees created smaller construct-irrelevant variance while unfamiliar formats created larger irrelevant variance" (p. 84). They concluded that examinees who were familiar with the format of the tests had less irrelevant variance in their scores than examinees who were not familiar with the format of the tests. The research, however, only addressed English and not math and did not take into account the evolution of test question types and computer adaptive testing (CAT) that is currently used in the CAASPP testing in California.

The evolution of test question types in California has moved from solely using multiple choice on the California Standards Test (CST) to multiple choice with single correct response (MC), multiple choice with multiple correct responses (MSMC), matching tables (MA), short text (CR), drag and drop (DD),

hot spot (HS), table fill in (TI), graphing (G), and equation/numeric (EQ) that are currently used on the CAASPP test (*Smarter Balanced Question Types*, 2018). Given all of these types of question present on the CAASPP, the various formats may pose threat to the validity of a measure (Messick, 1996) and bias may arise as a result (Crocker & Algina, 2008). Through the lens of psychometricians, bias refers to construct validity rather than a question being “unfair” or “discriminatory.” When two examinees with the same ability level have different probabilities in answering the same test item correct, the test item is considered to be biased (Borsboom et al., 2002). This unequal probability might be unintentionally measuring a different dimension rather than the one intended by the test developers (Rhodes et al., 2017). In this respect, format familiarity might be a contributing factor to construct validity.

### Item Type

Given the different types test questions on the CAASPP, words problems are more prevalent, and especially on the performance tasks (*Smarter Balanced Question Types*, 2018). Research has shown that the construct response type word problems may be a contributing factor to inherent testing bias since low mathematics scores are assumed to be linked to low mathematics skills rather than being a multidimensional issue, including linguistics (Abedi & Lord, 2001; Rhodes et al., 2015) and the complexity of the English used in the mathematics items unequally impacting students with lower English proficiency (Abedi, 2004; Johnson & Monroe, 2004).

In a study of 1,174 eighth-grade students from 39 classes in 11 schools from the greater Los Angeles area, representing different language, socioeconomic, and ethnic backgrounds, Abedi and Lord (2001) selected 20 questions from the 69 released eighth-grade of the National Assessment of Educational Progress (NAEP) items, modified their language structure (making it simpler), and administered the questions to the students. When the questions were not modified, proficient English speakers scored significantly higher than ELLs. Abedi and Lord (2001) found that English language learners benefited more than students who were proficient in English in the modified version. Furthermore, Abedi and Lord (2001) found that students' performance was affected through the modification of the language structure and the mean differences were statistically significant.

This finding is consistent with the study by Rhodes et al. (2015) of 264 students from third to fifth grades with intellectual disability from the greater metro-Atlanta area; participants were given the KeyMath-Revised Diagnostic Inventory to measure mathematics achievement. Through their analysis, Rhodes et al. (2015) found that the difficulties encountered by students may have been due to the language structures of the items, and the limitations in language ability of the students affected mathematics performance.

Since there is a correlation between language ability and math performance, the process of language modifications on math assessments provided greater benefits to English language learners (ELL) and students in low

and average math classes (Abedi & Lord, 2001; Noonan, 1990). However, any assessment that is constructed and normed for native English speakers would yield lower reliability and validity when applied to the ELLs population (Abedi, 2003).

As a cautionary note, simplifying the language structure of mathematics items may not be a beneficial across the board accommodation for English Language Learners (Johnson & Monroe, 2004). In a study of 1,232 seventh-grade students in Washington (1,060 general education, 138 special education, 34 ELL), Johnson and Monroe (2004) analyzed the data of two forms of a 20-item math test (16 MC and 4 Constructed Response (CR) taken from the state education website. On one form, the even problems were written in simplified language. Students randomly received one of two forms. On the second form, the odd problems were written in simplified language. Students were randomly given either form of the assessment by the classroom teacher. MC items were scored on a right/wrong fashion while CR items were scored on a 0-2 rubric scale. Johnson and Monroe (2004) found that the simplified format only benefited special education students and did not make any difference in the performance of ELL students. They did acknowledge that the sample size of ELL students was small and the students population was limited which made generalization of results difficult. This acknowledgement in limitation have been echoed by previous research that also yielded no significant differences for ELL students when using simplified language (Rivera & Stansfield, 2001).

## Language Complexity

Not only does the new state mathematics assessment (CAASPP) introduce item formats unfamiliar to students, it also demanded higher English language proficiency of them in order to perform well (California Department of Education, 2019; “Common Core State Standards Initiative,” 2019; “Smarter Balanced Question Types,” 2018). Abedi and Levine (2013) surmised that “to perform well in math in English all students, including ELLs, must not only master math content knowledge, but they must also be quite proficient in all domains of English – including reading, writing, speaking and listening – to perform successfully in the assessments” (p. 27). As such, the Smarter Balanced Assessment Consortium (SBAC), responsible for creating and grading the CAASPP test, has taken measures to identify and eliminate test items exhibiting cultural and linguistic bias through a method known as Differential Item Functioning (Abedi & Levine, 2013). Furthermore, items deemed to be unnecessarily complex in linguistic structure are modified and the consortium planned to explore and incorporate accommodations for ELL students that would make assessments more accessible to ELL students (Abedi & Levine, 2013). Not every accommodation will necessarily serve its intended purposes (Abedi, 2014), some accommodation benefits ELL students (Abedi & Lord, 2001; Noonan, 1990), some accommodations yield no significance difference for ELL students (Johnson & Monroe, 2004; Rivera & Stansfield, 2001), and some accommodations change the focal construct of the test items (Abedi et al., 2004).

Ultimately, educators should look at students' level of English proficiency in order to choose the most appropriate accommodations for ELL students (Abedi, 2014).

An item's language structure affects the performance of ELL students (Abedi, 2003, 2004; Abedi et al., 2004; Abedi & Levine, 2013; Abedi & Lord, 2001; Johnson & Monroe, 2004; Rivera & Stansfield, 2001; Shaftel et al., 2006). "Large performance gaps between ELL students and their native English speaking peers are observed for items with high levels of language demand" (Abedi, 2014, p. 261) with the largest gap in English language arts items (40% to 60% lower), a smaller gap in mathematics problem solving items (8% to 25% lower), and a minimal gap in mathematics computation (0% to 10%) (Abedi, 2003). To reduce the gap, many different types of language-based accommodations have been used to reduce unnecessarily complex linguistic complexity and to provide equity and access to ELL students on standardized test (Abedi, 2014). Abedi (2014) provided a list of language-based accommodations through the facilitation of computers that have been explored by different studies and an analysis of the effectiveness and validity of each accommodation.

### Testing Accommodations

One type of accommodations is the usage a of dual language version of the test. The usage of a dual language version of the test requires additional time to be given for the test due the nature of the increased length of the test. Use of the dual language version of the test yielded mixed results. Duncan et al.

(2005) collected test data of 402 students, focus group data of 68 students, and interviews of 18 students to assess the effectiveness of dual language format as a testing accommodation. They assembled a 60 items assessment which was chosen from the 1990, 1992, and 1996 NAEP eight-grade mathematics item banks. Of the 60 items, 45 items were multiple choice and 15 items were constructed response. Additionally, they included 23 more items dealing with demographic information. An analysis of the test data, focus group data, and interviews revealed that the majority of the students considered the dual language booklet to be useful (Duncan et al., 2005). This result is in contrast to the findings in the meta-analysis conducted by Pennock-Roman and Rivera (2011). The meta-analysis included five studies dealing with this accommodation. Since all of the studies dealt with dual language accommodation for students with time constraints, generalizations of results are limited. While this accommodation might have produced mixed results for ELL students (Abedi, 2014), this is an accommodation that the CAASPP provides for ELL students. This accommodation is known as 'stacked translation'.

The CAASPP is a computer based assessment that provides students with the following embedded universal tools for all students with respect to language for the mathematics portion of the assessment: English dictionary, English Glossary, text-to-speech, and Spanish version of the test (*Embedded Universal Tools, Designated Supports, and Accommodations Video Tutorials*, n.d.).



Since the CAASPP is a computer-based assessment, providing students with an English dictionary isn't as impractical as Abedi (2014) found in his assessment of accommodations. There is no additional cost that would be associated with this accommodation. In reviewing the results of two studies using published dictionaries as a form of accommodation for ELL students, Abedi et al. (2004) found that dictionaries provided mixed benefit for students. What is good, Abedi et al. (2014) noted is that this accommodation does not have any impact on the focal construct. A study of 11<sup>th</sup> graders in New Jersey receiving accommodations: translation of instructions, extra time, and a bilingual dictionary on the High School Proficiency Assessment showed that students who received bilingual dictionary as an accommodation scored the lowest on the mathematics test (Miller, Okum, Sinai, & Miller, 1999). Yet, a study by Albus, Bielinski, Thurlow, and Liu (2001) of 133 Hmong students with limited English proficiency and 69 English-proficient students in Minneapolis examined on four reading passages, two passages with dictionary and two without, revealed that students with self-reported intermediate English reading proficiency benefited from the accommodation while students who self-reported poor English proficiency did not benefit from the accommodation. However, this study was about a reading assessment instead of a mathematics assessment.

With the CAASPP assessment, an English glossary is a universal tool for all students. A pop-up glossary can be accessed when students place the computer cursor over a glossed word. This type of accommodation has been

found to be conditionally successful in several studies (Abedi et al., 2000; Kopriva et al., 2007). Abedi et al. (2000) studied a group of 946 8<sup>th</sup>-grade students taking a test using items from the NAEP. The researchers studied the students under four accommodations: (a) modified English language of the test items, (b) glossary, (c) extra time, and (d) glossary plus extra time. The control group were administered the original NAEP items. These were the students who received the original booklet while the experimental group was given one of four accommodations. The distribution of the booklet was random for the sampled students. Using multiple regression analysis and a criterion scaling approach, Abedi et al. (2000) found that ELL students benefitted from all of the accommodations except for the glossary accommodation. ELL students who were given only the glossary were negatively affected by this accommodation due to information overload and generous time would resolve the issue (Abedi et al., 2000; Pennock-Roman & Rivera, 2011). The result found by Abedi et al. (2000) and Penrock-Roman and Rivera (2011) is echoed by the result obtained by Kopriva et al. (2007) when they studied 272 English-speaking ELL (152 3<sup>rd</sup> graders and 120 4<sup>th</sup> graders) in South Carolina. Students were randomly assigned to receive either no test accommodation, a bilingual dictionary, a bilingual glossary, oral reading of test items in English, both oral reading and a bilingual glossary, both a picture dictionary and a bilingual glossary, or oral reading, bilingual glossary, and picture dictionary; on average, ELL students who

received appropriate accommodations outperform ELL students who received no accommodation or inappropriate accommodations.

Text-to-speech is another embedded universal tool available to students as an accommodation on the CAASPP mathematics exam. This option would not only read the test directions but also the test questions to students. In his analysis of language based accommodation effectiveness and validity, Abedi (2014) stated that there is a lack of research on text-to-speech of test directions; however, studies dealing with this type of accommodation either suggested that this type of accommodation is likely to be responsive to the needs of ELLs (Kieffer et al., 2009) or yielded no significant differential item functioning (DIF) on test items (Young et al., 2008). Acosta, Rivera, and Willner (2008) and Kopriva et al. (2007), however, found that text-to-speech of the test items was effective for ELL students, especially for students with low English proficiency.

A Spanish version of the mathematics CAASPP assessment is available as an accommodation for students who have resided in the U.S. for less than one year prior to taking the CAASPP assessment (*Embedded Universal Tools, Designated Supports, and Accommodations Video Tutorials*, n.d.). A meta-analysis by Kieffer et al. (2009) of the few researches revealed that ELL students scored lower when given the assessment in Spanish rather than in English (Kieffer et al., 2009). This was an analysis of a study conducted by Hofstetter (2003). In this study, Hofstetter (2003) looked at data from 849 8<sup>th</sup> graders from Southern California, who self-reported themselves as Hispanic and identified

Spanish as a second language. The students were randomly assigned a NAEP mathematics test for eighth-grade students that were either of original English content, modified, or original Spanish translation. As a result, 35% of the students received the original version, 37% of the students received the modified version, and 28% received the Original Spanish version. Through the analysis of the data using descriptive analysis via SPSS and ANOVA, Hofstetter (2003) found that students who received the Original Spanish accommodation conditionally performed lower than students who received no accommodation. For students who received mathematics instruction in English, the accommodation had a negative but not significant effect on their NAEP mathematics test scores. In contrast, students instructed in Spanish who took the Original Spanish accommodation performed better than comparable students with no accommodation. The result was indicative of the strong evidence that “students perform better when the language of the mathematics test matches the students’ language of instruction” (Hofstetter, 2003, p. 183).

Understanding how item formats affect test-takers is necessary because it can provide potentially valuable information about students’ scores. Students may be receiving low scores as a result of the inappropriate item design or due to ability level (Moon et al., 2018). However, it is common practice for students to guess or skip on a test item because they do not know how to do the problem (Budescu & Bar-Hillel, 1993; Cronbach, 1941). Students react differently to different item format when they are unsure of the answer (Moon et al., 2018). In

an analysis of test data drawn from seven examinations developed for physicians, with the number of participants for each examination ranging between 132 to 948, Grosse and Wright (1985) found that students tend to respond True rather than False in true-false items when they are unsure of the answer. Cronbach (1941) also found that participants tended to lean toward choosing True when they were unsure of the answer. However, this was not the case when it came to multiple-selection multiple-choice (MSMC) items (Cronbach, 1941).

Moon et al. (2018) conducted a study of 1,091 adults between the age of 20 and 40 with bachelor's degree or higher from Amazon Mechanical Turk. Through a web based survey, participants were given a pretest to gauge the participants' prior math knowledge and followed by a test. The items on the pretest and test covered a range of math topics. The item format of the test differs than the format of the pretest. Moon et al. (2018) developed an assessment with five item format conditions: nonforced-choice grid (NFC) [219 participants], forced-choice grid (FC) [210 participants], multiple-selection multiple-choice (MSMC) [212 participants], forced-choice grid with do-not-know (DK) [225 participants], and grid with all possible options (APO) [225 participants]. After participants took the pretest, they were given the test with all of the items adhering to one of the five format conditions. "The items were content-equivalent regardless of format" (p. 57). The test did not have a time limit and participants were given a sample question with the same format as the

test prior to working on the test. The researchers used an analysis of covariance (ANCOVA) on affirmative selection rate and test scores and the Bonferroni method for each of the significant effects.

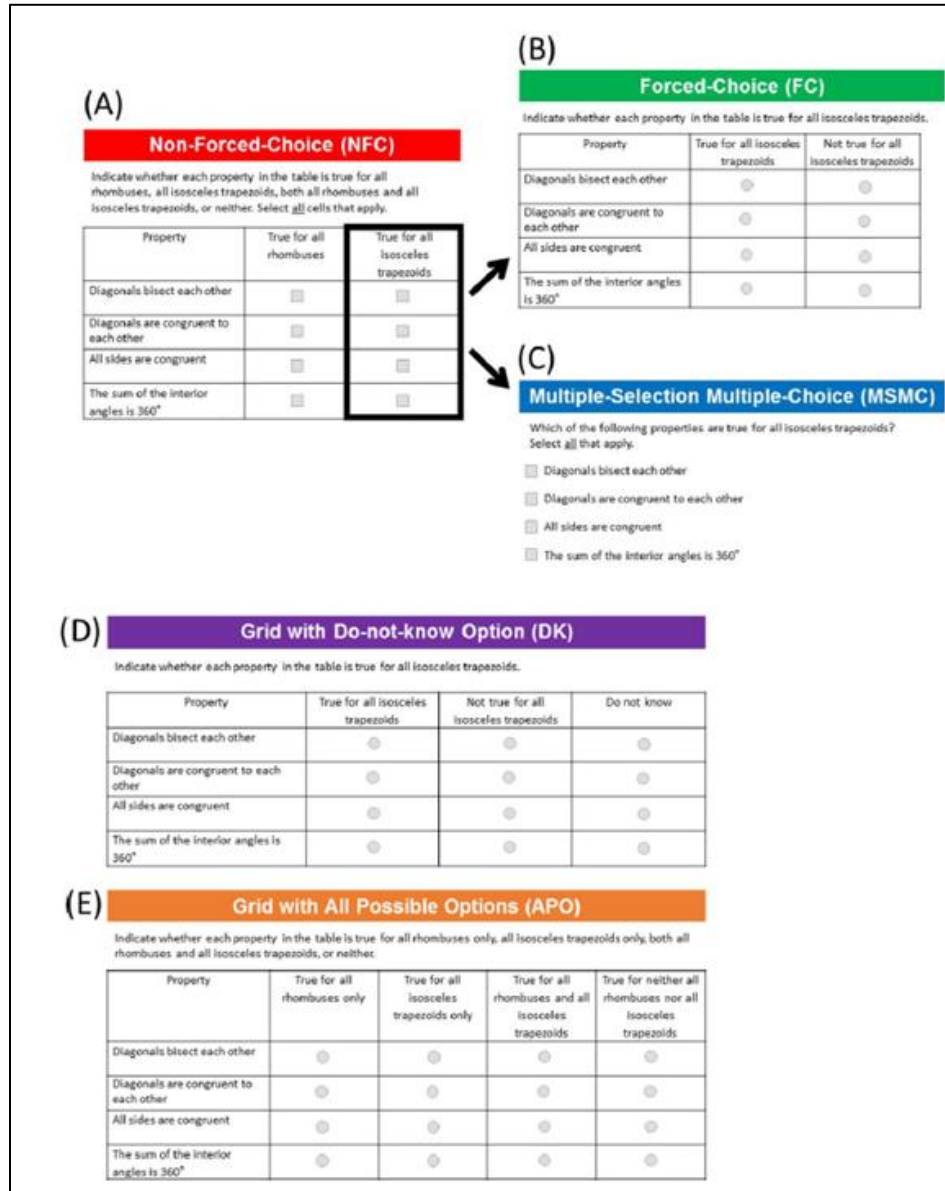


Figure 2. Item format conditions (Moon et al., 2018).

Moon et al. (2018) found that different item formats affected test-takers' willingness to choose an answer when they don't know, hence, affecting their scores. Their result was consistent with Grosse and Wright (1985) and Cronbach (1941) findings that participants had a tendency to respond affirmatively (choosing True) on true-false grid items and in MSMC items. When a do-not-know option was present, the difference no longer existed in the two formats. Furthermore, they confirmed that the presence of a grid resulted in more affirmative responses and different visual layouts (NFC and NPO) resulted in different pattern of answer even if the two formats had the same kinds of answers. The participants' tendency to select one answer per row was prevalent in NFC format but not in the NPO format.

#### Constructed Response Versus Multiple Choice Item Format

Ault (1972) conducted a study on the entire 8th grade class in a suburban New York school to test whether multiple choice (MC) and constructed response (CR) items provided equivalent measurement. Two format of the same test were created which differed in which items were to be represented as multiple choice and which item were to be represented as constructed response. The study showed that MC and CR items provided equivalent measurement and that MC items can be used as replacement for CR items without disrupting the measurement objective of the test (Ault, 1972). Ault (1972) revealed that CR items provided better item-test discrimination than MC items and could be used use to improve test reliability.

While Ault (1972) and Wainer and Thissen (Wainer & Thissen, 1993) argued that CR and MC item formats are interchangeable in assessments, Katsner and Stangla (2011) found that CR and MC questions with multiple responses are not equal when there are differing scoring rules. In a study of 13 graduate students from the Vienna University of Economics and Business, Katsner and Stangla (2011) gave the participants a 17 questions CR test items with varying complexity level and an equivalent level of difficulty MC test with multiple answers on the same day. The participants also received a Study Process Questionnaire after one week for additional insights. The CR questions were graded without knowing the identity of the examinees. There were no penalty for incorrect answers and partial credits were awarded. The MC tests were scored automatically using three different scoring rule: *All-or-Nothing (AN)*, students need to find all correct matches to get credit or else zero, *Number Correct (NC)*, students get credit for responses and incorrect responses are ignored, and *University-specific scoring rule (WU)*, students gets partial credit and guessing is prevented due to incorrect answer being penalized. Using many-facet Rasch measurement (MFRM) approach through the FACETS software developed by John Linacre, Kastner and Stangla (2011) found that CR tests and MC test with multiple responses are equal when *NC scoring* is used but students' ability level are hard to discriminate. However, the researchers found that the two test formats are not interchangeable when using the *AN* or *WU* grading rule. Additionally, they found that students' abilities are discriminated



better when the *WU scoring* rule is used. The researchers further acknowledged that the sample size was too small for generalization.

### Differential Item Functioning (DIF)

The interest in studying how items function differently for different groups started with the examination of item bias. It wasn't until 1986 that a more neutral term *Differential Item Functioning* was proposed to replace item bias since item bias "does not accurately describe the situation" of items with DIF (Holland & Thayer, 1986, p. 1). Differential item functioning refers to how students of the same ability level perform differently on the same test item. When that occurs, researchers state that a differential item functioning is present on the test item. A study on how different groups of participants score differently on the same test item may shed light on the test item as well as the backgrounds and experiences of the test takers. Furthermore, the identification of the test items that have DIF is important because these items pose a threat to equity and access to math education for these groups being compared.

