Readability, Complexity, and Representations: A Holistic Analysis of New York State Algebra Exams

by

Stephanie Leah Schaefer

February 24, 2020

A dissertation submitted to the
faculty of the Graduate School of
the University at Buffalo, The State University of New York
in partial fulfillment of the requirements for the
degree of

Doctor of Philosophy

Department of Learning and Instruction

ProQuest Number: 27830059

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent on the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.



ProQuest 27830059

Published by ProQuest LLC (2020). Copyright of the Dissertation is held by the Author.

All Rights Reserved.
This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Copyright by

Stephanie Leah Schaefer

2020

All Rights Reserved

Dedication

“For human beings this is impossible, but for God all things are possible.” (Matthew 19:26)

To my aunts Pat and Kathy

Acknowledgements

I would like to first of all thank my parents, brother Andy, and family who have supported me in my academic studies and have tirelessly encouraged me to continue teaching.

I would also like to thank my doctoral committee: Dr. Deborah Moore-Russo, Dr. Sunha Kim, and Dr. Claire Cameron who have been very supportive of my research. Their guidance has been immeasurable. I am very grateful that Dr. Deborah Moore-Russo encouraged me to go into this PhD program after taking a summer class with her, and has constantly supported me since Summer 2016. To Dr. Kim and Dr. Cameron, I knew that you would both pore over my drafts and go above and beyond to give feedback that would improve my conceptualization of my topic.

Also, I would like to thank my work places that have encouraged and supported me in my doctoral studies. These places have truly transformed my teaching practices: Seton Catholic Central, SUNY Broome, St. Paul's Church, iTutor, and Southern New Hampshire University. It is truly because of your flexibility that these ideas could come to fruition. Thank you for giving me the opportunity to educate students!

Abstract

The State of New York has experienced three major changes in state algebra standards and assessments since 2002. Readability, context, and cognitive complexity have been studied separately as important variables that affect algebra success. Mathematical representation and the subsequent relationships among representations have been an integral part of international, national, and statewide algebra standards movements. The current study aimed to establish a holistic approach to consider trends in the New York state algebra exam items using the coding categories of readability, context, complexity, representations, and representation relationships. Results indicated higher reading levels, more context items, more variability in complexity, and increased levels of representations and representational relationships for exams and item types (constructed response and multiple choice).

Table of Contents

Dedication iii

Acknowledgements iv

Abstract v

List of Tables xii

List of Figures xiv

List of Appendices xvii

Chapter 1 – Introduction 1

Readability 3

Real World Context 4

Cognitive Complexity 5

Representations and Multiple Representations 6

Problem Statement 9

Significance of the Study 9

Research Questions 11

Definition of Terms 12

Overview of Dissertation 18

Chapter 2 – Review of Literature 19

History of Standards and Accountability Movements 19

New York State. 24

Summary. 25

Research on Readability in Mathematics 26

History and research of readability formulas. 26

Linguistic factors for readability research. 28

Combining readability tests. 29

Benefits of measuring readability. 30
Contextual research. 31
Factors affecting context. 32
Research on Complexity of Mathematics 33
Research about Mathematical Representations 36
Types of representations. 37
Cognitive load benefits in representations. 41
Visual representations research. 42
Symbols representations research. 44
Numeric representations research. 46
Written representations research. 48
Relationships among representations. 50
Visual and written representations. 50
Connections of multiple representations. 51
Benefits of multiple representations. 52
Factors affecting multiple representations. 53
Limitations of Research: Standards, Mathematical Representations, Context, Readability, and Complexity 55
Theoretical Framework 56
The process of mathematical thinking. 63
Relating semiotics to the assessment process. 64
Summary 65
Chapter 3 – Methods 66
Study Design: Data Set of New York State Regents Standards and Examinations 66
Previous Critical Reading and Standards Research 69
Interrater Reliability 71