Previous studies have shown that there existed a variety of factors that influenced differential item functioning. One such factor occurred between gender in performance on mathematics and quantitative items (Scheuneman & Grima, 1997; Wang & Lane, 1994). Abedalaziz, Leng, and Alahmadi (2014) and Doolittle and Cleary (1987) showed that differences in item functioning were also related to item content. Research by Abedalaziz, Leng, and Alahmadi (2014) of 1400 eleventh grade students in Kuala Lumpur who took a 40-item instrument

consisting of basic arithmetic, verbal, arithmetic, elementary algebra, and geometry found that females performed better than males in Algebra and males performed better than females in Geometry. This finding was consistent with a previous study by Doolittle and Cleary (1987). Not only did Doolittle and Cleary (1987) looked at item content, they also found in their study of 8 randomly equivalent samples of 1,300-1,400 students taking the 40-item ACT Assessment Mathematics Usage Test that item type also caused differences in item functioning between males and females; they found that word problems were differentially easier for males than for females.

When looking at MC and CR items, even though Ault (1972), Wainer and Thissen (Wainer & Thissen, 1993), and Kastner and Stangla (2011) found the item formats to be conditionally interchangeable, Moses, Liu, Tan, Deng, and Dorans (2013) research showed DIF occurring between males and females. Moses et al. (2013) study of gender DIF using 14 DIF matching variables to evaluate CR items in six forms of three mixed-format tests found that gender DIF does occur depending on the type of tests. Moses et al (2013) analyzed SAT Math test scores from two SAT administration consisting of 235,756 females and 204,956 males test takers. Each SAT test had 10 dichotomously scored CR items and 44 MC items. They also analyzed 2 forms of the Praxis Principles of Learning & Teaching: Grades 7-12 which consisted of 12 4-point CR items and 24 MC items for Form 1 and 23 MC items for Form 2. Furthermore, they analyzed two forms of the Praxis School Leaders Licensure assessment each

with seven 6-point CR items and 76 MC items. With respect to the SAT math test results, Moses et al. (2013) found that, on average, males performed better than females on the major sections of the tests and that MC and CR items measure similar constructs. Additionally, the researchers found that females outperformed males on both sections of the Praxis tests with a greater performance differences in the CR items than on the MC items. Furthermore, the analysis suggested that the CR and MC items of the Praxis tests do not necessarily measure similar constructs.

While a large body of research on item bias and differential item functioning exists, the majority of the research focused on gender or ethnicities as the focal and reference groups. Little research used English Learners as the focal group in their study. As such, this research uses English Learners' English Language Proficiency Assessments for California (ELPAC) as a foundation to divide them into sub-focal groups and compare them to the reference group of all students. This research will focus on differential item functioning of EL students on item formats other than MC and CR items.

## CHAPTER THREE

### RESEARCH DESIGN AND METHODOLOGY

This chapter will present the methodology for the proposed study. The population, setting, sampling procedures, data collection procedures, as well as the instrumentation will be outlined. The chapter will also include the plan to analyze the data to answer the two proposed research questions. Furthermore, this chapter will also provide a section detailing the issue of validity and trustworthiness standards outlined by Creswell (2014). This chapter will conclude with an explanation of the positionality of the researcher in the context of the study.

The purpose of this cross-sectional non-experimental quantitative study is to examine the effect of different item format types on the 8<sup>th</sup> grade ALEKS chapter tests and its effect on students' scores with a special focus on EL students. Students are exposed to the different formats through the use of common formative assessment (CFA) since the beginning of the school year. The focal group for this study will be EL students and the reference group will be all other non-EL students. Two ALEKS chapter tests data will be collected to analyze for DIF between these groups on non MC and CR items.

#### Research Design

This study was a cross-sectional non-experimental quantitative study analyzing sets of math chapter tests data. The data collected for this study came

from the district chapter tests that students took through the district supplemental program that was purchased through the parent company McGraw-Hill. The program is known as ALEKS. The materials that students learn in this course came from the Course 3 Math 8 textbook by McGraw-Hill.

Since testing data was used to analyze for DIF, uniform testing conditions and items was necessary. As such, participants took their chapter tests as they progressed through the instructional year. In order to prevent the issue of format familiarity being a confounding variable in this study, ensuring that students were familiar with the format of the assessments and the interface of the ALEKS program, data were not collected during the first quarter, August – October, of the school year. This step reduced construct-irrelevant variance in the scores (Baghaei & Aryadoust, 2015) and reduced the threat to construct validity (Rhodes et al., 2017).

While other research designs were considered, none met the need for the purpose of this study.

The following research questions guided the research design of this study:

Question 1: Does Differential Item Functioning (DIF) exist between English Learners (EL) and native English speakers on item formats other than MC and CR items?

Question 2: If DIF exists between these groups on item formats other than MC and CR, did DIF exist among students with different levels of English proficiency as determined by the ELPAC?

## Research Setting

The school district is located in the Inland Empire of Southern California. It has an ethnically rich and community with its population being approximately 82% Hispanic or Latino, 11% African-American, 4% Caucasian, and 3% other groups. The school district is the 42nd largest among California's 1,028 school districts. The District serves approximately 25,000 students, pre-school through 12th grade. The District's leadership is committed to promoting continued increased student achievement, fiscal responsibility and solvency, and a safe learning and working environment for enrichment and support to our students, staff and communities. On-going staff development, teacher training, and the recruitment of the most knowledgeable, highly energetic, and committed personnel will promote a model working and learning environment throughout the District.

The participants in this study were limited to all 8<sup>th</sup> grade students currently attending a Middle School with approximately 1,600 students spread almost equally across grades 6, 7, and 8. According to the information the school submit to AVID Center, the students population composed of 91% Hispanic or Latino, 4.5% African American, 1.5% Asian, 2% White, 0.25% American Indian or Alaskan Native, 0.125% Native Hawaiian or other Pacific Islander, and 0.56% identified with two or more races. All of the students in the school received free lunch. In terms of gender break down, the school had 807 female students and 795 male students.

## Research Sample

Only eighth graders from a single school were selected for this study.

This school had the highest population in the district with teachers who were briefed of the study and expressed their interest in participating. Creswell (2014) noted that convenience sampling, while less desirable and is nonprobabilistic, can be used to choose participants based on their convenience and availability. The participants in this study were chosen based on accessibility and geographical proximity.

While there were five middle schools in the district, only 8<sup>th</sup> graders from this school were used because there was inconsistency across the district when it came to the selection of items for chapter tests. While the district provided an assessment template for each chapter, each site could modify the template to fit the needs of the site resulting in no two sites having the same items or the number of items on any given chapter tests. Furthermore, each grade level had its own assessment templates which did not have common items. For this reason, it was not feasible to get participants from multiple sites or grade levels.

The reason that Okuno Middle School (pseudonym) was chosen for this study was because of its geographical proximity to the district office and accessibility of participants for data collection purposes. Eighth graders were chosen due to two important factors. First, the district longitudinal math CAASPP data, in Figure 3 below, had shown that there existed a parabolic trend in scores with 8<sup>th</sup> grade scores being at the minimum of the vertex. The trend started high

in 3<sup>rd</sup> grade, with performance gradually trending downward until 8<sup>th</sup> grade and then jumping back upward in 11<sup>th</sup> grade. CAASPP testing only happens in grade 3 through 8 and 11.

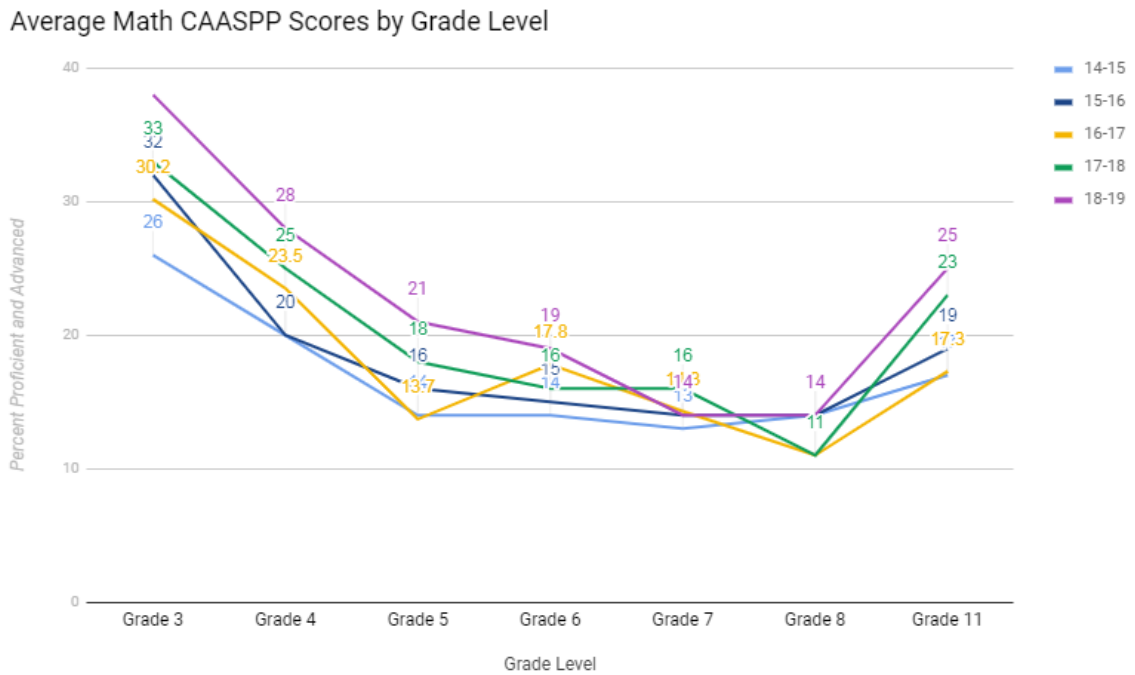


Figure 3. District longitudinal math CAASPP results of percentage of students getting a score of proficient or advanced.

Second, 8<sup>th</sup> grade math CAASPP for the district was significantly lower than the county and state average as shown in Figure 4. The focal group will be EL learners while the reference group will be the non-EL 8<sup>th</sup> grade students at Okuno Middle School.



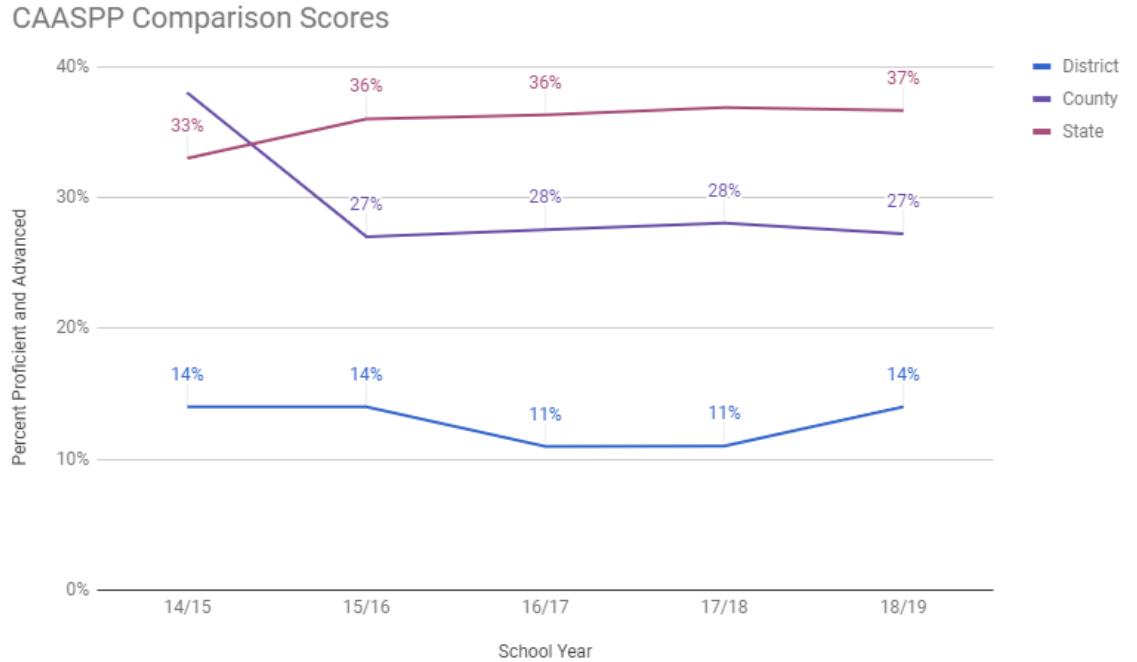


Figure 4. 8<sup>th</sup> grade longitudinal math CAASPP results compared with county and state.

### Research Data

The data for this research came from two ALEKS chapter tests. The first test came from the second quarter (chapter 3 test) and the second test (chapter 5) came from the third quarter. The reason that quarter two and three data was used was to allow students to become familiar with item formats in the first quarter and thus avoided any issue with format familiarity affecting scores. Baghaei and Aryadoust (2015) found that results from students who are not familiar with the format of the items on the assessments have more construct-irrelevant variance. Furthermore, Rhodes et al. (2017) determined that format familiarity might be a contributing factor to construct validity.

### Partial Credits

ALEKS allows teachers the option to provide partial credit to student on items that have multiple parts. As such, teachers at Okuno Middle School used this option when administering the tests to the students. In the process of cleaning up the data for this study, 21 out of the 29 items were given the option of partial credits. The partial credits had a wide range of values: 0.1, 0.17, 0.25, 0.3, 0.33, 0.5, 0.6, 0.67, etc.

The WinStep program uses whole number as valid entry. In order to ensure that the program could successfully analyze the data set, this study used a rounding mechanism to address the issue of partial credits. The rounding method was used uniformly for all of the items in order to create a dichotomous data set. For any data points with partial credit between 0 and 0.49, those data points were rounded down to 0. For any data points with partial credit between 0.5 and 1, those data points were rounded to 1. A score of 0 indicated that the participant got a wrong answer or skipped the question. A score of 1 indicated that the participant got the correct answer. Any students who did not take the test would end up with missing scores for all items on that test and the missing scores would be given a designation of (.), indicating that the test was not administered and the score was not counted against the participants.

## Instrumentation

ALEKS is an online math supplemental program that is associated with the district math adoption of the McGraw-Hill math textbooks during the 2013-2014 school year. The contents of ALEKS were aligned with the materials from the textbooks for grade 6-12. Teachers were encouraged to use the program as supplemental resource. Teachers used the ALEKS diagnostic assessment results to provide students with appropriate instructions. Since the online program was aligned to the classroom textbook, teachers often assigned practice tasks, quizzes, and tests to measure students' performance as they related to the standards rather than used traditional pencil and paper format. The item formats on the practice tasks, quizzes, and tests were similar to what students would encounter on the chapter tests. By exposing students to the different item formats via practice tasks, quizzes, and tests prior to taking the chapter tests, students were familiar with the format of the items; thus, reducing construct-irrelevant variance and reducing threat to construct validity.

One of the characteristic features of the problems on ALEKS was that each problem was accompanied with step-by-step explanation. While it was not a type of feedback that came from teachers, the online feedback was still valuable because it provided students with explanations on how to correctly complete the problem. The feedback was provided to all students and was uniform across tests.

Elewar and Corno (1985) found that teacher feedback provided to students not only improves students' achievement but also attitude toward mathematics. The immediate feedback provided to students through the ALEKS program served as a substitute for the one-on-one teachers' feedback which was often limited or infrequent given the lack of time for such personal one-on-one interaction with students and could potentially replicate positive effects found by Elewar and Corno (1985).

ALEKS monitors student progress and signals when a student is to progress to the next set of problems. It is a self-paced learning and assessment system. When a student successfully completes five problems correctly on a concept, the program will mark the concept as mastered. Furthermore, after spending five hours actively working on the learning path in ALEKS, the program will initiate a knowledge check to assess the students on mathematical concepts that they have mastered. If students miss any of the question(s) on the knowledge check, the program will make students learn the concept(s) associated with the missed problem(s).

Another feature about the program is the type of questions presented. The questions from the program address two out of the three claims (*Figure 5*) dictated by the CAASPP guideline: concepts and procedures and problem solving/modeling and data analysis. The exposure to the type of questions that are similar with the CAASPP questions will provide students the

opportunity to interact with mathematical problems that enable students to learn mathematical concepts at a deeper level.

Area (Claim) Descriptors	Above Standard	Near Standard	Below Standard
<p><b>Concepts and Procedures</b></p> <p>Applying mathematical concepts and procedures</p>	<p><b>The student demonstrates a thorough ability to consistently</b> explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency.</p>	<p><b>The student demonstrates some ability to</b> explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency.</p>	<p><b>The student does not demonstrate the ability to</b> explain and apply mathematical concepts and interpret and carry out mathematical procedures with precision and fluency.</p>
<p><b>Problem Solving/Modeling and Data Analysis</b></p> <p>Using appropriate tools and strategies to solve real world and mathematical problems</p>	<p><b>The student demonstrates the thorough ability to consistently</b> solve a range of complex, well-posed problems in pure and applied mathematics, making productive use of knowledge and problem-solving strategies. The student demonstrates the ability to <b>consistently</b> analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems.</p>	<p><b>The student demonstrates some ability to</b> solve a range of complex, well-posed problems in pure and applied mathematics, making productive use of knowledge and problem-solving strategies. <b>The student demonstrates some ability to</b> analyze complex, real-world scenarios and can construct and use mathematical models to interpret and solve problems.</p>	<p><b>The student does not demonstrate the ability to</b> solve a range of complex, well-posed problems in pure and applied mathematics, making productive use of knowledge and problem-solving strategies. <b>The student does not demonstrate the ability to</b> analyze complex, real-world scenarios and construct and use mathematical models to interpret and solve problems.</p>
<p><b>Communicating Reasoning</b></p> <p>Demonstrating ability to support mathematical conclusions</p>	<p><b>The student demonstrates the thorough ability to</b> clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others.</p>	<p><b>The student demonstrates some ability to</b> clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others.</p>	<p><b>The student does not demonstrate the ability to</b> clearly and precisely construct viable arguments to support their own reasoning and to critique the reasoning of others.</p>

Figure 5. A description of the CAASPP claims that provide a summary about what students are able to do (California Department of Education, n.d.).

### Item Categorization

The purpose of this study was to identify if DIF existed between EL and non-EL students and among students with different ELPAC level on items other than MC and CR items. To correctly answer the research questions, it was necessary to categorize the items on chapter 3 and chapter 5 tests. To differentiate the items between the two tests, the study used the following naming mechanism. Items from chapter 3 would start with the name *Item 301* and end with *Item 314*. Items from chapter 5 would start with the name *Item 501* and end with *Item 515*. There are a total of 29 items for this data set.

Even though ALEKS is an online platform, it does not possess the capability of providing all 9 item types that the CAASPP has. The ALEKS program does not have any items belonging to the short text (CR), drag and drop (DD), or hot spot (HS) category. Given the nature of the content of chapter 3 and chapter 5, only multiple choice (MC), graphing (G), and Equation/Numeric were used. The items on the two chapter tests were divided into the following three item types: multiple choice with single correct response (MC), graphing (G), equation/numeric (EQ). For the purpose of this study, CR items are operationally defined as test items in which students had to provide a short text explanation. Items in which students had to solve and provide an answer received a label of EQ.

## Data Collection

This study collected the test data from two chapter tests administered to 8<sup>th</sup> graders at Okuno Middle School during the 2019-2020 school year. The first chapter test was chapter 3 that was administered during the second quarter. The second chapter test was chapter 5 that was administered during the third quarter. Appropriate steps were taken to obtain permission to use the data for the purposes of the research. The data analyses and resulting findings from the study will be shared with the site to better address the needs of EL students if DIF occurs.

The test data were housed online as part of the ALEKS system and were accessible by the students' teachers. The researcher had to manually obtain the data from each teacher for each period and then merged them together into one file. This needed to be done for each chapter test.

ALEKS did not provide any demographic data about the students except for first name, last name, and local identification number. As such, the researcher needed to extract all relevant demographic data (i.e., gender, EL status, resolved ethnicities, grade level, and local identification number (ID)) for the students from a separate data repository called Illuminate that the school district has used since 2010.

Once the test data and demographic data were obtained, the researcher merged the two sets of data together, using students' ID as a matching criterion. Upon successful completion of the data merger, all students' names and ID

numbers were removed and replaced with arbitrary assigned numbers for the purpose of anonymity and to ensure that the identity of the participants could not be readily linked directly or through identifiers to the participants.

For the purpose of this study, the researcher kept the gender information of each participant and if each student was a native English speaker or an English Learner; and if an English Learner, what ELPAC score did each student had. All other information beside gender, arbitrarily assigned pseudo number and ELPAC score were deleted. The researcher will not need to contact the participants and will not need to re-identify the participants once each participant receives an arbitrarily assigned number. All data were stored on a password protected Rialto Unified School District Google Drive and was accessed through a password protected computer at home. Any files that have students' data were encrypted with password prior to uploading them into the Google Drive. All data relevant to this study will be destroyed after 2 years from the date of first collection.

### Data Analysis

Once both chapter test raw scores were collected and merged with demographic data, the researcher began analyzing the data using the WinStep program. The WinStep program allowed the researcher to conduct a Rasch analysis with the data to determine if DIF exists between ELs and native English speakers on the non-MC and non-CR items.



DIF occurring for EL students does not mean that EL students do better or worse on a particular item. DIF occurs for EL students if they perform better or worse compared to what they are expected to perform relative to their overall performance on the rest of the assessments. Furthermore, the size of DIF must be large enough to be unlikely due to chance and reflect a substantive difference in performance between groups. The analysis compared the results between EL students (focal group) to non-EL students (reference group).

The items on the two chapter tests were divided into the following nine item type: multiple choice with single correct response (MC), multiple choice with multiple correct responses (MSMC), matching tables (MA), short text (CR), drag and drop (DD), hot spot (HS), table fill in (TI), graphing (G), equation/numeric (EQ). All items would be used on the analysis but only items that are non-MC and non-CR would be used to answer the research question since previous researchers studying DIF due to item format have primarily focused on MC and CR items.

A variable map for persons and items was generated to provide information on the logit measure of each item as a mean to compare the difficulty of each item as well as students' ability level on the same scale. The variable map showed two categories (CATS) – right and wrong. The variable map had the Logit scale on the extreme left. Theoretically, the logit scale ranges from negative infinity to positive infinity and have equal intervals between units (Bond & Fox, 2015).

On the variable map, the label <more> was located at the top left and <less> at the bottom left of the vertical line/scale to indicate that participants higher up the scale have higher ability level than the participants lower down the scale. Similarly, the label <rare> and <freq> at the top right and bottom right of the vertical line suggested that fewer students were successful with the items toward the top while more successful with items toward the bottom. Students located toward the top of the scale were most able students while students located toward the bottom of the scale were the least able students. Similarly, items located toward the top of the scale were most difficult while items located toward the bottom of the scale were the least difficult.

The zero of the logit scale is always located as the item mean (Bond & Fox, 2015). This was an arbitrary location for the 0 of the scale. Zero does not mean an absence of ability level nor does negative logits mean a deficit of ability. It is simply a measurement used to compare item difficulty and ability level.

The variable map was used to determine if items would be well “targeted” to the ability of the students. Well “targeted” items would be used to properly separate the ability of students who are clustered together.

Summary statistics for extreme and non-extreme persons and items was provided along with separation value and Cronbach Alpha value. WinStep provided the summary descriptive statistics for the data for all EL and non-EL students. The program would be able to generate 4 possible tables: non-extreme students, extreme and non-extreme students, non-extreme items, and

extreme and non-extreme items. Extreme students were those who missed all of the problems or received full points on the 29 items. Extreme items were items in which all students either missed or got correct.

Each table provided the Mean, Standard Deviation, Maximum and Minimum statistics for the Raw Scores, Logit Measure, Standard Error (standard deviation of the errors) obtained with the Rasch model, the Mean Squares and Standardized-Z for infit and outfit, separation value, and Cronbach Alpha. High separation value would result in high Cronbach Alpha. The higher the separation value, the better it is to separate students' ability (Bond & Fox, 2015). Assuming that these students take the same test over again for reliability, they would end up in the same order. Hence, the test-retest scenario described by Traub and Rowley (1991) would yield a high reliability coefficient.

An item z-fit statistics was ran to determine the existence of any overfitting or underfitting items. WinStep was used to generate a bubble map and a z-fit statistics table for the items and for persons. In the bubble map, there were bubbles of different sizes. The size of each bubble represented its standard error of measurement (SEM). The bigger the SEM value, the bigger the size of the bubble. Any bubbles that fell within the range of -2 to +2 Zstd (Z standard deviation) were considered to be in the fit zone (Bond & Fox, 2015). Any that fell above +2 Zstd were considered to be underfit (too much variability) for the Rasch model and any fell below -2 Zstd were considered to be overfit (fit too well to the Rasch model).

Furthermore, an item dimensionality table was used to show the explained and unexplained variance of the persons and items used in the analysis. This dimensionality table was used to check the reliability coefficient. Given any set of data, persons or items, there is always variance in the data. Total variance comprises of those that can be explained and variance that cannot be explained. The analysis looked at 100% of the variance and divided them into explained and unexplained categories. A good situation is to have data with 50%+ variance belonging in the explained area; anything less than 50% is cause for concerns (Bond & Fox, 2015).

Pearson correlation was generated to see how well EL DIF measures correlated with non-EL DIF measures as well as among students with different ELPAC level.

To answer research question 1: Did Differential Item Functioning (DIF) exist between English Learners (EL) and Native English speakers on item type other than MC and CR items? DIF Pairwise – Rasch-Welch analysis was used to analyze all items and the result let the researcher see if DIF exists between the focal and reference group. The result assisted in answering the null hypothesis of “no DIF” between the focal and reference group on non-MC and non-CR items. Any item with a big enough DIF contrast ( $p < 0.05$ ) allowed for the rejection of the null hypothesis and the acceptance of the alternative hypothesis that DIF between focal and reference group was present.

To answer research question 2: If DIF existed between the focal and reference groups on item type other than MC and CR, did DIF exist among students with different levels of English proficiency as determined by the ELPAC? A t-test was used to analyze all items of EL students and to determine if DIF occurred among students with different ELPAC levels in the focal group. The result assisted in answering the null hypothesis of “no DIF” among students with different ELPAC levels in the focal group on non-MC and non-CR items. Any item with a big enough DIF size ( $p < 0.05$ ) allowed for the rejection of the null hypothesis and acceptance of the alternative hypothesis that DIF existed among students with different ELPAC levels in the focal group. Furthermore, DIF size revealed if the item was harder or easier for certain ELPAC level.

### Summary

The purpose of this cross-sectional non-experimental quantitative study was to determine if DIF occurred between EL and non-EL students and if DIF existed among students with different levels of English language proficiency as determined by the ELPAC. This chapter presented the methodology for the proposed study. The population, setting, sampling procedures, data collection procedures, as well as the instrumentation was outlined. The chapter also included the plan to analyze the data to answer the two proposed research questions. Furthermore, this chapter also provided a section detailing the issue of validity and trustworthiness standards outlined by Creswell (2014). This chapter concluded with an explanation of the positionality of the researcher in the

context of the study. The next chapter will present the findings of the data collection and analysis.

## CHAPTER FOUR

### RESULTS

#### Introduction

The purpose of this cross-sectional non-experimental quantitative study was to determine if DIF occurred between EL and non-EL students and if DIF existed among students with different levels of English language proficiency as determined by the ELPAC. In this chapter, the result of the study will be presented. The chapter includes a description of the sample, steps used to analyze the data using WinSteps and a report of the findings that answer the research questions. Each item on the test was divided into different item type(s) according to their format. WinSteps was used to transform raw scores to Rasch measures. WinSteps was used to run the analysis for this study because it uses the Rasch model as the basis of analysis allowing the transformation of data set from ordinal scale to interval scale (Bond & Fox, 2015).

The software was used to obtain a variable map that showed the difficulty of the items and the ability level of the students. WinSteps analyses also provided summary statistics tables for the students and items. Item separation index and Cronbach Alpha as measures of reliability of measures and scores was calculated. Furthermore, z-fit statistics report is reported to examine the fit of items and persons to the Rasch model. Item dimensionality, the amount of explained and unexplained variance in the responses of items and students, was obtained to examine if the items collectively addressed a single construct.

Finally, DIF Pairwise-Rasch-Welch analysis was used to find if DIF occurs between EL and non-EL students and a t-test on DIF scores was used to test if DIF existed among students with different ELPAC level for any of the test items.

### Sample Demographics and Data Consolidation

#### Sample Demographics

The sample consisted of 545 8<sup>th</sup> graders at Okuno Middle School including 261 males (47.9%) and 284 females (52.1%). This number was the actual number of 8<sup>th</sup> graders enrolled at the time of data collection. The ethnicity of the participants was majority Hispanic ( $n = 497$ , 91.2%). There were a total of 11 White/Caucasian students (2%), 24 Black/African American students (4.4%), and 13 students of Other ethnicities (2.4%). Of the 545 students in the sample, 398 students were categorized as non-EL (73%) and 147 students were EL (27%).

The EL distribution was as follows: 19 students had an ELPAC level 1 (3.5%), 27 students had an ELPAC level 2 (5%), 61 students had an ELPAC level 3 (11.2%), and 40 students had an ELPAC level 4 (7.3%). This demographic information is presented in Table 1.



Table 1.  
Frequency Table for Demographic Information

Categories	<i>n</i> (545)	%
Gender		
Male	261	47.9
Female	284	52.1
Ethnicity		
White/Caucasian	11	2.0
Black/African American	24	4.4
Hispanic	497	91.2
Other	13	2.4
EL Status		
non-EL Students	398	73.0
ELPAC1	19	3.5
ELPAC2	27	5.0
ELPAC3	61	11.2
ELPAC4	40	7.3

#### Participants Data Consolidation

Upon initial data collection, there were a total of 545 8<sup>th</sup> graders enrolled at Okuno Middle School. Since the school site offers an accelerated track for students, 70 of the 8<sup>th</sup> graders were enrolled into this accelerated pathway and take a different course with different chapter tests. After removing the students in the accelerated pathway from the list, 475 students were left. None of the students in the accelerated pathway had an EL designation, i.e., they were all English proficient.

The list of 475 eligible participants was used to extract their test scores. Twelve students who did not have test scores from chapter 3 and chapter 5 were

excluded from the list. Students who only had one of two chapter test scores were kept in the data set. Missing scores for either of the two tests were given the designation of (.) indicating that the test was not administered to them. Consequently, the data set ended up with 463 participants with valid data. Participants data consolidation is shown on Table 2.

Table 2.  
Valid Data Points After Removing Entry without Test Scores

Variable	<i>Initial Collection (475)</i>	<i>%</i>	<i>Participants with Valid Data (463)</i>	<i>%</i>
EL Status				
non-EL	328	69.1	321	69.3
EL	147	30.9	142	30.7
ELPAC1	19	4.0	18	3.9
ELPAC2	27	5.7	26	5.6
ELPAC3	61	12.8	59	12.7
ELPAC4	40	8.4	39	8.4

*Note:* Due to rounding errors, percentage may not add up to 100%.

### Test Data Consolidation

For this study, data from chapter 3 and chapter 5 were collected. Chapter 3 covered the following concepts: rate of change, slope, writing and graphing equations, and solving systems of equations. Chapter 5 covered the following concepts: lines, angles of triangles, understanding and using the Pythagorean Theorem, and distance on the coordinate plane. While the district provided all the middle schools with the same item template, each site made their own changes which resulted in different sites having their own set of items.

Chapter 3 has 14 questions. Of the 14 questions from chapter 3 test, there are eight Equation/Numeric (EQ) questions, four Graphing (G) questions, one question that is both Multiple Choice (MC) and Equation/Numeric (EQ), and one question that is both Graphing (G) and Multiple Choice (MC). Item breakdown by type for chapter 3 is shown in Table 3.

Table 3.  
Item Breakdown by Type for Chapter 3

	<i>Item Type</i>		
	<i>MC</i>	<i>G</i>	<i>EQ</i>
Item 301			x
Item 302			x
Item 303			x
Item 304			x
Item 305			x
Item 306			x
Item 307		x	x
Item 308		x	
Item 309		x	
Item 310		x	
Item 311			x
Item 312	x		x
Item 313	x	x	
Item 314			x

Chapter 5 has 15 questions. Of the 15 questions from this chapter test, twelve are EQ questions, one question that is MC, and two questions that are both MC and EQ. Item breakdown by type for chapter 5 is shown in Table 4.

Table 4.  
Item Breakdown by Type for Chapter 5

	<i>Item Type</i>	
	<i>MC</i>	<i>EQ</i>
Item 501		x
Item 502		x
Item 503		x
Item 504		x
Item 505	x	x
Item 506	x	x
Item 507		x
Item 508		x
Item 509		x
Item 510	x	
Item 511		x
Item 512		x
Item 513		x
Item 514		x
Item 515		x

## Results of the Study

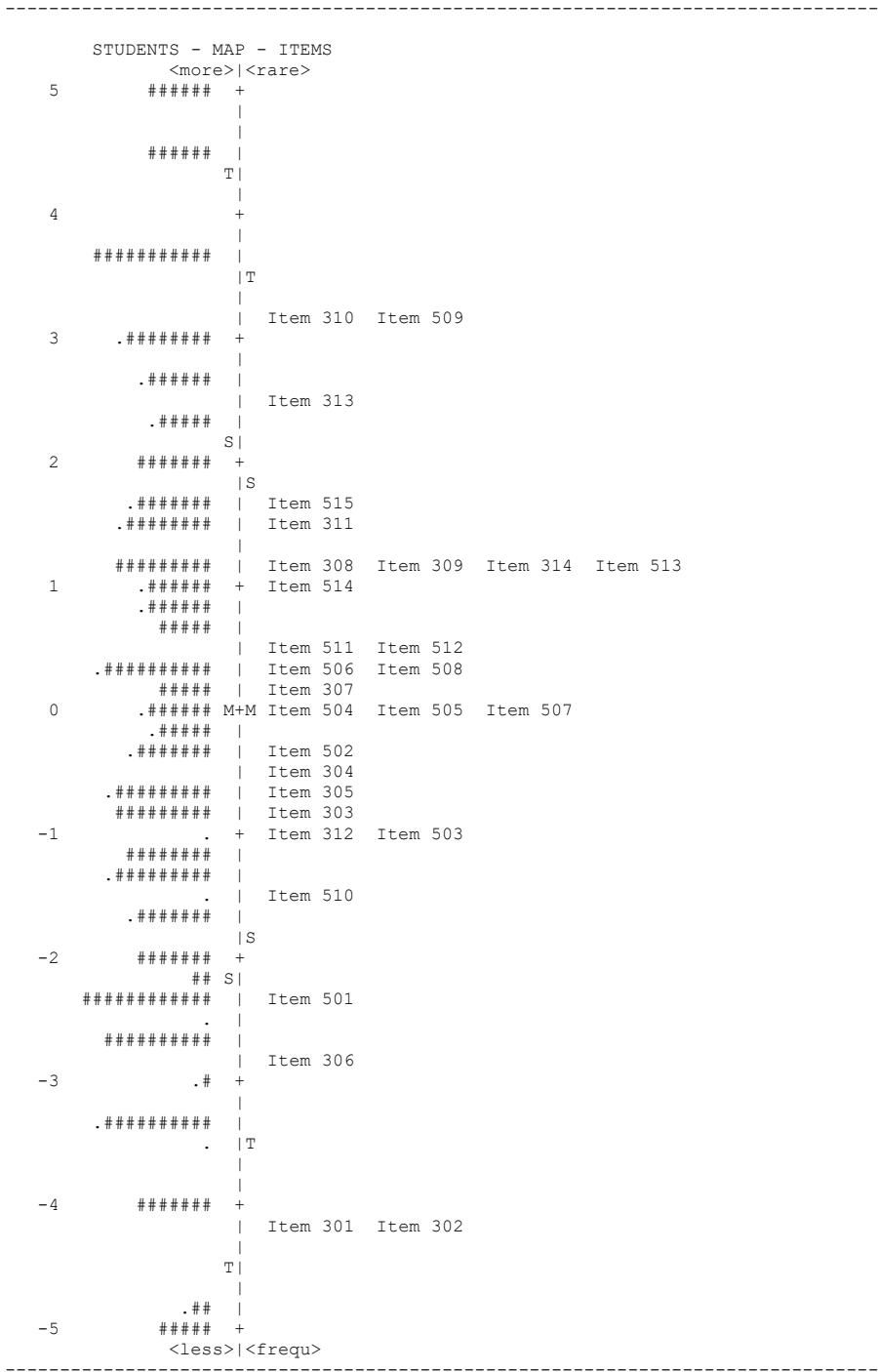
### Research Question One

Variable Map. Using the data of 463 students and 29 items, a variable map or Wright Map (*Figure 6*) was generated that provided the difficulty of each test item as compared to the students' ability (Bond & Fox, 2015). The map shows the item difficulties and student abilities on a common interval scale in logit units. Students' ability measures are identified on the left of the line in the middle of the map while the item calibrations are identified to the right. The scale for this map extends from -5 logits to 5 logits. The test items, organized by difficulty level (logit measure in descending order) with the most difficult items at

the top of the list and the least difficult at the bottom, is presented in Appendix D. On the vertical scale, the mean difficulty of the Items and the mean of the ability level of the students were both located at the '0' mark of the logit scale. The students were located between -5 and 5 while the items were located between -4.5 and 3.5 logits.

A close examination of the map shows that the mean of the participants' abilities was very close to the mean of the item difficulties indicating that the average abilities of the participants matched the average item difficulty. Generally speaking, the items were well 'targeted' to the abilities of all students except at the extreme ends. There were no items that were 'targeted' to the most able students and to the least able students. When items are not well targeted to the abilities of the students, the measures are associated with larger measure errors than those student measures that are that well targeted.

Item 310 and Item 509 were located at the top of the scale indicating that these two items were the most difficult items out of the 29 items. Conversely, the variable map also showed Item 301 and Item 302 to be the least difficult.



Note: Each '#' is 2 and each "." is 1.

Figure 6. A variable map of 463 students and 29 items.

Summary Statistics. Winsteps provides two types of analyses: (a) analyses with non-extreme students/items and (b) analyses with extreme students.

Table 5 shows the scores of non-extreme students. The mean logit measures of student abilities was -0.07 and the standard deviation was 2.18 logits. The ability measures ranged from a maximum value of 4.42 logits to a minimum value of -4.90 logits. On average the data appear to fit the Rasch model with average infit mean-square (MNSQ) value of 0.97 and average outfit MNSQ value of 0.98. These values were close to the expected mean-square value of 1.0. The outfit MNSQ being lower than 1.0 indicates that the data slightly overfit the model; there was more predictability and less variability in the data than expected under the Rasch model (Bond & Fox, 2015).

Table 5.  
Summary Statistics for 441 Measured (Non-Extreme) Students

	Raw Score	Model Count	Measure (logits)	MNSQ	
				Infit	Outfit
Mean	13.9	28.6	-0.07	0.97	0.98
S.D.	8.2	2.5	2.18	0.29	0.96
Maximum	28.0	29.0	4.42	2.56	9.90
Minimum	1.0	14.0	-4.90	0.27	0.09
Real RMSE	= 0.64				
Separation Index	= 3.24		Student Reliability = 0.91		

*Note: Extreme students are those who received full points or 0 point.*

*Maximum extreme score: 12 students*

*Minimum extreme score: 10 students*

*Valid responses: 98.5%*

The analysis of extreme and non-extreme students (Table 6) showed that the mean of the Rasch logit measure was -0.03, the standard deviation was 2.46 logits, with measure ranging from a maximum of 5.73 logits to a minimum of -6.29 logits. The real root mean-square error (RMSE) was 0.75 and the separation index was 3.13. Based on the Rasch measures, the student reliability coefficient was 0.91. The score reliability, i.e. Cronbach Alpha (KR-20), was 0.95 suggested high internal consistency of measures and scores. The Cronbach alpha value of 0.95 indicated that the scores were highly reliable (Traub & Rowley, 1991). This means that there was a high probability that students estimated as having high measures actually did have high measures and students estimated as having low measures actually did have low measures. Furthermore, a Cronbach alpha value of 0.95 can also be interpreted as 95 percent of the observed variance in scores was associated with systematic differences in the performances of the students and 5 percent to errors (Traub & Rowley, 1991).



Table 6.  
Summary Statistics for 463 Measured (Extreme and Non-Extreme) Students

	Raw Score	Model Count	Measure (logits)	MNSQ	
				Infit	Outfit
Mean	14.0	28.4	-0.03		
S.D.	8.6	2.8	2.46		
Maximum	29.0	29.0	5.73		
Minimum	0.0	14.0	-6.29		
Real RMSE = 0.75					
Separation Index = 3.13			Student Reliability = 0.91		
Cronbach Alpha (KR-20) Students Raw Score Reliability = 0.95					

*Note: Extreme students are those who received full points or 0 point.*

For the analysis of the 29 items in Table 7, there were no extreme cases implying that there was no item which all students correctly answered and no item which all students answered incorrectly. The summary descriptive statistics of the 29 items showed a mean of the logit measure to be 0.00, set arbitrarily by the Rasch model, with a standard deviation of 1.77. The placement of the items ranged from a minimum logit value of -4.13 to a maximum logit value of 3.16. On average, the data appear to fit the Rasch model with average infit mean-square (MNSQ) value of 1.01 and average outfit MNSQ value of 1.03. The outfit MNSQ being higher than 1.0 indicates that the data slightly underfit the model; there was less predictability and more variability in the data than what would be expected under the Rasch model (Bond & Fox, 2015). The root mean-square error (RMSE) was 0.15 and the separation index of the 29 items was extremely high at 11.71. This resulted in the item reliability of 0.99. Such reliability coefficients are not uncommon for item measures.

Table 7.  
Summary Statistics for 29 Measured (Non-Extreme) Items

	Raw Score	Model Count	Measure (logits)	MNSQ	
				<i>Infit</i>	<i>Outfit</i>
Mean	223.3	453.8	0.00	1.01	1.03
S.D.	88.9	7.5	1.77	0.15	0.44
Maximum	409.0	461.0	3.16	1.34	2.49
Minimum	70.0	446.0	-4.13	0.75	0.55
Real RMSE	= 0.15				
Separation Index	= 11.71		Item Reliability = 0.99		

*Note: Extreme items were questions in which all students received all or no points.*

Item Fit Analysis. Further analysis of the Outfit ZSTD provided us with a bubble map (*Figure 7*) and an output shown on Table 8. This bubble map only displayed the items and not the persons.

Given the bubble map (*Figure 7*) and z-fit statistics (Table 8) below, the result showed that Item 303, Item 313, Item 512, Item 514, Item 513, and Item 511 landed partially or completely outside of the “fit” zone. Item 303 and item 313 located above of the 2 ZSTD range. These items had more variability than expected suggesting that the additional variance might not be due to chance. In contrast, item 512, item 514, item 513, and item 511 located below the -2 ZSTD range. These four items exhibited more predictability and less variability than expected.