Research Instruments 73
Preliminary Research 75
Schema 1: Readability 75
Schema 2: Context 76
Item Reliability and Context 76
Schema 3: Complexity 77
Item Readability, Context, and Complexity 77
Schema 4: Representations 78
Schema 5: Representational Relationships 80
Research Methodology and Rationale 81
Data Analysis 81
Summary 83
Chapter 4 – Results 84
Preliminary Analysis of Items and Item Parts 85
Preliminary Standards Research 86
Old exams. 86
Recent exams. 87
Current exams. 88
Research Question 1: Readability for Exam Versions and Standards 90
By exam version. 91
By item type. 98
Additional testing. 99
Research Question 2: "Real World" Context for Exam Versions and Standards 103
By exam version. 104
By item type. 105

Research Question 3: Relationship between Item Readability and Context.....108
Research Question 4: Cognitive Complexity for Exam Versions and Standards.....109
By exam version. 109
By item type. 111
Research Question 5: Relationships among Readability, Context, and Cognitive Complexity.....114
Research Question 6: Type and Number of Representations for Exam Versions 115
Number of representations by exam version. 115
Types of representations by exam version. 121
Visual-graph (V-G) representations by exam and item type. 122
Symbolic-algebraic representations (S-A) by exam and item type. 126
Numeric-Numeral representations (N-N) by exam and item type. 131
Written representations by exam and item type. 134
Research Question 7: Stem-Response Representation Relationships for Exam Versions and Standards.....138
Old exams. 140
Recent exams. 142
Current exams. 144
Additional tests for Visual-Graph stems. 146
Additional tests for Symbolic-Algebraic stems. 148
Additional tests for Numeric-Numeral stems. 149
Additional tests for Written stems. 151
Research Question 8: Representation Relationships.....152
By exam version and item type. 152
Summary.....158
Chapter 5 – Summary, Discussions, Implications 160
Discussion.....160

Research Question 1: What is the average grade level and average level of reading ease for each exam (*old, recent,* or *current)* and item part type (constructed response, multiple choice)? 161

Research Question 2: Do certain exams (*old*, *recent*, or *current*) rely more on a "real word" context? Which item part types (constructed response, multiple choice) rely the most on using "real world" contexts? 162

Research Question 3: Is there a correlation between item readability and presence of context? 163

Research Question 4: What is the average complexity of the items labeled by exam (*old, recent*, and *current*)? What is the average complexity of items labeled by item part types (constructed response, multiple choice)? 163

Research Question 5: Are there correlations among readability, context, and complexity? .164

Research Question 6: How many representations appear on average for each standard and overall for each exam (*old*, *recent*, or *current*)? What types of mathematical representations appear most frequently by item part types (constructed response, multiple choice)? 166

Research Question 7: How do representations interact with one another in the *stem* and *response* of the item? Do certain types of relationships occur with representations occur on certain exams (*old*, *recent*, or *current*) or by item part types (constructed response, multiple choice)? 168

Research Question 8: What types of representation relationships occur most frequently according to exam (*old, recent,* or *current*) and item part types (constructed response, multiple choice)? 168

Summary by Exam Type 169

Old exams. 169

Recent exams. 171

Current exams. 173

Limitations of Findings 176

Implications 178

Standards based education, cognitive load, and semiotics. 178

Algebra is more than a collection of standards. 181

Cognition and brain-based research. 182

Are the *current* standards more complex? 184
Recommendations 185
Teaching practices. 186
Improving reading fluency. 186
"Real world" context. 186
Cognitive demand. 187
Connections among representations. 187
Conclusion 188
References 190
Appendix 216

List of Tables

Table 1 *Item Counts from Old, Recent, and Current State Algebra Examinations by Item Type and Value* 15

Table 2 *List of Topic Areas for* Old, Recent, *and* Current *State Algebra Exams.* 15

Table 3 *List of* Current *Subtopics, by Topic Area, Noted on Exam Items Found on the* Current *State Algebra Exams* 16

Table 4 *Number of Item Counts for* Old, Recent, *and* Current *Regents Examinations* 69

Table 5 *Topics and Subtopic Areas of* Old, Recent, *and* Current *Items* 70

Table 6 *Coder Agreement for Stems and Responses* 72

Table 7 *Overall Coding Schema* 74

Table 8 *Representation Types for Stems and Responses of State Algebra Exam Items* 78