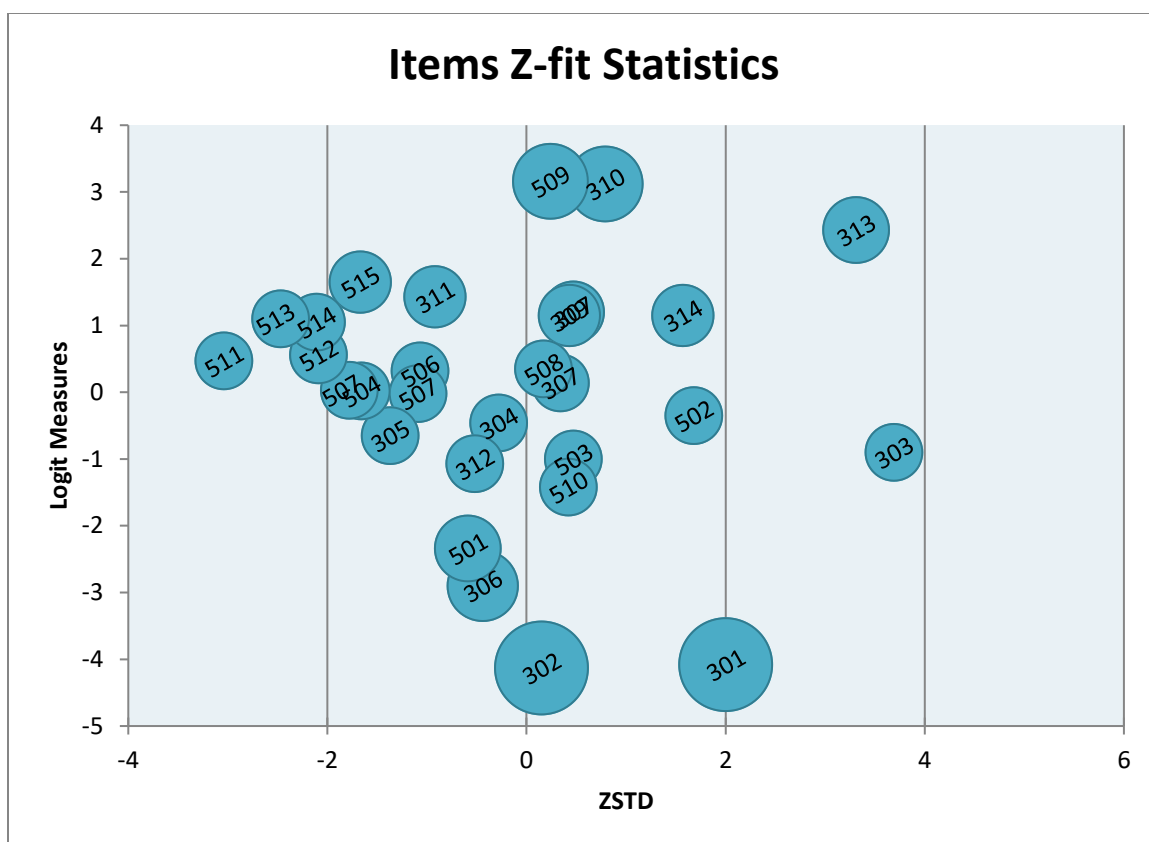


Figure 7. A bubble map of the items z-fit statistics for the 29 items.

Table 8.  
Summary of Z-fit Statistics of Items Falling Above and Below 2 Zstd

	<i>Logit Measures</i>	<i>ZSTD</i>	<i>SEM</i>
Item 303	-0.90	3.69	0.13
Item 313	2.43	3.31	0.15
Item 512	0.56	-2.09	0.13
Item 514	1.05	-2.11	0.13
Item 513	1.10	-2.47	0.13
Item 511	0.47	-3.04	0.13

An analysis of the point-measure correlation was performed to ensure that the response-level scoring makes sense. Negative observed correlation would indicate that something may have gone wrong. Negative observed correlation

would indicate that something may have gone wrong. The analysis of 463 students and 29 items showed that the observed point-measure correlation ranged from 0.34 to 0.73 and the expected point-measure correlation ranged from 0.43 to 0.66. The results showed that the observed and expected correlations were all positive indicating that the response-level score made sense.

#### Item Dimensionality Summary of Explained and Unexplained Variance.

Considering the high reliability value of 0.99 in the item descriptive analysis, further analysis of the variance factor of the data was performed. The analysis of the 463 students and 29 items looked at 100% of the variance in observations (Table 9). The variance in this data set comprised of 52.9% explained variance and 47.1% unexplained variance. The 52.9% of the variance explained by the Rasch model was partitioned into variance explained by the person (i.e., 29.4% of the total variance) and variance explained by the items (i.e., 23.5% of the total variance). Additionally, the first contrast of the unexplained variance was below 3 eigenvalues so there was not a need to explore other dimensions (Linacre, 2006).

The data set of this study had a higher percentage of explained variance, 52.9%, than the recommended value of 50% indicating that the data fit the Rasch model. As such, the variable map has a better capacity for predicting the ability level of the persons and the difficulty level of the items (Bond & Fox, 2015).

Table 9.  
Item Dimensionality Summary of 463 Students and 29 Items

	=	<u>Empirical</u>		Modeled
		$\lambda$	%	%
Total raw variance in observations	=	61.6	100.0%	100.0%
Raw variance explained by measures	=	32.6	52.9%	49.2%
Raw variance explained by persons	=	18.1	29.4%	27.4%
Raw variance explained by items	=	14.5	23.5%	21.9%
Raw unexplained variance (total)	=	29.0	47.1%	50.8%

*Note:* Table of Standardized Residual variance (in Eigenvalue units).

Differential Item Functioning Analysis. The purpose of this cross-sectional non-experimental quantitative study was to answer two research questions. The first question was to determine if DIF exists between English Learners (EL) and native English speakers on item formats other than MC and CR items. To accomplish this task, DIF Pairwise-Rasch-Welch analysis was used to determine DIF between EL and non EL students. A summary of the DIF analysis is showed below in Table 10 and a visual is shown in *Figure 8*. A correlation analysis between the DIF measures of EL and non-EL students showed that it has a Pearson correlation coefficient value of 0.98.

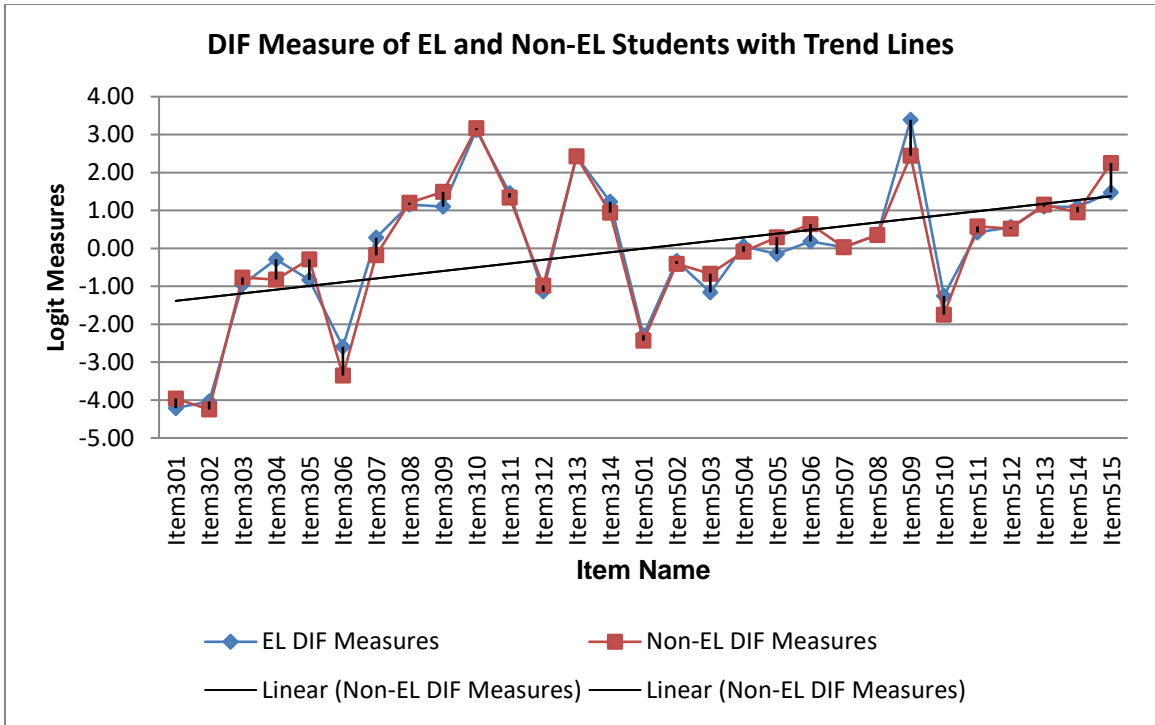


Figure 8. A graph of the DIF measures of EL and non-EL students with trend lines.

Table 10.  
Summary of DIF Analysis by EL Status (Rasch-Welch Analysis)

Item	DIF measures		DIF contrast	<i>t</i>	<i>df</i>	<i>p</i>
	EL	N-EL				
Item 301	-3.96	-4.21	0.24	0.58	369	0.562
Item 302	-4.25	-4.04	-0.21	-0.49	352	0.626
Item 303	-0.77	-0.96	0.19	0.68	320	0.495
Item 304	-0.82	-0.29	-0.53	-1.91	315	0.058
Item 305	-0.29	-0.83	0.54	1.91	317	0.057
<b>Item 306</b>	<b>-3.35</b>	<b>-2.59</b>	<b>-0.75</b>	<b>-2.27</b>	<b>332</b>	<b>0.024</b>
Item 307	-0.18	0.28	-0.46	-1.63	311	0.105
Item 308	1.20	1.15	0.06	0.18	298	0.858
Item 309	1.49	1.10	0.39	1.22	293	0.223
Item 310	3.17	3.12	0.05	0.12	275	0.906
Item 311	1.34	1.46	-0.12	-0.39	298	0.696
Item 312	-0.98	-1.13	0.15	0.51	322	0.609
Item 313	2.43	2.43	0.00	0.00	283	1.000
Item 314	0.94	1.23	-0.29	-0.98	303	0.329
Item 501	-2.43	-2.29	-0.14	-0.48	349	0.633
Item 502	-0.41	-0.33	-0.09	-0.32	323	0.749
Item 503	-0.67	-1.16	0.49	1.75	331	0.082
Item 504	-0.09	0.06	-0.16	-0.56	320	0.574
Item 505	0.29	-0.15	0.45	1.57	317	0.117
Item 506	0.64	0.18	0.46	1.58	312	0.115
Item 507	0.03	0.03	0.00	0.00	319	1.000
Item 508	0.35	0.35	0.00	0.00	316	1.000
<b>Item 509</b>	<b>2.44</b>	<b>3.39</b>	<b>-0.95</b>	<b>-2.54</b>	<b>314</b>	<b>0.012</b>
Item 510	-1.75	-1.25	-0.5	-1.80	332	0.073
Item 511	0.58	0.42	0.16	0.55	313	0.585
Item 512	0.52	0.56	-0.03	-0.11	314	0.910
Item 513	1.15	1.10	0.05	0.15	307	0.879
Item 514	0.95	1.09	-0.13	-0.45	310	0.655
<b>Item 515</b>	<b>2.25</b>	<b>1.47</b>	<b>0.78</b>	<b>2.26</b>	<b>286</b>	<b>0.025</b>

Note: EL = EL students, N-EL = non-EL students.

The analysis used the DIF contrast and Joint Standard Error (S.E.) to test the null hypothesis.

H<sub>0</sub>: There was no Differential Item Functioning between EL and non-EL students on non-MC and non-CR items.

H<sub>a</sub>: Differential Item Functioning existed between EL and non-EL students on non-MC and non-CR items.

DIF Pairwise-Rasch-Welch analysis determined that the result was statistically significant for three items: Item 306, Item 509, and Item 515. These three items are all of the Equation/Numeric (EQ) item type. The DIF Pairwise-Rasch-Welch analysis did not show any MC items to have DIF between EL and non-EL students.

The results showed that there was DIF between EL and non-EL students on some of the EQ item types. The result showed DIF exists between EL and non-EL students on Item 306 (DIF contrast = -0.75 logit,  $t(332) = -2.27$ ,  $p = 0.024$ ). The result showed DIF exists between EL and non-EL students on Item 509 (DIF contrast = -0.95,  $t(314) = -2.54$ ,  $p = 0.012$ ). The result showed DIF exists between EL and non-EL students on Item 515 (DIF contrast = 0.78,  $t(286) = 2.26$ ,  $p = 0.025$ ). Therefore, the study rejected the null hypothesis. There existed DIF between English Learners and native English speaker on item type other than MC and CR. In this case, DIF exists between EL and non-EL students on EQ items.



When placing the focal (EL) group against the reference group (non-EL) in the analysis (Table 11), there was negative DIF contrast for item 306 and item 509. This result indicated that item 306 (DIF contrast = -0.75) and item 509 (DIF contrast = -0.95) were 0.75 logits and 0.95 logits less difficult for EL students than non-EL students, respectively. Conversely, the DIF contrast was positive for item 515 (DIF contrast = 0.78) indicating that this item was 0.78 logits more difficult for EL students than non-EL students.

Table 11.  
Items Meeting or the Criteria to Reject the Null Hypothesis

	<i>Item Type</i>	<i>EL Status</i>		<i>DIF Contrast</i>	<i>Pairwise-Rasch-Welch</i>		
		<i>EL Measure</i>	<i>non-EL Measure</i>		<i>t</i>	<i>df</i>	<i>p</i>
Item 306	EQ	-3.35	-2.59	-0.75	-2.27	332	0.024
Item 509	EQ	2.44	3.39	-0.95	-2.54	314	0.012
Item 515	EQ	2.25	1.47	0.78	2.26	286	0.025

Note: EL = Focal Group, non-EL = Reference Group

### Research Question Two

Variable Map. To answer the second research question, the test data of 142 EL students were used for the analysis. Nineteen (19) students had an ELPAC level 1, twenty-seven (27) students had an ELPAC level 2, sixty-one (61) students had an ELPAC level 3, and forty (40) students had an ELPAC level 4.

Using the data of 142 students and 29 items, a variable map (*Figure 7*) was generated that provided the difficulty of each test item as compared to the students' ability (Bond & Fox, 2015). The scale for this map extends from -5

logits to 5 logits. The test items, organized by difficulty level (logit measure in descending order) with the most difficult items at the top of the list and the least difficult at the bottom, is presented in Appendix E. On the vertical scale, the mean difficulty of the Items was set at 0 while mean of the ability level of the students was located near -1 logit. The students were located between -5 and 5 logits while the items were located between -4.5 and 3.5 logits.

The mean of the participants' abilities was at the about 1 logit below the mean of the items difficulty indicating that, on average, the participants found the test to be more difficult. Furthermore, the items were generally well 'targeted' to the abilities of all the students. Mistargeting was not observed at the lower end of the scale but was observed at the upper end. There were no items that were well 'targeted' to the most able students. The ability level of the most able students in this analysis was based on the item below them. Also, Item 302 was located at the bottom most of the scale and no students on the other side of the scale. This means Item 302 was too easy for all of the students.

Item 310 was located at the top most of the scale meaning that it was the most difficult item out of the 29 items. Most items were distributed within one standard deviation on either side of the mean item calibration. The standard deviation for the item calibration was slightly smaller than the standard deviation of the student abilities.



Summary Statistics. The output for this analysis through WinSteps produced three tables of descriptive statistics. The first and second tables gave the summary statistics for measured students; the third table gave the summary statistics for measured items. WinSteps provides two types of analyses: (a) analyses with non-extreme students/items and (b) analyses with extreme students.

Table 12 shows the scores of non-extreme students. The mean logit measures of student abilities was -0.84 logits and the standard deviation was 2.10 logits. The ability measures ranged over 8.47 logits from a maximum value of 4.41 logits to a minimum value of -4.06 logits. On the average the data appear to fit the Rasch model with average infit mean-square (MNSQ) value of 0.98 and average outfit MNSQ value of 0.89. These values were close to the expected mean-square value of 1.0. The outfit MNSQ being lower than 1.0 indicated that the data slightly overfit the model; there was more predictability and less variability in the data than expected under the Rasch model (Bond & Fox, 2015).

Table 12.  
Summary Statistics for 140 Measured (Non-Extreme) Students

	Raw Score	Model Count	Measure (logits)	MNSQ	
				Infit	Outfit
Mean	11.2	28.6	-0.84	0.98	0.89
S.D.	7.7	2.3	2.10	0.34	0.74
Maximum	28.0	29.0	4.41	2.73	5.27
Minimum	1.0	15.0	-4.06	0.35	0.10
Real RMSE	= 0.66				
Separation Index	= 3.01		Student Reliability = 0.90		

*Note: Extreme students were those who received full points or 0 point.*

*Maximum extreme score: 12 students*

*Valid responses: 98.5%*

The output in Table 13 analyzed all students, extreme and non-extreme showed the mean of the Rasch logit measure was -0.74 logits, the standard deviation was 2.23 logits, with measure ranging from a maximum of 5.70 logits to a minimum of -4.06 logits. The root mean-square error (RMSE) was 0.69 and the separation index was 3.05. Based on the Rasch measures, the student reliability coefficient was 0.90 and Cronbach Alpha (KR-20) value was 0.94. The Cronbach alpha value of 0.94 indicated that the scores were highly reliable (Traub & Rowley, 1991). This means that there was high probability that students estimated with high measures actually did have high measures and students estimated with low measures actually did have low measures. Furthermore, a Cronbach alpha value of 0.94 can also be interpreted as 94 percent of the observed variance in scores was associated systematic differences in the performances of the students and 6 percent was errors (Traub & Rowley, 1991).

Table 13.  
Summary Statistics for 142 Measured (Extreme and Non-Extreme) Students

	Raw Score	Model Count	Measure (logits)	MNSQ	
				Infit	Outfit
Mean	11.4	28.6	-0.74		
S.D.	7.9	2.3	2.23		
Maximum	29.0	29.0	5.70		
Minimum	1.0	15.0	-4.06		
Real RMSE = 0.69					
Separation Index = 3.05			Student Reliability = 0.90		
Cronbach Alpha (KR-20) Students Raw Score Reliability = 0.94					

*Note: Extreme students were those who received full points or 0 point.*

For the analysis of the 29 items in Table 14, there were no extreme cases implying that there was no item which all students correctly answered and no item which all students answered incorrectly. The summary descriptive statistics of the 29 items showed a mean of the logit measure to be 0.00, set arbitrarily by the Rasch model, with a standard deviation of 1.83. The placement of the items ranges from a minimum logit value of -4.32 to a maximum logit value of 3.20. On average, the data appear to fit the Rasch model with average infit mean-square (MNSQ) value of 1.01 and average outfit MNSQ value of 0.90. The outfit MNSQ being lower than 1.0 indicated that the data slightly overfit the model; there was more predictability and less variability in the data than what would be expect under the Rasch model (Bond & Fox, 2015). The root mean-square error (RMSE) was 0.27 and the separation index of the 29 items was high at 6.60. This resulted in the item reliability of 0.98. Such reliability coefficients are not uncommon for item measures.

Table 14.  
Summary Statistics for 29 Measured (Non-Extreme) Items

	Raw Score	Model Count	Measure (logits)	MNSQ	
				<i>Infit</i>	<i>Outfit</i>
Mean	55.9	140.1	0.00	1.01	0.90
S.D.	29.6	2.0	1.83	0.17	0.38
Maximum	125.0	142.0	3.20	1.43	2.04
Minimum	13.0	138.0	-4.32	0.75	0.52
Real RMSE	= 0.27				
Separation Index	= 6.60		Item Reliability = 0.98		

*Note: Extreme items were questions in which all students received all or no points.*

Item Fit Analysis. Further analysis of the Outfit ZSTD provided us with a bubble map (*Figure 8*) and an output shown on Table 15. This bubble map only displayed the items and not the persons.

Given the bubble map (*Figure 7*) and z-fit statistics (Table 15) below, the result showed that Item 303 completely outside of the “fit” zone. Item 303 located above of the 2 ZSTD range indicating that it had too much variability and the variance might not be due to chance. No item fell below the -2 ZSTD range.

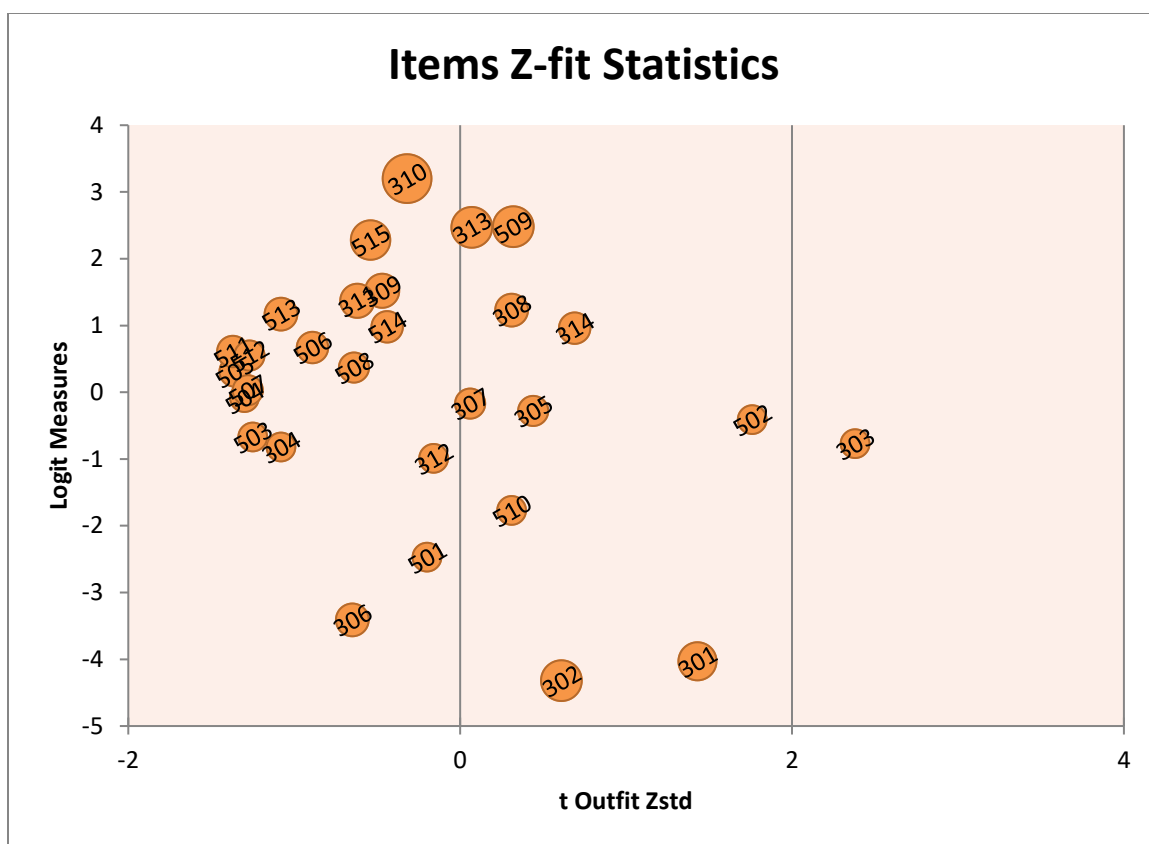


Figure 10. A bubble map of the items z-fit statistics for the 29 items.

Table 15.  
Summary of Z-fit Statistics of Items Falling Above and Below 2 Zstd

	<i>Logit Measures</i>	<i>ZSTD</i>	<i>SEM</i>
Item 303	-0.77	2.38	0.23

An analysis of the point-measure correlation was performed to ensure that the response-level scoring makes sense. Negative observed correlation would indicate that something may have gone wrong. Negative observed correlation would indicate that something may have gone wrong. The analysis of 142 students and 29 items showed that the observed point-measure correlation



ranged from 0.26 to 0.71 and the expected point-measure correlation ranged from 0.37 to 0.66. The results showed that the observed and expected correlations were all positive indicating that the response-level score made sense.

#### Item Dimensionality Summary of Explained and Unexplained Variance.

Considering the high reliability value of 0.98 in the item descriptive analysis, we further looked into analyzing the variance factor of the data. The analysis of the 142 students and 29 items looked at 100% of the variance in observations (Table 16). The variance in this data set comprised of 52.9% explained variance and 47.1% unexplained variance. The 52.9% of the variance explained by the Rasch model was partitioned into variance explained by the person (i.e., 28% of the total variance) and variance explained by the items (i.e., 24.8% of the total variance). Additionally, the first contrast of the unexplained variance was below 3 eigenvalues so there was not a need to explore other dimensions (Linacre, 2006).

The data set of this study had a higher percentage of explained variance, 52.9%, than the recommended value of 50% (Bond & Fox, 2015) indicating that the data fit the Rasch model. As such, the variable map has a better capacity for predicting the ability level of the persons and the difficulty level of the items.

Table 16.  
Item Dimensionality Summary of 142 Students and 29 Items

		Empirical		Modeled
		$\lambda$	%	%
Total raw variance in observations	=	61.5	100.0%	100.0%
Raw variance explained by measures	=	32.5	52.9%	49.5%
Raw variance explained by persons	=	17.3	28.0%	26.2%
Raw variance explained by items	=	15.3	24.8%	23.2%
Raw unexplained variance (total)	=	29.0	47.1%	50.5%

Note: Table of Standardized Residual variance (in Eigenvalue units).

Differential Item Functioning Analysis. The purpose of this cross-sectional non-experimental quantitative study was to answer two research questions. The second of which was to determine if DIF exists among students with different levels of English proficiency as determined by the ELPAC on item formats other than MC and CR items. A t-test on DIF size was used to determine DIF existed among students with different ELPAC level. *Figure 9* provided a graph of the DIF measures of students with different ELPAC level.

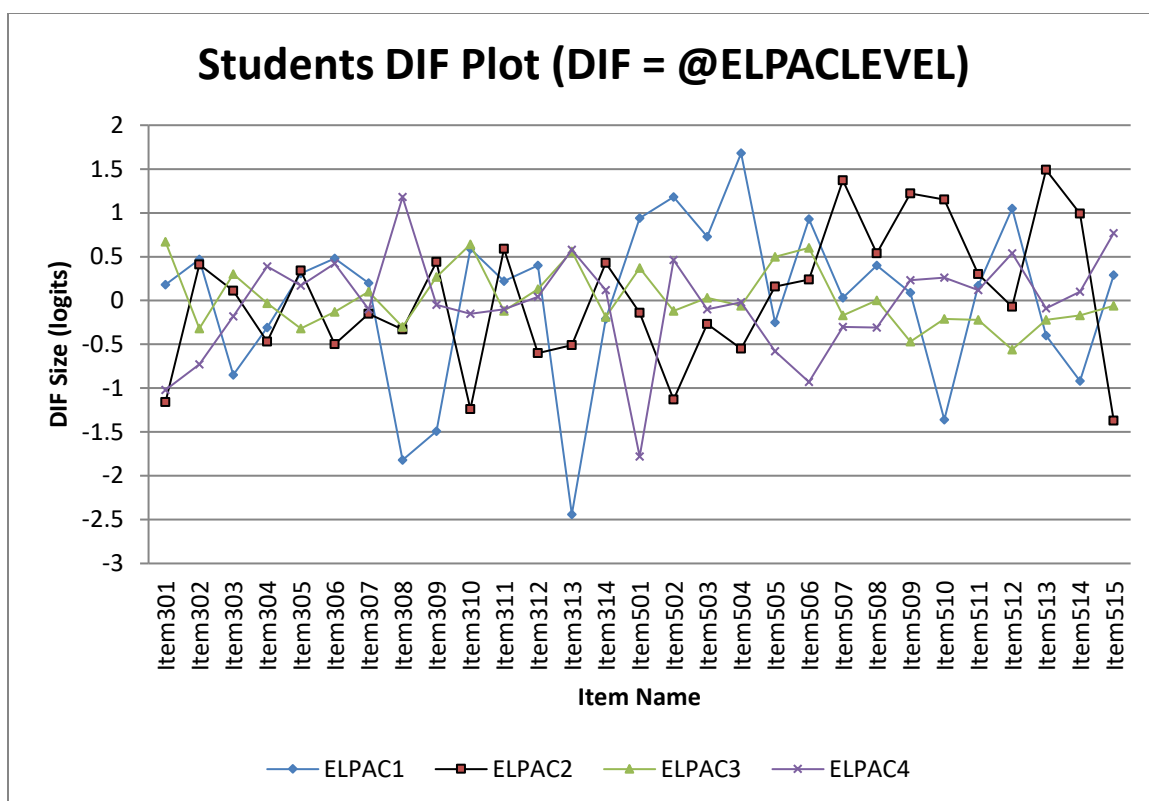


Figure 11. A graph of the DIF size for student with different ELPAC level.

Using the DIF size, WinSteps used the t-test to test the null hypothesis.

$H_0$ : There was no Differential Item Functioning among students with different ELPAC levels on non-MC and non-CR items.

$H_a$ : Differential Item Functioning occurred among students with different ELPAC levels on non-MC and non-CR items.

The T-test analysis was statistically significant for five items: Item 308 (G), Item 313 (MC &G), Item 501 (EQ), item 502 (EQ), and Item 510 (MC). The t-test

analysis showed that DIF existed among students with different ELPAC level on MC, G, and EQ item types.

Table 17.  
Summary of Items Meeting the Criteria to Reject the Null Hypothesis

	Item Type	Observed Average	Expected Average	DIF Score	DIF Measure	DIF Size	<i>p</i>
Item 308	G						
1		0.38	0.21	0.16	-0.59	-1.82	0.031
2		0.20	0.17	0.03	0.90	-0.33	0.633
3		0.26	0.23	0.03	0.93	-0.30	0.443
4		0.16	0.29	-0.14	2.41	1.18	0.033
Item 313	MC & G						
1		0.31	0.13	0.18	0.03	-2.44	0.010
2		0.12	0.09	0.03	1.96	-0.51	0.522
3		0.09	0.12	-0.03	3.03	0.56	0.335
4		0.11	0.15	-0.05	3.05	0.58	0.351
Item 501	EQ						
1		0.47	0.61	-0.13	-1.53	0.94	0.179
2		0.62	0.60	0.02	-2.61	-0.14	0.791
3		0.66	0.71	-0.05	-2.10	0.37	0.294
4		0.95	0.82	0.13	-4.25	-1.78	0.030
Item 502	EQ						
1		0.24	0.34	-0.10	0.77	1.18	0.200
2		0.46	0.32	0.14	-1.53	-1.13	0.044
3		0.44	0.43	0.02	-0.52	-0.12	0.744
4		0.47	0.55	-0.07	0.06	0.46	0.258

Note: Students with ELPAC score of 1 = 1, ELPAC2 = 2, ELPAC3 = 3, ELPAC4 = 4.

Table 18 showed that there was DIF among students with different ELPAC level on Item 308 for ELPAC 1 (DIF size = -1.82, *p* = 0.031) and ELPAC4 (DIF size = 1.18, *p* = 0.033), on Item 313 for ELPAC1 (DIF size = -2.44, *p* = 0.010), on Item 501 for ELPAC4 (DIF size = -1.78, *p* = 0.030), on Item 502 for ELPAC2 (DIF size = -1.13, *p* = 0.044), on Item 510 for ELPAC1 (DIF size = -1.36, *p* = 0.047).

ELPAC group with big enough DIF size and  $p < 0.05$  on any particular item would mean that the item was significantly biased against that group. Therefore, this study rejects the null hypothesis. There exists DIF among students of different ELPAC level on item type other than MC and CR. In this case, DIF existed among students with different ELPAC level on Equation/Numeric (EQ) and Graphing (G) item type. The complete table can be found on Appendix F.

Comparing the DIF score of students with different ELPAC level (Table 17) would show if the item was more or less difficult for students of a particular ELPAC level. DIF score positive means that the item was easier. The DIF size would reveal the logit difference. The result for ELPAC1 on item 308 (DIF score = 0.16, DIF size -1.82) showed that the item less difficult for ELPAC1 students than expected and that it was less difficult by 1.82 logits. For ELPAC4 on item 308 (DIF score = -0.14, DIF size = 1.18), the question was more difficult by 1.18 logits. ELPAC1 on item 313 (DIF score = 0.18, DIF size = -2.44), the question was less difficult by 2.44 logits. ELPAC4 on item 501 (DIF score = 0.13, DIF size = -1.78), the question was less difficult by 1.78 logits. For ELPAC2 on item 502 (DIF score = 0.14, DIF size = -1.13), the question was less difficult by 1.13 logits.

A correlation analysis between the DIF measures of students with different ELPAC level showed that the correlation coefficient between ELPAC1 and ELPAC2 was the lowest at 0.77 while the correlation coefficient between ELPAC3 and ELPAC4 was highest at 0.93. The full summary is shown below on Table 18.

Table 18.  
Correlational Analysis of Item Measures for Different ELPAC Level

Categories		<i>Pearson</i> <i>r</i>
Data Set 1	Data Set 2	
ELPAC1	ELPAC2	0.77
ELPAC1	ELPAC3	0.85
ELPAC1	ELPAC4	0.81
ELPAC2	ELPAC3	0.88
ELPAC2	ELPAC4	0.88
ELPAC3	ELPAC4	0.93

### Summary

The purpose of this cross-sectional non-experimental quantitative study was to determine if DIF occurred between EL and non-EL students and if DIF existed among students with different levels of English language proficiency as determined by the ELPAC. WinSteps was used to generate a variable map for all of the students and EL students only, respectively, to show the difficulty of the items and the ability level of the students. Summary statistics tables for the persons and items were generated and analyzed for separation value and Cronbach Alpha for all students and for EL students exclusively. Furthermore, item z-fit statistics report was used to analyze for overfitting and underfitting items for all students test data and for EL students test data. Item dimensionality report was looked at for explained and unexplained variance of the items and persons. DIF Pairwise-Rasch-Welch analysis and t-test were used to find if DIF

occurred between EL and non-EL students and among students with different ELPAC level for any of the test items, respectively.

The study found that DIF did occur between EL and non-EL students on items other than MC and CR. Furthermore, the study found that DIF also existed among students with different ELPAC level on items other than MC and CR.

## CHAPTER FIVE

### CONCLUSIONS AND RECOMMENDATIONS

#### Introduction

This study was a cross-sectional non-experimental study aimed to understand Differential Item Functioning in items that are not MC or CR between EL and non-EL students and among students with different ELPAC levels. WinSteps (Linacre, 2006) was used to run the analysis. The program is able to transform raw scores that are ordinal measure to Rasch measure that are along an interval scale (Bond & Fox, 2015). Furthermore, WinSteps was able to produce variable maps in which the item difficulties were placed on the same scale as the ability measures of the students.

#### Limitations of Study

This study had two major limitations. The first limitation was the size of the target samples, mainly, the size of the EL students at each ELPAC level. With sample sizes of or above 30, the violation of the normality assumption should not cause major problems given that the Central Limit Theorem takes effect at  $n = 30$  (Elliott & Woodward, 2007; Pallant, 2007). Given that there were 19 ELPAC1 students and 27 ELPAC2 students, the results of the study should not be generalized to the population of ELPAC1 and ELPAC2 students.

The second limitation of the study was the limited number of item types presented in the two chapters of the ALEKS program selected for the study (i.e.,



chapters 3 and 5). The ALEKS program has only six different item types but chapter 3 and 5 included only three item types. These item types were graphing (G), multiple choice with single correct response (MC), and Equation/Numeric (EQ). The CAASPP tests developed in response to the CCSSM (2010) as the state assessment for California has nine different item type, namely, multiple choice with single correct response (MC), multiple choice with multiple correct responses (MSMC), matching tables (MA), short text (CR), drag and drop (DD), hot spot (HS), table fill in (TI), graphing (G), and equation/numeric (EQ) that are currently used on the CAASPP test (*Smarter Balanced Question Types*, 2018).

The purpose of this study was to look at DIF in item types other than MC and CR. Given the above limitations, the study included an examination of DIF in Graphing items and Equation/Numeric items. The study was not able to gain any insight on multiple choice with multiple correct responses (MSMC) items, matching tables (MA) items, or table fill in (TI) items.

### Characteristics of Assessments

The variable maps (Figure 6 and 7) showed that the test items were generally the same level of difficulty as the ability level of all the students while being about 1 logit more difficult for the EL students. Furthermore, the variable maps showed that the tests were not highly-targeted. A well-targeted instrument “has a distribution of items that matches the range of the test candidates’ abilities. Ideally, the mean and *SDs* of items and persons would match closely” (Bond & Fox, 2015, p. 372). The average difficulty of the items did not align with

the average ability level of the students (Figure 7) and the range of the measures of the items did not match the range of the measures of the students (Figure 6 and 7).

The item fit analysis of 463 students and 29 items showed that there were six misfitting items: two underfitting (Item 301 and 302) and four overfitting items (Item 511, 512, 513, and 514). Misfitting items either had too much variability (underfit) or too little variability (overfit) to fit the Rasch model (Linacre, 2006). Item 301 (-4.08 logits) and 302 (-4.13 logits) were the easiest items. Item 511 (0.47 logits), 512 (0.56 logits), 513 (1.10 logits) and 514 (1.05 logits) were around half of a standard deviation above the mean difficulty of the items.

The item fit analysis of 142 students and 29 items showed only one underfitting item (Item 303) and no overfitting items. Item 303 (-0.77 logits) was about half of a logit below the mean difficulty of the items and was in line with the mean ability of the students. It was the eighth easiest item.

The presence of the misfitting items in the two analyses could potentially point to the distortion of the unidimensionality assumption of the instrument. Measures fitting the Rasch model must be unidimensional (Bond & Fox, 2015). As such, there might have been more than one underlying latent trait distorting the assumption of unidimensionality of these items, thus affecting students' performance. These potential underlying traits could possibly be the presence of mathematics self-efficacy (Betz & Hackett, 1983), mathematics anxiety (Richardson & Suinn, 1972), and/or mathematical mindset (Rattan et al., 2012) in

the students. An in-depth look at the misfitting items alongside classroom cultures and teaching pedagogies, as a recommendation for future research, might be able to shed some lights into this issue.

## Interpretation of Results

### Research Question One

Although the analysis of 463 students and 29 items showed that there were no items that were 'targeted' to the most able students and to the least able students, the students' scores (separation index = 3.13 and Rash reliability coefficient = 0.91) and the items (separation index = 11.71 and Rash reliability coefficient = 0.99) were highly reliable. High reliability of students and items in this study meant that there was a high probability that students and items estimated with high measures did have higher measures than students and items estimated with low measures (Linacre, 2006). Furthermore, the explained variance in the observation was 52.9%; an explained variance of less than 50% would be cause for concern (Bond & Fox, 2015).

The analysis showed that DIF occurred between EL and non-EL students for Equation/Numeric (EQ) items and not for Graphing items (G). Out of the 29 items, 24 of them were EQ item type. Only three of those EQ items had DIF between EL and non-EL students. That meant 12.5% of the items showed presence of DIF. The three items were Items 306, 509, and 515. Item 306 (DIF contrast = -0.75) and Item 509 (DIF contrast = -0.95) were easier for EL students

than non-EL students while Item 515 (DIF contrast = 0.78) was harder for the EL students than the non-EL students.

An examination of Item 306 (Figure 10) showed that the problem had minimal amount of verbal text. Thus, language complexity was not likely to have been a contributing factor to item bias, negatively impacting scores for EL students (Abedi & Lord, 2001). Students could have gotten the answer to this problem by locating the point of intersection between the graph and the two axes. With the DIF contrast of this item to be -0.75, EL students found this problem to be 0.75 logits less difficult than non-EL students.

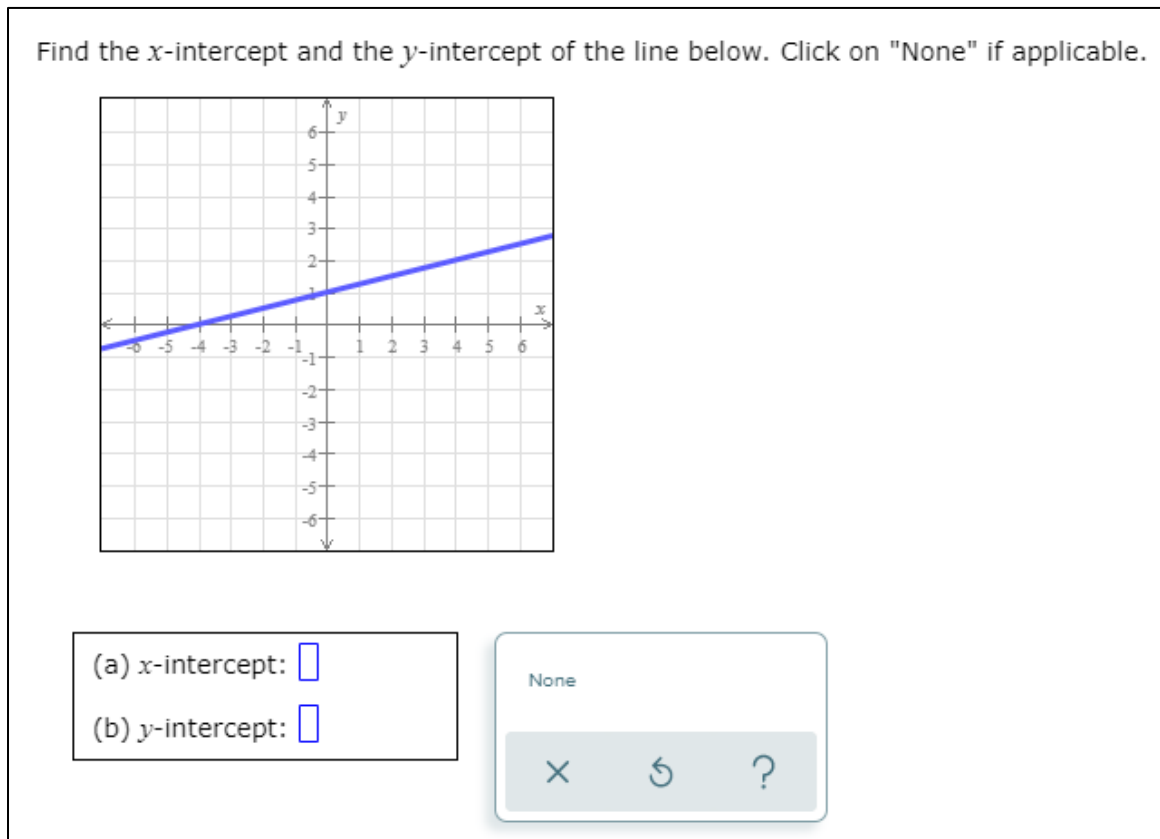


Figure 10. Item 306 – the 6<sup>th</sup> question in chapter 3 test.

In stark contrast to Item 306, Item 509 (Figure 11) was considered a word problem due to the amount of language present. Abedi and Lord (2001) would consider the presence of text in this problem to be an inherent contributor to testing bias for EL students. However, according to the analysis of DIF contrast (-0.95), EL students found this problem to be almost 1 logit less difficult than non-EL students.