Table 9 *Counts and Relative Counts of Items on* Old *State Algebra Exams by* Old *Topic Areas* 87

Table 10 *Counts and Relative Counts of Items on* Recent *State Algebra Exams by* Recent *Topic Areas* 88

Table 11 *Counts and Relative Counts of Items on* Current *State Algebra Exams by* Current *Topic and Subtopic Areas* 89

Table 12 *Readability Test Medians for* Old, Recent, *and* Current *Exams* 91

Table 13 *Readability Test Medians by Item Type for* Old, Recent, *and* Current *Exams* 99

Table 14 *Relative Context Percentage by Item Type (Constructed Response, Multiple Choice) for* Old, Recent, *and* Current *Exams* 106

Table 15 *Cognitive Complexity by Item Part Type (Constructed Response, Multiple Choice) for* Old, Recent, *and* Current *Exams* 113

Table 16 Current *August 2018, item part 26a responses* 117

Table 17 Current *August 2018, item part 26b responses* 119

Table 18 *Number of Representational Subcategories Present in Item Parts Type for* Old, Recent, *and* Current Exams 120

Table 19 *Average Number of Representational Subcategories Present in Item Stems and Responses by Item Parts Type for* Old, Recent, *and* Current *Exams* 121

Table 20 *Prominent Representations in Stems and Responses on* Old, Recent, *and* Current *Exams* 122

Table 21 *Relative Visual-Graph Percentage for Stems and Responses by Item Type (Constructed Response, Multiple Choice) for* Old, Recent, *and* Current *Exams* 126

Table 22 *Relative Symbolic-Algebraic Percentage for Stems and Responses by Item Type (Constructed Response, Multiple Choice) for* Old, Recent, *and* Current *Exams* 130

Table 23 *Relative Numeric-Numeral Percentage for Stems and Responses by Item Type (Constructed Response, Multiple Choice) for* Old, Recent, *and* Current *Exams* 134

Table 24 *Relative Written Percentage for Stems and Responses by Item Type (Constructed Response, Multiple Choice) for* Old, Recent, *and* Current *Exams* 138

Table 25 *Stem-to-Response Relationship Medians for* Old *Exams.* 142

Table 26 *Stem-to-Response Relationship Medians for* Recent *Exams.* 144

Table 27 *Stem-to-Response Relationship Medians for* Current *Exams.* 146

Table 28 *Relative Representation Treatments Percentage Medians by Item Type for* Old, Recent, and Current *Exams* 154

Table 29 *Relative Representation Conversions Percentage Medians by Item Type for* Old, Recent, and Current *Exams* 156

Table 30 *Representation Synergy by Item Type for* Old, Recent, *and* Current *Exams* 158

Table 31 *Major Findings for* Old, Recent, *and* Current *Exams Organized by Schema* 175

List of Figures

Figure 1. *Current* Exam, August 2015, #25 45

Figure 2. The sign meanings of the function f(x) = 2x + 9. 58

Figure 3. Example of the semiotic bundle on the *current* June 2015 exam, item 1. 60

Figure 4. Example of the semiotic chain. 61

Figure 5. Example of the semiotic chain on the *Current* June 2015 exam, item 24. 62

Figure 6. Box and whisker plots of number of item parts for *old*, *recent*, and *current* exams. . 86

Figure 7. Box and whisker plots of Flesch Reading Ease for *old*, *recent*, and *current* exams... 92

Figure 8. Box and whisker plots of Flesch-Kincaid Grade Level for *old*, *recent*, and *current* exams. 93

Figure 9. Box and whisker plots of ATOS Level for *old*, *recent*, and *current* exams. 94

Figure 10. Box and whisker plots of Average Word Length for *old*, *recent*, and *current* exams. 95

Figure 11. Box and whisker plots of Average Sentence Length in words for *old*, *recent*, and *current* exams..... 96

Figure 12. Box and whisker plots of Average Vocabulary Level for *old*, *recent*, and *current* exams. 97