A triangular pyramid is formed from three right triangles as shown below. Use the information given in the figure to find the length  $SU$ . If applicable, round your answer to the nearest whole number. The lengths on the figure are not drawn accurately.

Figure 11. Item 509 – the 9<sup>th</sup> question in chapter 5 test.

The language presence in Item 515 (Figure 12) was very similar to Item 306. Students could have gotten the answer by either calculating the distance using the distance formula or counting the units on the provided graph. With

minimal amount of verbal text present in the problem, language complexity was not likely to have been a contributing factor to item bias, negatively impacting scores for EL students (Abedi & Lord, 2001). However, unlike Item 306, EL students found Item 515, with DIF contrast being 0.78, to be 0.78 logits more difficult than non-EL students.

Calculate the distance between the points  $B = (-9, 9)$  and  $L = (-1, 1)$  in the coordinate plane. Round your answer to the nearest hundredth.



Distance:

Figure 12. Item 515 – the 15<sup>th</sup> question in chapter 5 test.

Looking at the three items that had DIF, there does not seem to exist any commonalities as to why DIF might have occurred between EL and non-EL students. Theoretically, the hardest of the three items with the presence of DIF for EL students should have been Item 509 due to the large presence of verbal text; yet, the results showed otherwise. Additionally, while the language presence for Item 306 and 515 were similar, EL students found the former to be easier and the latter to be harder than non-EL students.

An in-depth analysis of the items showed that DIF on these items might have been random rather than systematic. As such, it is recommended that a more systematic choice of items should be used for future studies to examine DIF between EL and non-EL students.

### Research Question Two

The analysis done for the second research question examined the test data of EL students (n = 142). The data from the non-ELs were not included in the analysis. Both the students' scores (separation index = 3.05 and Rasch reliability coefficient = 0.90) and the items (separation index = 6.60 and Rasch reliability coefficient = 0.98) were highly reliable. High reliability of students and items in this study meant that there was a high probability that students and items estimated with high measures did have higher measures than students and items estimated with low measures (Linacre, 2006). Furthermore, the explained variance in this analysis was at 52.9%.

General observation should be that the expected average measures of all the items should be in ascending order from ELPAC1 to ELPAC4. That is to say, ELPAC4 students were expected to have the highest score follow by ELPAC3, ELPAC2, and ELPAC1. However, the results showed that the observed average measure for ELPAC2 was lower than the observed average measure of ELPAC1 students on sixteen of the twenty-nine items (see Appendix F). More than 55% of the items had this pattern.

The t-test of DIF size showed that DIF existed among students with different ELPAC levels for Equation/Numeric (EQ) and Graphing (G) items. Out of the twenty-nine items in the analysis, twenty-four of items were Equation/Numeric type and five were Graphing type. Two of the Equation/Numeric items and two of the Graphing items showed DIF among students with different ELPAC levels. The two Graphing items were Item 308 and 313. The two Equation/Numeric items were Item 501 and 502. That is 8.3% of Equation/Numeric items and 40% of Graphing items showed presence of DIF among students with different ELPAC levels.

For Item 308 (Figure 13), ELPAC1 (observed average = 0.38, expected average 0.21) students scored higher than expected while ELPAC4 (observed average = 0.16, expected average = 0.29) students scored worse than expected. Furthermore, ELPAC1 students scored higher than ELPAC4 students who had a much higher level of English proficiency than the ELPAC1 students. The problem is solved by the graphing methods. The analysis showed that ELPAC1 students scored the highest among all four levels of the EL students. Proficiency with the English language did not seem to have an impact on their performance.



Graph the line with slope  $-\frac{2}{3}$  passing through the point  $(5, -4)$ .

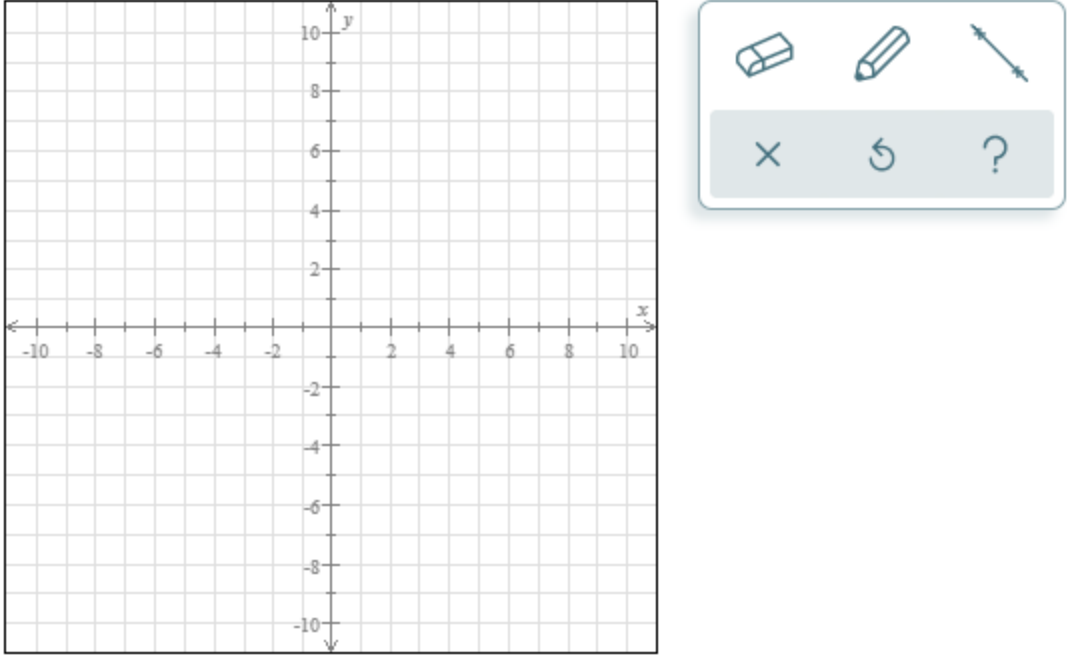


Figure 13. Item 308 – the 8<sup>th</sup> question in chapter 3 test.

For Item 313 (Figure 14), ELPAC1 (observed average = 0.31, expected average = 0.13) students scored higher than expected. In fact, ELPAC 1 students scored highest of all of the EL students. The problem is solved by graphing the lines and identifying the point of intersection. Proficiency with the English language did not seem to have an impact on their performance.

Here is a system of equations.

$$\begin{cases} y = 3x - 1 \\ y = -x - 5 \end{cases}$$

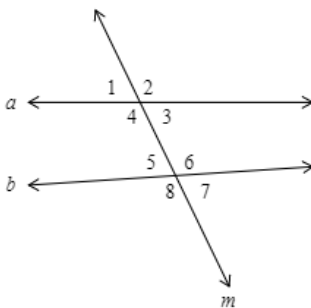
Graph the system. Then write its solution. Note that you can also answer "No solution" or "Infinitely many" solutions.

Figure 14. Item 313 – the 13<sup>th</sup> question in chapter 3 test.

For Item 501 (Figure 15), the observed average showed that students with higher ELPAC level did score higher than students with lower ELPAC level. Furthermore, ELPAC4 (observed average = 0.94, expected average = 0.82) students scored higher than expected; the DIF size on this item was -1.78 logits. To answer this question, students needed to identify appropriate angles meeting the requirements specified by the vocabulary terms. Students with lower ELPAC level scored lower than students with higher ELPAC level. The presence of the

vocabulary words could have been a potential contributor to the overall performance (bias) of the EL students (Abedi et al., 2000).

Give a pair of alternate exterior angles, a pair of alternate interior angles, and a pair of corresponding angles.



(a) Alternate exterior angles:  $\angle$   and  $\angle$

(b) Alternate interior angles:  $\angle$   and  $\angle$

(c) Corresponding angles:  $\angle$   and  $\angle$

Figure 15. Item 501 – the 1<sup>st</sup> question in chapter 5 test.

For Item 502 (Figure 16), it was ELPAC2 (observed average = 0.46, expected average = 0.32) students who scored much higher than expected; the DIF size on this item was -1.13. The observed score of 0.46 was the second highest observed average for the students with different ELPAC levels. This item was an Equation/Numeric item in which students had to solve for  $x$  and  $y$  and input the results. The problem had minimal amount of verbal text. Thus, language complexity was not likely to have been a contributing factor to item bias, negatively impacting scores for EL students (Abedi & Lord, 2001). Thus, proficiency with the English language should not have impacted the overall performance of ELPAC2 students on this item.

In the figure below,  $l \parallel o$ . Find the values of  $y$  and  $x$ .

$y = \square$   
 $x = \square$

Figure 16. Item 502 – the 2<sup>nd</sup> question in chapter 5 test.

The analysis of the four items that had DIF among students with different ELPAC levels revealed that ELPAC1 students performed better on Graphing (G) items than the rest of the EL students. ELPAC4 students had better than expected result in the Equation/Numeric (Item 501) problem than the other EL students; this result could have been a result of the language complexity present in the problem. Furthermore, ELPAC2 students performed better than expected on Item 502; however, the observed average on this item was similar to that of ELPAC3 and ELPAC4 students. This could potentially have been due to random error or small sample size.

## Recommendations for Educational Leaders

Educational leaders should desire to create balanced assessments by having the difficulties of the items match the abilities of all the students. The variable maps showed that the assessments were not well-targeted. It is recommended that assessments leaders consider adding items of appropriate difficulty level to the assessments so that they can be more balanced and more well-targeted.

Furthermore, assessment leaders should consider taking a second look at the misfitting items and decide whether to keep or remove from future assessments as there might be more than one underlying latent trait distorting the assumption of unidimensionality of these items and affecting students' performance. Since the item fit analysis of all the students showed different misfitting items than those of only EL students, the recommendation would be to create assessments for EL students separate from the assessments for non-EL students. Using this approach, modifications to the assessments can be done for each group individually to fit the needs of each group.

Additionally, it is recommended that teachers be made aware of potential DIF across test items and that they practice the routine usage of testing accommodations for EL students on assessments that are appropriate to their ELPAC level thus reducing the potential for DIF.

Educational leaders should encourage their teachers to provide appropriate accommodations based on individual students' academic

background; English proficiency is one of the most important criteria in selecting appropriate accommodation (Abedi, 2014). Additionally, any accommodations which reduces and/or eliminate testing bias should enhance the validity of the assessment results (Abedi & Lord, 2001). Possible accommodations for EL students, as long as it's appropriate to the students' academic background, could be: allowing students access to the Spanish version of the assessment (Hofstetter, 2003), providing dual language (stacked translation) test (Duncan et al., 2005), providing students with a modified language version of the assessment (Abedi et al., 2000; Abedi & Lord, 2001), providing EL students with an option for text-to-speech (Kopriva et al., 2007), and allowing EL students to get access to a glossary (Abedi et al., 2000). The effect of DIF on items with language complexity can be further reduced if students are to be provided with nonlinguistic schematic representations that would assist them in making meaning of the text (Martiniello, 2009)

Ultimately, any accommodation(s), when use correctly, would be most effective if students are also given extra time to take advantage of the accommodations (Abedi, 2014; Abedi et al., 2000, 2004; Duncan et al., 2005; Pennock-Roman & Rivera, 2011).

#### Recommendations for Future Research

There are two recommendations for future research: (1) use a larger sample size of EL students, especially ELPAC1 and ELPAC2 students and (2)

use more chapter tests to incorporate more item formats that might potentially be problematic to EL students.

The larger population would allow for more generalizability of the results. There are two different approaches in which the sample size can be increased for this type of study. The first approach is to expand the population to include 6<sup>th</sup> and 7<sup>th</sup> graders. While the students in each grade level learn different content and take different assessments, the use of linking questions can be employed. That is to say, use common items on the assessments that students from all three grade levels can take. WinSteps analyses can be used to determine DIF occurs between EL and non-EL, between grade levels, and among ELPAC level.

The second approach to getting a larger population for the study would be to standardize assessments across the same grade level in the district. Representatives from each site would meet to come up with a common assessment, common items, that all sites would use. This way, the sites will not need to modify the assessments from its original templates. If this can be done, then testing data can be collected for all 8<sup>th</sup> graders in the district rather than 8<sup>th</sup> graders from only one school site.

The second recommendation for future research would be to use more chapter tests data in the analysis, especially tests that have diverse item types. Having more diverse item types would allow for a more thorough DIF analysis and potentially discovering more items types that might be problematic for EL students.

## Conclusion

This cross-sectional non-experimental quantitative study was to determine if DIF occurred between EL and non-EL students on test items that were not MC or CR types. Furthermore, the study was designed to explore whether DIF existed between students with different ELPAC level on test items other than MC and CR types.

The analysis of the 463 students and 29 items showed that DIF did occur between EL and non-EL students on Equation/Numeric (EQ) item type. DIF was identified in three of the twenty-four EQ test items. In depth analysis of the items did not reveal any systematic commonalities as to why DIF would have occurred for those problems. EL students should have done worse than non-EL students on problem with language complexity (Item 509) but did not. On items where language complexity was not present, EL students had mixed results.

The analysis of the 142 EL students and 29 items showed that DIF existed among students with different ELPAC level. Of the items in which there was DIF among the student with different ELPAC level, ELPAC1 students outperformed other EL students on Graphing (G) type items while ELPAC4 students outperformed the rest of the EL students on problem with language complexity.



APPENDIX A  
INSTITUTIONAL REVIEW BOARD APPROVAL LETTER



March 30, 2020

**CSUSB INSTITUTIONAL REVIEW BOARD**

Administrative/Exempt Review Determination

Status: Determined Exempt

IRB-FY2020-291

Mr. Michael Nguyen and Prof. Joseph Jesunathadas  
Doctoral Studies Program and Department of Teacher Education & Foundation  
California State University, San Bernardino  
5500 University Parkway  
San Bernardino, California 92407

Dear Mr. Nguyen and Prof. Jesunathadas:

Your application to use human subjects, titled "Differential Item Functioning of English Learners on Item Format other than Constructed Response (CR) and Multiple Choice (MC) Items" has been reviewed and approved by the Chair of the Institutional Review Board (IRB) of CSU, San Bernardino has determined your application meets the federal requirements for exempt status under 45 CFR 46.104. The CSUSB IRB has not evaluated your proposal for scientific merit, except to weigh the risk and benefits of the study to ensure the protection of human participants. The exempt determination does not replace any departmental or additional approvals which may be required.

You are required to notify the IRB of the following as mandated by the Office of Human Research Protections (OHRP) federal regulations 45 CFR 46 and CSUSB IRB policy. The forms (modification, renewal, unanticipated/adverse event, study closure) are located in the Cayuse IRB System with instructions provided on the IRB Applications, Forms, and Submission webpage. Failure to notify the IRB of the following requirements may result in disciplinary action.

- Ensure your CITI Human Subjects Training is kept up-to-date and current throughout the study
- Submit a protocol modification (change) if any changes (no matter how minor) are proposed in your study for review and approval by the IRB before being implemented in your study.
- Notify the IRB within 5 days of any unanticipated or adverse events are experienced by subjects during your research.
- Submit a study closure through the Cayuse IRB submission system once your study has ended.

If you have any questions regarding the IRB decision, please contact Michael Gillespie, the Research Compliance Officer. Mr. Michael Gillespie can be reached by phone at (909) 537-7588, by fax at (909) 537-7028, or by email at [mgillesp@csusb.edu](mailto:mgillesp@csusb.edu). Please include your application approval number IRB-FY2020-291 in all correspondence. Any complaints you receive from participants and/or others related to your research may be directed to Mr. Gillespie.

Best of luck with your research.

Sincerely,

*Donna Garcia*

Donna Garcia, Ph.D., IRB Chair  
CSUSB Institutional Review Board

DG/MG

APPENDIX B  
CHAPTER 3 TEST

**Question 1 of 14**

For each table, determine whether it shows that  $x$  and  $y$  are proportional.

If  $x$  and  $y$  are proportional, fill in the blank with a number in simplest form.

Table 1				Table 2			
$x$	5	8	11	$x$	4	6	8
$y$	10	32	66	$y$	16	24	32
<input type="radio"/> Proportional				<input type="radio"/> Proportional			
y is ____ times x				y is ____ times x			
<input type="radio"/> Not proportional				<input type="radio"/> Not proportional			

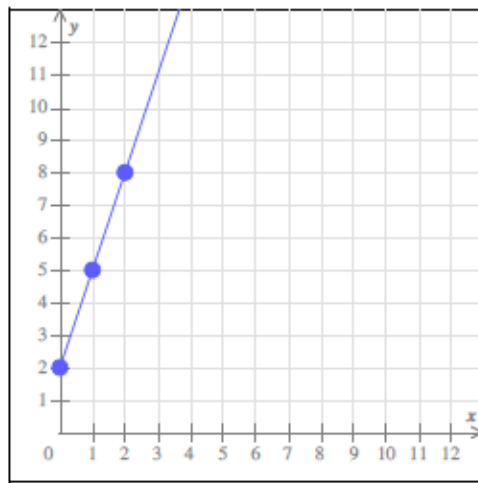
**Question 2 of 14**

Each graph below shows a relationship between  $x$  and  $y$ .

For each graph, determine whether  $x$  and  $y$  are proportional.

If  $x$  and  $y$  are proportional, fill in the blank with a number.

Graph 1

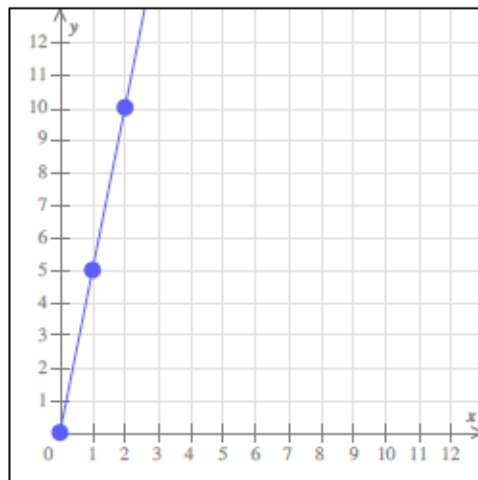


Proportional

$y$  is \_\_\_\_\_ times  $x$

Not proportional

Graph 2



Proportional

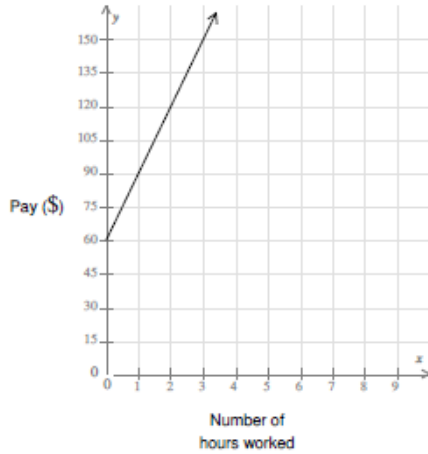
$y$  is \_\_\_\_\_ times  $x$

Not proportional

Question 3 of 14

Tammy makes house calls. For each, she is paid a base amount and makes additional money for each hour she works. The graph below shows her pay (in dollars) versus the number of hours worked.

Use the graph to answer the questions.



(a) How much does her pay increase for each hour worked?

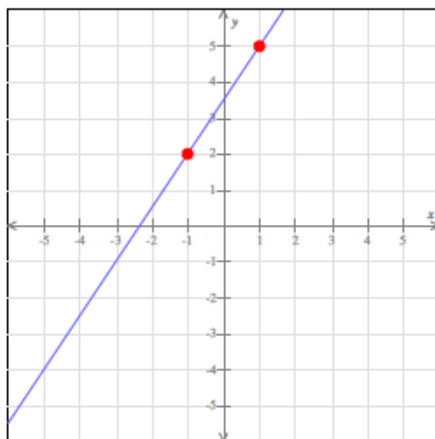
\$ \_\_\_\_\_

(b) What is the slope of the line?

\_\_\_\_\_

**Question 4 of 14**

Find the slope of the line graphed below.

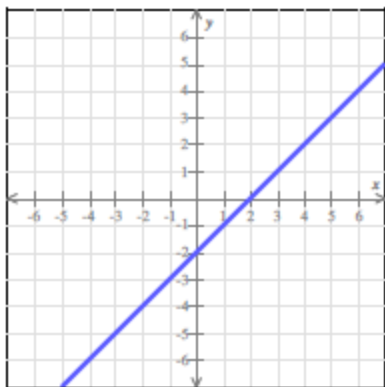


**Question 5 of 14**

Write an equation in slope-intercept form for the line with slope  $-\frac{1}{2}$  and y-intercept  $-7$ .

**Question 6 of 14**

Find the x-intercept and the y-intercept of the line below. Click on "None" if applicable.



**Question 7 of 14**

The equation of a line is given below.

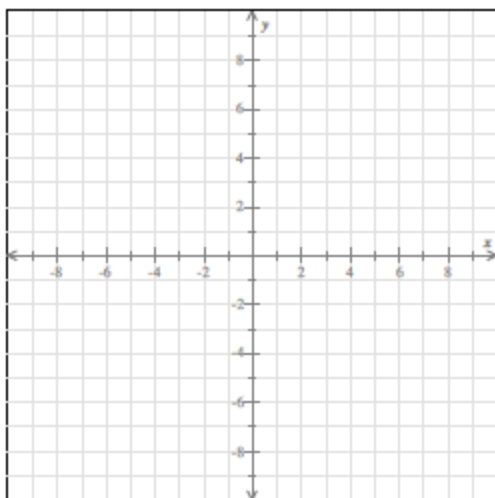
$$2x + 3y = -12$$

Find the x-intercept and the y-intercept.

Then use them to graph the line.

x-intercept: \_\_\_\_\_

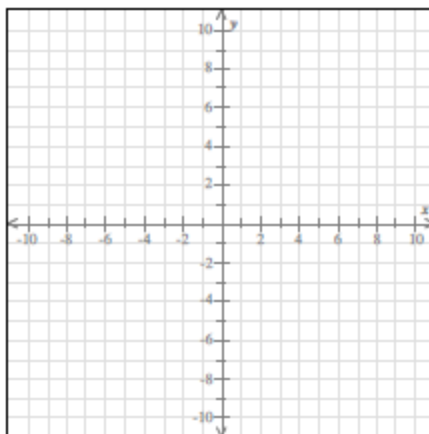
y-intercept: \_\_\_\_\_



**Question 8 of 14**

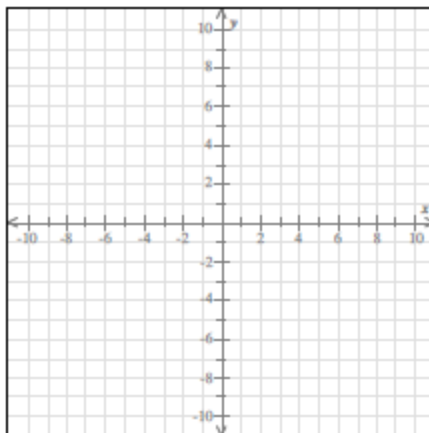


Graph the line with slope  $-\frac{3}{4}$  passing through the point  $(-3, -4)$ .



**Question 9 of 14**

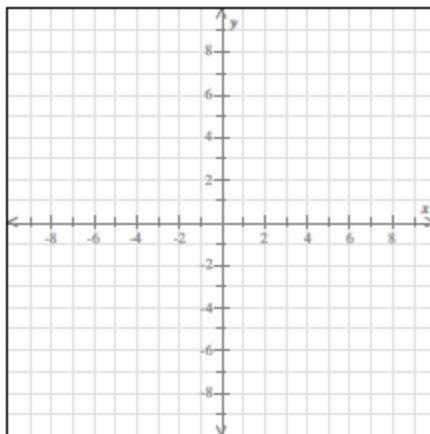
Graph the line with slope 3 passing through the point  $(-5, 3)$ .



**Question 10 of 14**

Graph the line.

$$y + 2 = \frac{4}{3}(x + 5)$$



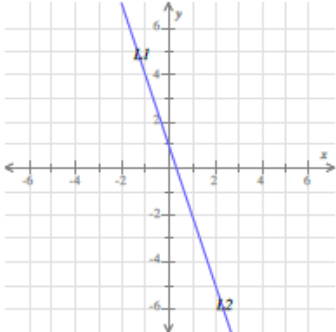
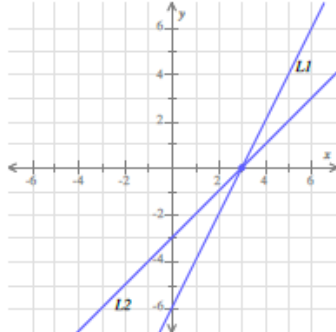
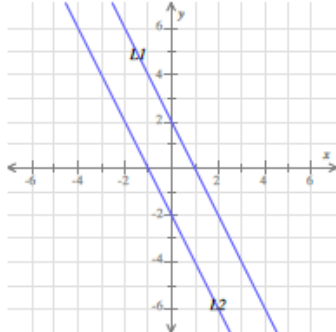
**Question 11 of 14**

A line passes through the point  $(2, -6)$  and has a slope of  $-\frac{5}{2}$ .

Write an equation in point-slope form for this line.

**Question 12 of 14**

Three systems of linear equations are shown.  
 For each system, choose the best description of its solution.  
 If the system has exactly one solution, give its solution.

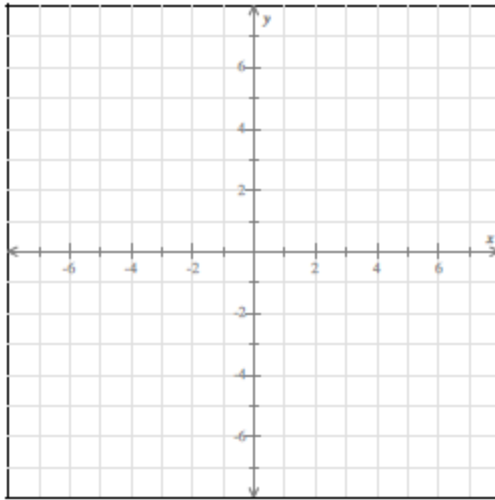
System A	System B	System C
Line 1: $y = -3x + 1$	Line 1: $y = 2x - 6$	Line 1: $y = -2x + 2$
Line 2: $3x + y = 1$	Line 2: $y = x - 3$	Line 2: $y = -2x - 2$
		
<input type="radio"/> The system has exactly one solution. Solution: ( __ , __ ) <input type="radio"/> The system has infinitely many solutions. <input type="radio"/> The system has no solution.	<input type="radio"/> The system has exactly one solution. Solution: ( __ , __ ) <input type="radio"/> The system has infinitely many solutions. <input type="radio"/> The system has no solution.	<input type="radio"/> The system has exactly one solution. Solution: ( __ , __ ) <input type="radio"/> The system has infinitely many solutions. <input type="radio"/> The system has no solution.

Question 13 of 14

Here is a system of equations.

$$\begin{cases} y = -2x - 6 \\ y = -x - 4 \end{cases}$$

Graph the system. Then write its solution. Note that you can also answer "No solution" or "Infinitely many" solutions.



**Question 14 of 14**

Solve the system of equations.

$$y = 7x$$

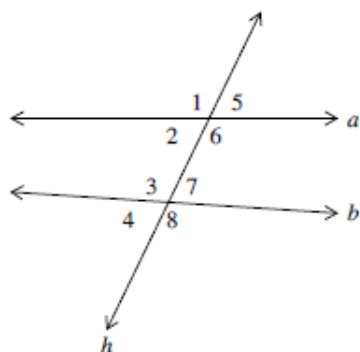
$$y = 3x + 20$$

(McGraw-Hill: ALEKS, n.d.)

APPENDIX C  
CHAPTER 5 TEST

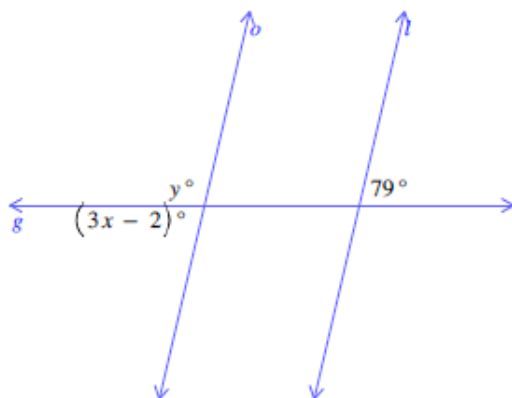
**Question 1 of 15**

Give a pair of alternate interior angles, a pair of alternate exterior angles, and a pair of corresponding angles.



**Question 2 of 15**

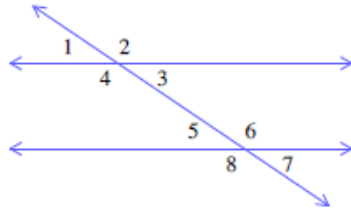
In the figure below,  $o \parallel l$ . Find the values of  $y$  and  $x$ .



**Question 3 of 15**

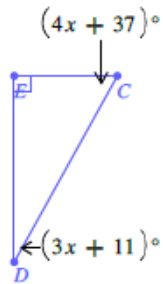
Two parallel lines are cut by a transversal as shown below.

Suppose  $m\angle 6 = 146^\circ$ . Find  $m\angle 1$  and  $m\angle 3$ .



**Question 4 of 15**

In the triangle below,  $\angle E$  is a right angle. Suppose that  $m\angle C = (4x + 37)^\circ$  and  $m\angle D = (3x + 11)^\circ$ .



(a) Write an equation to find  $x$ . Make sure you use an "=" sign in your answer.

Equation:

(b) Find the degree measure of each angle.

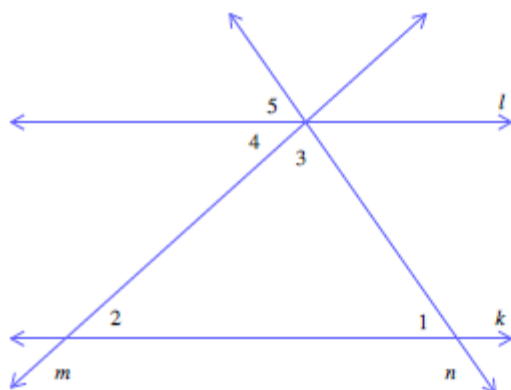
$$m\angle C = \underline{\hspace{1cm}}^\circ$$

$$m\angle D = \underline{\hspace{1cm}}^\circ$$

$$m\angle E = \underline{\hspace{1cm}}^\circ$$

**Question 5 of 15**

In the figure below, lines  $l$  and  $k$  are parallel.  
 Suppose that  $m\angle 1 = 55^\circ$  and  $m\angle 4 = 42^\circ$ .



Complete the statements below.

We see that  $\angle 2$  and  $\angle 4$  are \_\_\_\_\_.  
 And since lines  $l$  and  $k$  are parallel,  $\angle 2$  and  $\angle 4$  are \_\_\_\_\_.  
 So,  $m\angle 2 = \square^\circ$ .

We see that  $\angle 1$  and  $\angle 5$  are \_\_\_\_\_.  
 And since lines  $l$  and  $k$  are parallel,  $\angle 1$  and  $\angle 5$  are \_\_\_\_\_.  
 So,  $m\angle 5 = \square^\circ$ .

By the angle addition property,  $m\angle 5 + m\angle 4 + m\angle 3 = \square^\circ$ .  
 Note that  $m\angle 5 = 55^\circ$  and  $m\angle 4 = 42^\circ$ , so  $m\angle 3 = \square^\circ$ .

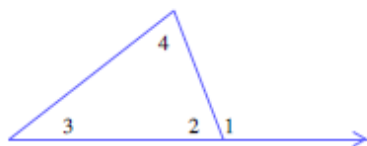
Therefore,  $m\angle 1 + m\angle 2 + m\angle 3 = \square^\circ$ .

The relationship between  $\angle 1$ ,  $\angle 2$ , and  $\angle 3$  is an example of the following rule.  
 The sum of the interior angle measures of a triangle is  $\square^\circ$ .

Question 6 of 15



In the figure below, suppose  $m\angle 3 = 38^\circ$  and  $m\angle 4 = 73^\circ$ .



Complete the statements below.

The sum of the interior angle measures of a triangle must be °.

So,  $m\angle 2 + m\angle 3 + m\angle 4 =$  °.

We are given that  $m\angle 3 = 38^\circ$  and  $m\angle 4 = 73^\circ$ .

Therefore,  $m\angle 3 + m\angle 4 =$  °.

And so  $m\angle 2 =$  °.

From the figure, we can see that  $m\angle 1 + m\angle 2 =$  °.

Using the value we already found for  $m\angle 2$ , we find that  $m\angle 1 =$  °.

Therefore,  $m\angle 1$        $m\angle 3 + m\angle 4$ .

This result is an example of the Exterior Angle Property of Triangles.  
For any triangle, the measure of an exterior angle \_\_\_\_\_.

**Question 7 of 15**

Find the sum of the interior angle measures of a convex 14-gon (a fourteen-sided polygon).

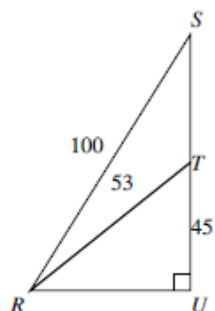
**Question 8 of 15**

The sum of the interior angle measures of a convex polygon is  $1440^\circ$ . How many sides does it have?

**Question 9 of 15**

Use the information given in the figure to find the length  $SU$ .  
If applicable, round your answer to the nearest whole number.

The lengths on the figure are not drawn accurately.



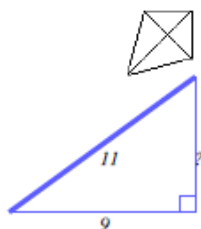
**Question 10 of 15**

Determine whether a triangle with the given side lengths is a right triangle.

Side lengths	Right triangle	Not a right triangle	Not enough information
7, 24, 25	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
15, 35, 40	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
10, 24, 26	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6, 7, 9	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

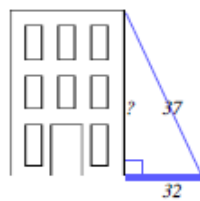
**Question 11 of 15**

A kite flying in the air has an 11-ft line attached to it. Its line is pulled taut and casts a 9-ft shadow. Find the height of the kite. If necessary, round your answer to the nearest tenth.



**Question 12 of 15**

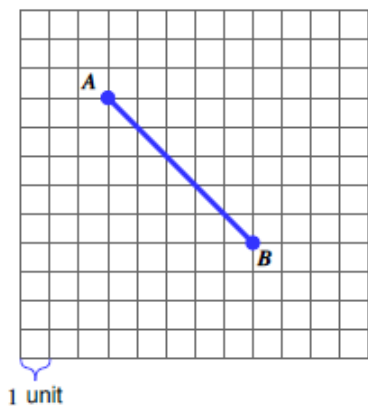
The length of a shadow of a building is 32 m. The distance from the top of the building to the tip of the shadow is 37 m. Find the height of the building. If necessary, round your answer to the nearest tenth.



**Question 13 of 15**

Find the distance between the points  $A$  and  $B$  given below.  
(That is, find the length of the segment connecting  $A$  and  $B$ .)

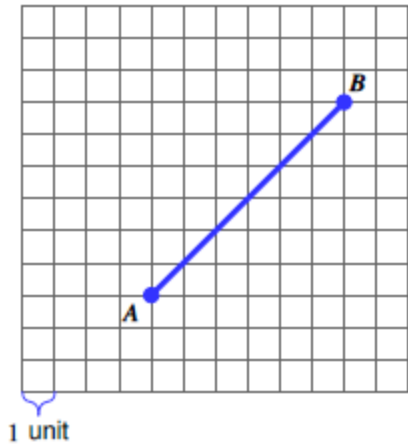
Round your answer to the nearest hundredth.



**Question 14 of 15**

Find the distance between the points  $A$  and  $B$  given below.  
(That is, find the length of the segment connecting  $A$  and  $B$ .)

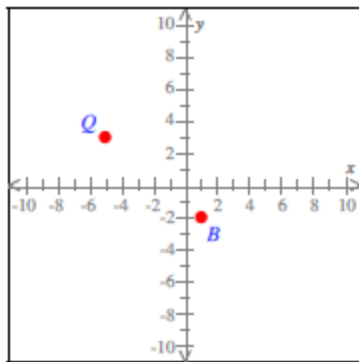
Round your answer to the nearest hundredth.



**Question 15 of 15**

Calculate the distance between the points  $Q = (-5, 3)$  and  $B = (1, -2)$  in the coordinate plane.

Round your answer to the nearest hundredth.



(McGraw-Hill: ALEKS, n.d.)

APPENDIX D

MEASURE ORDER OF 463 STUDENTS AND 29 ITEMS

*Measure Order of 463 Students and 29 Items*

Item	Total Score	Count	Logit Measure	Infit MNSQ	Outfit MNSQ
Item 509	70	461	3.16	1.11	1.04
Item 310	70	446	3.12	1.11	1.27
Item 313	96	446	2.43	1.24	2.49
Item 515	134	461	1.65	0.90	0.61
Item 311	143	446	1.43	1.02	0.78
Item 309	155	446	1.20	0.95	1.09
Item 308	158	446	1.15	1.18	1.08
Item 314	158	446	1.15	1.15	1.36
Item 513	163	461	1.10	0.75	0.55
Item 514	166	461	1.05	0.77	0.61
Item 512	194	461	0.56	0.83	0.66
Item 511	199	461	0.47	0.77	0.55
Item 508	206	461	0.35	1.11	1.02
Item 506	208	461	0.32	0.91	0.82
Item 307	215	446	0.14	1.13	1.05
Item 507	225	461	0.03	0.93	0.73
Item 504	226	461	0.02	0.87	0.75
Item 505	228	461	-0.02	0.95	0.82
Item 502	248	461	-0.35	1.01	1.30
Item 304	250	446	-0.46	0.99	0.94
Item 305	261	446	-0.65	0.90	0.76
Item 303	275	446	-0.90	1.34	1.94
Item 503	286	461	-1.00	0.91	1.09
Item 312	285	446	-1.07	1.02	0.87
Item 510	310	461	-1.42	1.14	1.08
Item 501	357	461	-2.34	1.05	0.80
Item 306	372	446	-2.90	0.90	0.82
Item 301	408	446	-4.08	1.32	2.00
Item 302	409	446	-4.13	1.02	1.00
<i>Mean</i>	223.3	453.0	0.00	1.01	1.03
<i>S.D.</i>	88.9	7.5	1.77	0.15	0.44

APPENDIX E

MEASURE ORDER OF 142 STUDENTS AND 29 ITEMS

Measure Order of 142 Students and 29 Items

Item	Total Score	Count	Logit Measure	Infit MNSQ	Outfit MNSQ
Item 310	13	138	3.20	0.89	0.60
Item 509	19	142	2.48	1.01	1.05
Item 313	19	138	2.47	1.43	0.86
Item 515	21	142	2.28	0.89	0.53
Item 309	30	138	1.52	0.93	0.67
Item 311	32	138	1.37	0.94	0.63
Item 308	34	138	1.23	1.21	1.07
Item 513	35	142	1.17	0.75	0.52
Item 514	38	142	0.98	0.86	0.77
Item 314	38	138	0.96	1.15	1.26
Item 506	43	142	0.67	0.92	0.67
Item 511	44	142	0.61	0.81	0.55
Item 512	45	142	0.55	0.80	0.58
Item 508	48	142	0.37	0.90	0.77
Item 505	49	142	0.31	0.91	0.60
Item 507	54	142	0.03	0.89	0.65
Item 504	56	142	-0.08	0.91	0.65
Item 307	57	138	-0.17	1.14	0.99
Item 305	59	138	-0.28	1.11	1.10
Item 502	62	142	-0.41	1.07	1.52
Item 503	67	142	-0.67	0.90	0.68
Item 303	68	138	-0.77	1.35	1.79
Item 304	69	138	-0.82	0.95	0.70
Item 312	72	138	-0.99	0.98	0.93
Item 510	88	142	-1.77	1.19	1.07
Item 501	101	142	-2.47	1.16	0.86
Item 306	114	138	-3.41	0.88	0.61
Item 301	122	138	-4.03	1.28	2.04
Item 302	125	138	-4.32	1.03	1.29
<i>Mean</i>	55.9	140.1	0	1.01	0.90
<i>S.D.</i>	29.6	2.0	2	0.17	0.38



APPENDIX F  
SUMMARY OF DIF ANALYSIS BY ELPAC LEVEL

*Summary of DIF Analysis by ELPAC Level*

Item	Observed Average	Expected Average	DIF Score	DIF Measure	DIF Size	<i>p</i>
Item 301						
1	0.81	0.83	-0.02	-3.85	0.18	0.804
2	0.92	0.81	0.11	-5.19	-1.16	0.152
3	0.82	0.89	-0.06	-3.37	0.67	0.104
4	0.97	0.94	0.04	-5.05	-1.02	0.340
Item 302						
1	0.81	0.86	-0.05	-3.85	0.47	0.526
2	0.80	0.85	-0.05	-3.91	0.41	0.484
3	0.93	0.91	0.02	-4.64	-0.32	0.570
4	0.97	0.95	0.02	-5.05	-0.73	0.493
Item 303						
1	0.50	0.40	0.10	-1.62	-0.85	0.235
2	0.36	0.37	-0.01	-0.66	0.11	0.846
3	0.44	0.48	-0.04	-0.46	0.30	0.403
4	0.63	0.60	0.03	-0.95	-0.18	0.671
Item 304						
1	0.44	0.40	0.04	-1.13	-0.31	0.675
2	0.44	0.38	0.06	-1.29	-0.47	0.406
3	0.49	0.49	0.00	-0.85	-0.03	0.936
4	0.55	0.61	-0.06	-0.43	0.39	0.348
Item 305						
1	0.31	0.34	-0.03	0.03	0.31	0.711
2	0.28	0.32	-0.04	0.06	0.34	0.594
3	0.46	0.41	0.04	-0.59	-0.32	0.382
4	0.50	0.53	-0.03	-0.11	0.17	0.674
Item 306						
1	0.69	0.76	-0.07	-2.93	0.48	0.479
2	0.80	0.74	0.06	-3.91	-0.50	0.395
3	0.84	0.83	0.01	-3.53	-0.13	0.761
4	0.87	0.90	-0.03	-2.99	0.42	0.464
Item 307						
1	0.31	0.33	-0.02	0.03	0.20	0.813
2	0.32	0.30	0.02	-0.31	-0.15	0.810
3	0.39	0.40	-0.01	-0.07	0.10	0.790
4	0.53	0.51	0.02	-0.27	-0.10	0.800

Summary of DIF Analysis by ELPAC Level Cont.

Item	Observed Average	Expected Average	DIF Score	DIF Measure	DIF Size	$\rho$
Item 308						
1	0.38	0.21	0.16	-0.59	-1.82	0.031
2	0.20	0.17	0.03	0.90	-0.33	0.633
3	0.26	0.23	0.03	0.93	-0.30	0.443
4	0.16	0.29	-0.14	2.41	1.18	0.033
Item 309						
1	0.31	0.19	0.12	0.03	-1.49	0.091
2	0.12	0.15	-0.03	1.96	0.44	0.580
3	0.18	0.20	-0.03	1.78	0.27	0.549
4	0.26	0.26	0.01	1.47	-0.05	0.911
Item 310						
1	0.06	0.09	-0.03	3.78	0.59	0.638
2	0.12	0.06	0.06	1.96	-1.24	0.130
3	0.05	0.08	-0.03	3.84	0.64	0.365
4	0.11	0.10	0.01	3.05	-0.15	0.813
Item 311						
1	0.19	0.20	-0.02	1.59	0.22	0.820
2	0.12	0.16	-0.04	1.96	0.59	0.463
3	0.23	0.22	0.01	1.25	-0.12	0.760
4	0.29	0.28	0.01	1.27	-0.10	0.818
Item 312						
1	0.38	0.42	-0.05	-0.59	0.40	0.609
2	0.48	0.40	0.08	-1.59	-0.60	0.278
3	0.49	0.51	-0.02	-0.85	0.13	0.711
4	0.63	0.64	-0.01	-0.95	0.04	0.934
Item 313						
1	0.31	0.13	0.18	0.03	-2.44	0.010
2	0.12	0.09	0.03	1.96	-0.51	0.522
3	0.09	0.12	-0.03	3.03	0.56	0.335
4	0.11	0.15	-0.05	3.05	0.58	0.351
Item 314						
1	0.25	0.23	0.02	0.75	-0.21	0.816
2	0.16	0.19	-0.03	1.39	0.43	0.559
3	0.28	0.26	0.02	0.78	-0.19	0.630
4	0.32	0.33	-0.02	1.08	0.12	0.784

*Summary of DIF Analysis by ELPAC Level Cont.*

Item	Observed Average	Expected Average	DIF Score	DIF Measure	DIF Size	$\rho$
Item 501						
1	0.47	0.61	-0.13	-1.53	0.94	0.179
2	0.62	0.60	0.02	-2.61	-0.14	0.791
3	0.66	0.71	-0.05	-2.10	0.37	0.294
4	0.95	0.82	0.13	-4.25	-1.78	0.029
Item 502						
1	0.24	0.34	-0.10	0.77	1.18	0.200
2	0.46	0.32	0.14	-1.53	-1.13	0.044
3	0.44	0.43	0.02	-0.52	-0.12	0.744
4	0.47	0.55	-0.07	0.06	0.46	0.258
Item 503						
1	0.29	0.37	-0.07	0.06	0.73	0.379
2	0.38	0.35	0.03	-0.95	-0.27	0.628
3	0.46	0.46	0.00	-0.65	0.03	0.941
4	0.61	0.59	0.02	-0.77	-0.10	0.809
Item 504						
1	0.18	0.31	-0.13	1.60	1.68	0.095
2	0.35	0.29	0.06	-0.63	-0.55	0.348
3	0.39	0.38	0.01	-0.14	-0.06	0.864
4	0.50	0.50	0.00	-0.11	-0.02	0.952
Item 505						
1	0.29	0.27	0.02	0.06	-0.25	0.760
2	0.23	0.25	-0.02	0.47	0.16	0.806
3	0.27	0.33	-0.06	0.81	0.50	0.200
4	0.53	0.43	0.09	-0.27	-0.58	0.161
Item 506						
1	0.18	0.24	-0.07	1.60	0.93	0.338
2	0.19	0.21	-0.02	0.91	0.24	0.723
3	0.22	0.29	-0.07	1.27	0.60	0.140
4	0.53	0.38	0.15	-0.27	-0.93	0.027
Item 507						
1	0.29	0.30	0.00	0.06	0.03	0.971
2	0.15	0.27	-0.12	1.40	1.37	0.070
3	0.39	0.37	0.02	-0.14	-0.17	0.633
4	0.53	0.48	0.05	-0.27	-0.30	0.466

*Summary of DIF Analysis by ELPAC Level Cont.*

Item	Observed Average	Expected Average	DIF Score	DIF Measure	DIF Size	$p$
Item 508						
1	0.24	0.27	-0.03	0.77	0.40	0.654
2	0.19	0.24	-0.05	0.91	0.54	0.434
3	0.32	0.32	0.00	0.37	0.00	1.000
4	0.47	0.42	0.05	0.06	-0.31	0.446
Item 509						
1	0.12	0.12	-0.01	2.57	0.09	0.931
2	0.04	0.09	-0.05	3.70	1.22	0.305
3	0.15	0.12	0.03	2.00	-0.47	0.305
4	0.13	0.15	-0.02	2.70	0.23	0.689
Item 510						
1	0.71	0.50	0.20	-3.14	-1.36	0.047
2	0.35	0.49	-0.15	-0.63	1.15	0.056
3	0.64	0.62	0.03	-1.98	-0.21	0.556
4	0.71	0.74	-0.03	-1.52	0.26	0.568
Item 511						
1	0.24	0.25	-0.01	0.77	0.17	0.853
2	0.19	0.22	-0.03	0.91	0.30	0.658
3	0.32	0.30	0.03	0.39	-0.22	0.555
4	0.37	0.39	-0.02	0.73	0.12	0.775
Item 512						
1	0.18	0.25	-0.08	1.60	1.05	0.282
2	0.23	0.22	0.01	0.47	-0.07	0.910
3	0.37	0.30	0.07	-0.01	-0.56	0.127
4	0.32	0.40	-0.08	1.08	0.54	0.220
Item 513						
1	0.24	0.21	0.03	0.77	-0.40	0.651
2	0.08	0.17	-0.09	2.67	1.49	0.112
3	0.25	0.23	0.02	0.96	-0.22	0.577
4	0.32	0.30	0.01	1.08	-0.09	0.828
Item 514						
1	0.29	0.22	0.07	0.06	-0.92	0.274
2	0.12	0.19	-0.07	1.96	0.99	0.222
3	0.27	0.25	0.02	0.81	-0.17	0.660
4	0.32	0.33	-0.01	1.08	0.10	0.811

*Summary of DIF Analysis by ELPAC Level Cont.*

Item	Observed Average	Expected Average	DIF Score	DIF Measure	DIF Size	$p$
Item 515						
1	0.12	0.13	-0.02	2.57	0.29	0.782
2	0.19	0.10	0.10	0.91	-1.37	0.055
3	0.14	0.13	0.00	2.22	-0.06	0.907
4	0.11	0.17	-0.06	3.05	0.77	0.216

*Note: Students with ELPAC score of 1 = 1, ELPAC2 = 2, ELPAC3 = 3, ELPAC4 = 4.*

APPENDIX G

COMMAND CODE FOR ANALYSIS OF 463 STUDENTS AND 29 ITEMS

```

TITLE= 'DIF ANALYSIS BETWEEN EL AND NON-EL STUDENTS'

NI= 29 ; 29 items
ITEM1= 1 ; responses start in column 1 of the data
NAME1= 30 ; person-label starts in column 30 of the data
NAMELENGTH = 5 ; Length of person label
ITEM= ITEM ; items are called "items"
PERSON= STUDENTS ; persons are called "students"
CODES= 01 ; valid response codes (ratings) are 0, 1
MISSSCORE = -1 ; all coes not listed in CODES are to be treated
as "not administered"

CLFILE= * ; label the response categories
0 Wrong ; names of the response categories
1 Right
* ; "*" means the end of a list
@ELSTATUS = $s5W1 ; EL Status indicator in column 5 of student data record
DIF = @ELSTATUS
PSUBTOTAL = @ELSTATUS ; Subtotal by EL Status

&END ; this ends the control specifications

Item 301 ; These are brief descriptions of the 25 items
Item 302
Item 303
Item 304
Item 305
Item 306
Item 307
Item 308
Item 309
Item 310
Item 311
Item 312
Item 313
Item 314
Item 501
Item 502
Item 503
Item 504
Item 505
Item 506
Item 507
Item 508
Item 509
Item 510
Item 511
Item 512
Item 513
Item 514
Item 515
END NAMES ;this follows the item names

```



APPENDIX H

COMMAND CODE FOR ANALYSIS OF 142 STUDENTS AND 29 ITEMS

```

TITLE= 'DIF ANALYSIS AMONG STUDENTS WITH DIFFERENT ELPAC LEVEL'

NI= 29 ; 29 items
ITEM1= 1 ; responses start in column 1 of the data
NAME1= 30 ; person-label starts in column 30 of the data
NAMELENGTH = 5 ; Length of person label
ITEM= ITEM ; items are called "items"
PERSON= STUDENTS ; persons are called "students"
CODES= 01 ; valid response codes (ratings) are 0, 1
MISSSCORE = -1 ; all coes not listed in CODES are to
; be treated as "not administered"

CLFILE= * ; label the response categories
0 Wrong ; names of the response categories
1 Right
* ; "*" means the end of a list
@ELPACLEVEL = $s5W1 ; ELPACLEVEL indicator in column 5 of student data
; record

DIF = @ELPACLEVEL
PSUBTOTAL = @ELPACLEVEL ; Subtotal by ELPACLEVEL

&END ; this ends the control specifications

Item 301 ; These are brief descriptions of the 25
; items
Item 302
Item 303
Item 304
Item 305
Item 306
Item 307
Item 308
Item 309
Item 310
Item 311
Item 312
Item 313
Item 314
Item 501
Item 502
Item 503
Item 504
Item 505
Item 506
Item 507
Item 508
Item 509
Item 510
Item 511
Item 512
Item 513
Item 514
Item 515
END NAMES ;this follows the item names

```

## REFERENCES

- 2013 mathematics framework chapters—Curriculum frameworks (CA dept of education). (n.d.). Retrieved November 2, 2017, from <https://www.cde.ca.gov/ci/ma/cf/mathfwchapters.asp>
- Abedalaziz, N., Leng, C. H., & Alahmadi, A. (2014). Detecting a gender-related differential item functioning using transformed item difficulty. *Malaysian Online Journal of Educational Sciences*, 2(1), 16–22.
- Abedi, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data* (No. 603; p. 167). Los Angeles, CA: Center for the Study of Evaluation.
- Abedi, J. (2004). The No Child Left Behind Act and English Language Learners: Assessment and accountability issues. *Educational Researcher*, 33(1), 4–14.
- Abedi, J. (2014). The use of computer technology in designing appropriate test accommodations for English Language Learners. *Applied Measurement in Education*, 27(4), 261–272.
- Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, 74(1), 1–28.
- Abedi, J., & Levine, H. G. (2013). Fairness in assessment of English learners. *Leadership*, 42(3), 26–28.

- Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234.
- Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16–26.
- Acosta, B. D., Rivera, C., & Willner, L. S. (2008). *Best practices in state assessment policies for accommodating English language learners: A delphi study*. Arlington, VA: George Washington University Center for Equity and Excellence in Education.
- Akkus, M. (2016). The common core state standards for mathematics. *International Journal of Research in Education and Science, 2*(1), 49–54.
- Albus, D., Bielinski, J., Thurlow, M., & Liu, K. (2001). *The effect of a simplified English language dictionary on a reading test*. (LEP Projects No. 1).
- Alcocer, P. (n.d.). History of standardized testing in the United States. Retrieved November 18, 2019, from National Education Association website: <http://www.nea.org/home/66139.htm>
- Ault, L. H. (1972). *Multiple-choice versus created-response test items*.
- Baghaei, P., & Aryadoust, V. (2015). Modeling local item dependence due to common test format with a multidimensional rasch model. *International Journal of Testing, 15*(1), 71–87.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: WHFreeman.

- Betz, N. E., & Hackett, G. (1983). The relationship of mathematics self-efficacy expectations to the selection of science-based college majors. *Journal of Vocational Behavior, 23*(3), 329–345.
- Bond, T., & Fox, C. M. (2015). *Applying the rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.
- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2002). Different kinds of DIF: A distinction between absolute and relative forms of measurement invariance and bias. *Applied Psychological Measurement, 26*(4), 433–450.
- Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement, 30*(4), 277–291.
- Butler, R. (2000). Making judgments about ability: The role of implicit theories of ability in moderating inferences from temporal and social comparison information. *Journal of Personality and Social Psychology, 78*(5), 965–978.
- California Department of Education. (2017). Retrieved November 7, 2017, from <https://www.cde.ca.gov/>
- California Department of Education. (n.d.). Understanding California assessment of student progress and performance (CAASPP) summary reports. Retrieved June 10, 2018, from California Assessment of Student Performance and Progress website: <https://caaspp.cde.ca.gov/sb2016/UnderstandingCAASPPReports>

- Chinn, S. (2009). Mathematics anxiety in secondary students in England. *Dyslexia*, 15(1), 61–68.
- Common Core State Standards Initiative. (2010). *Common Core State Standards for Mathematics*. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers.
- Common Core state standards for mathematics. (2010). Washington D.C.
- Common Core State Standards Initiative. (n.d.). Retrieved August 11, 2018, from Mathematics standards website: <http://www.corestandards.org/Math/>
- Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.
- Cronbach, L. J. (1941). An experimental comparison of the multiple true-false and multiple multiple-choice tests. *Journal of Educational Psychology*, 32(7), 533–543.
- Doolittle, A. E., & Cleary, T. A. (1987). Gender-based Differential Item Performance in mathematics achievement items. *Journal of Educational Measurement*, 24(2), 157–166.
- Duncan, T. G., Parent, L. del R., Chen, W.-H., Ferrara, S., Johnson, E., Oppler, S., & Shieh, Y.-Y. (2005). Study of a dual-language test booklet in eighth-grade mathematics. *Applied Measurement in Education*, 18(2), 129–161.
- Elawar, M. C., & Corno, L. (1985). A factorial experiment in teachers' written feedback on student homework: Changing teacher behavior a little rather than a lot. *Journal of Educational Psychology*, 77(2), 162–173.

- Elliott A.C., & Woodward W.A. (2007). *Statistical analysis quick reference guidebook with SPSS examples*. London: Sage Publications.
- Embedded universal tools, designated supports, and accommodations video tutorials. (n.d.). Retrieved November 20, 2019, from California Assessment of Student Performance and Progress website:  
<http://www.caaspp.org/training/caaspp/uaag.html>
- Parent/Guardian resources*. (n.d.). English Language Proficiency Assessments for California. Retrieved March 10, 2020, from  
<https://www.elpac.org/resources/parent-resources/>
- Grosse, M. E., & Wright, B. D. (1985). Validity and reliability of true–false tests. *Educational and Psychological Measurement, 45*(1), 1–13.
- Hackett, G., & Betz, N. E. (1989). An exploration of the mathematics self-efficacy/mathematics performance correspondence. *Journal for Research in Mathematics Education, 20*(3), 261–273.
- Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education, 21*(1), 33–46.
- Hoaglin, D.C., Iglewicz, B., & Tukey, J.W. (1986). Performance of some resistant rules for outlier labeling. *Journal of American Association, 81*, 991-999.
- Hofstetter, C. H. (2003). Contextual and mathematics accommodation test effects for English-language learners. *Applied Measurement in Education, 16*(2), 159–188.

- Holland, P. W., & Thayer, D. T. (1986). *Differential item performance and the Mantel-Haenszel procedure*.
- Inside Mathematics. (2018). Retrieved February 24, 2018, from <http://www.insidemathematics.org/common-core-resources/mathematical-practice-standards>
- Johnson, E., & Monroe, B. (2004). Simplified language as an accommodation on math tests. *Assessment for Effective Intervention, 29*(3), 35–45.
- Kastner, M., & Stangla, B. (2011). Multiple choice and constructed response tests: Do test format and scoring matter? *Procedia - Social and Behavioral Sciences, 12*, 263–273.
- Khaliqi, D. (2016). How common is the common core? A global analysis of math teaching and learning. *School Science and Mathematics, 116*(4), 199–211.
- Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for english language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research, 79*(3), 1168–1201.
- Kopriva, R. J., Emick, J. E., Hipolito-Delgado, C. P., & Cameron, C. A. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educational Measurement: Issues and Practice, 26*(3), 11–20.



- Linacre, J. M. (2006). WINSTEPS Rasch measurement computer program. *Chicago: WINSTEPS. com.*
- Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational assessment, 14*(3-4), 160-179.
- McGraw-Hill: ALEKS. (n.d.). Retrieved May 13, 2020, from <https://www.aleks.com/>
- Messick, S. (1996). Validity of performance assessment. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1-18). Washington, DC: National Center for Education Statistics (NCES 96-802).
- Miller, E. R., Okum, I., Sinai, R., & Miller, K. S. (1999, April). A study of the English language readiness of limited English proficient students to participate in New Jersey's statewide assessment system. *In annual meeting of the National Council of Measurement in Education, Montreal, Canada.*
- Moon, J. A., Keehner, M., & Katz, I. R. (2018). Affordances of item formats and their effects on test-taker cognition under uncertainty. *Educational Measurement: Issues and Practice.*
- Moses, T., Liu, J., Tan, A., Deng, W., & Dorans, N. J. (2013). Constructed-response DIF evaluations for mixed-format tests. *ETS Research Report Series.*

National Governors Association Center for Best Practices, Council of Chief State School Officers (NGA/CCSSO). (2010). *Common Core State Standards for Mathematics, Appendix A*. Washington, DC: National Governor's Association Center for Best Practices, Council of Chief State School Officers. Retrieved March 5, 2018 from [http://www.corestandards.org/assets/CCSSI\\_Mathematics\\_Appendix\\_A.pdf](http://www.corestandards.org/assets/CCSSI_Mathematics_Appendix_A.pdf)

Newstead, K. (1998). Aspects of children's mathematics anxiety. *Educational Studies in Mathematics*, 36(1), 53–71.

Ng, L. K. (2012). Mathematics anxiety in secondary school students. *Mathematics Education Research Group of Australasia*, 35.

Noonan, J. (1990). Readability problems presented by mathematics text. *Early Child Development and Care*, 54, 57–81.

Oztuna, D., Elhan, A.H., & Tuccar, E. (2006). Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. *Turkish Journal of Medical Sciences*, 36(3), 171-176.

Pajares, F., & Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. *Contemporary Educational Psychology*, 24(2), 124–139.

Pajares, F., & Kranzler, J. (1995). Self-efficacy beliefs and general mental ability in mathematical problem-solving. *Contemporary Educational Psychology*, 20(4), 426–443.

- Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology, 86*(2), 193–203.
- Pallant, J. (2007). *SPSS survival manual, a step by step guide to data analysis using SPSS for windows* (3<sup>rd</sup> ed., pp. 179-200). Sydney: McGraw Hill.
- Peat, J., & Barton, B. (2005). *Medical statistics: A guide to data analysis and critical appraisal*. Blackwell Publishing.
- Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice, 30*(3), 10–28.
- Rattan, A., Good, C., & Dweck, C. S. (2012). “It’s ok — Not everyone can be good at math”: Instructors with an entity theory comfort (and demotivate) students. *Journal of Experimental Social Psychology, 48*(3), 731–737.
- Rhodes, K. T., Branum-Martin, L., Morris, R. D., Ronski, M., & Sevcik, R. A. (2015). Testing math or testing language? The construct validity of the KeyMath-Revised for children with intellectual disability and language difficulties. *American Journal on Intellectual and Developmental Disabilities, 120*(6), 542–568.
- Rhodes, K. T., Branum-Martin, L., Washington, J. A., & Fuchs, L. S. (2017). Measuring arithmetic: A psychometric approach to understanding formatting effects and domain specificity. *Journal of Educational Psychology, 109*(7), 956–976.

- Richardson, F. C., & Suinn, R. M. (1972). The mathematics anxiety rating scale: Psychometric data. *Journal of Counseling Psychology, 19*(6), 551–554.
- Rivera, C., & Stansfield, C. W. (2001). *The effects of linguistic simplification of science test items on performance of limited English proficient and monolingual English-speaking students.*
- Scheuneman, J. D., & Grima, A. (1997). Characteristics of quantitative word items associated with differential performance for female and black examinees. *Applied Measurement in Education, 10*(4), 299–319.
- Schoenfeld, A. H. (1989). Explorations of students' mathematical beliefs and behavior. *Journal for Research in Mathematics Education, 20*(4), 338–355.
- Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of english language learners and students with disabilities. *Educational Assessment, 11*(2), 105–126.
- Smarter Balanced Question Types. (2018, March 8). Retrieved March 15, 2019, from California Department of Education website:  
<https://www.cde.ca.gov/ta/tg/sa/question-types.asp>
- Springer, R., Pugalee, D., & Algozzine, B. (2007). Improving mathematics skills of high school students. *Clearing House: A Journal of Educational Strategies, Issues and Ideas, 81*(1), 37–44.

- Traub, R., & Rowley, G. (1991). Understanding reliability. *Educational Measurement: Issues and Practice*, 10(1), 37–45.
- United Nations Educational, Scientific and Cultural Organization (UNESCO). (2017). Retrieved March 9, 2018, from <https://en.unesco.org/news/new-unesco-report-sheds-light-gender-inequality-stem-education>
- Usher, E. L., & Pajares, F. (2009). Sources of self-efficacy in mathematics: A validation study. *Contemporary Educational Psychology*, 34(1), 89–101.
- Wang, N., & Lane, S. (1994). *Detection of gender-based differential item functioning in a mathematics performance assessment*. New York, New York: Ford Foundation.
- Yee, F. P. (1987). Anxiety and mathematics performance in female secondary school students in Singapore. *Singapore Journal of Education*, 8(2), 22–31.
- Young, J. W., Cho, Y., Ling, G., Cline, F., Steinberg, J., & Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. *Educational Assessment*, 13(2–3), 170–192.
- Zhao, Y. (2005). Increasing math and science achievement: The best and worst of the east and west. *Phi Delta Kappan*, 87(3), 219–222.