Figure 13. Box and whisker plots of Word Count for *old*, *recent*, and *current* exams. 98

Figure 14. Box and whisker plots of relative context percentage for *old*, *recent*, and *current* exams. 105

Figure 15. Stacked bar graph of the Relative Context Percentage by item type (constructed response) for *old*, *recent*, and *current* exams..... 107

Figure 16. Stacked bar graph of the Relative Context Percentage by item type (multiple choice) for *old*, *recent*, and *current* exams. 107

Figure 17. Scatterplot of ATOS Readability Scores versus Relative Context Percentage for *old*, *recent*, and *current* exams..... 108

Figure 18. Box and whisker plots of percentage of items with high Cognitive Complexity for *old*, *recent*, and *current* exams..... 110

Figure 19. Box and whisker plots of percentage of items with low complexity for *old*, *recent*, and *current* exams. 111

Figure 20. Stacked bar graph of cognitive complexity (for constructed response items) for *old*, *recent*, and *current* exams. 113

Figure 21. Stacked bar graph of cognitive complexity (for multiple choice items) for *old*, *recent*, and *current* exams. 114

Figure 22. Example of the labeling of representations on the *current* August 2018 exam, item part 26a. 116

Figure 23. Example of the labeling of representations on the *current* August 2018 exam, item part 26b. 118

Figure 24. Box and whisker plots of relative visual-graph percentage for stems on *old*, *recent*, and *current* exams. 123

Figure 25. Box and whisker plots of relative visual-graph percentage for responses on *old*, *recent*, and *current* exams. 124

Figure 26. Box and whisker plots of relative symbolic-algebraic percentages for stems on *old*, *recent*, and *current* exams. 127

Figure 27. Box and whisker plots of relative symbolic-algebraic percentages for responses on *old*, *recent*, and *current* exams. 128

Figure 28. Box and whisker plots of relative numeric-numeral percentages for stems on *old*, *recent*, and *current* exams. 131

Figure 29. Box and whisker plots of relative numeric-numeral percentages for responses on *old*, *recent*, and *current* exams. 132

Figure 30. Box and whisker plots of relative percentages of written stems for *old*, *recent*, and *current* exams. 135

Figure 31. Box and whisker plots of relative percentages of written responses for *old*, *recent*, and *current* exams. 136

Figure 32. Example of the labeling of representations on the *current* August 2017 exam, item 15. 139

Figure 33. Prevalence of representation relationships per exam administration for *old*, *recent*, and *current* exam versions. 140

Figure 34. Sankey diagram of the medians of stem-to-response relationships for *old* exams. .. 141

Figure 35. Sankey diagram of the medians of stem-to-response relationships for *recent* exams. 143

Figure 36. Sankey diagram of the medians of stem-to-response relationships for *current* exams. 145

Figure 37. Box and whisker plots of relative representation treatment percentage for *old*, *recent*, and *current* exams. 154

Figure 38. Box and whisker plots of relative representation conversion percentage for *old*, *recent*, and *current* exams. 156

Figure 39. Box and whisker plots of relative representation synergy percentage for *old*, *recent*, and *current* exams. 158

List of Appendices

Appendix A *Directions for Readability Formulas* 216

Appendix B *Definition of Contextual Situations* 217

Appendix C *Cognitive Demand Complexity Framework* 218

Appendix D *Conceptual Categories and Subcategories for Multiple Representations* 221

Appendix E *Representations Relationships Framework* 228

Readability, Complexity, and Representations: A Holistic Analysis of New York State Algebra Exams

Chapter 1 – Introduction

Algebra is important for students to master as it encourages students to recognize patterns, understand the rules of algebraic representations, and manipulate algebraic symbols that will be present in future courses in mathematics ("Algebra as a strand of school mathematics for all students", 2014). The earliest algebra can be found in Mesopotamia from 4000 years ago, and since then, its rules and wording have evolved based on educational theory and research of the time, and more recently, standards movements (Katz & Barton, 2007). Over the years, algebra was considered as a "gatekeeper" for more advanced mathematics classes (Martin, Gholson, & Leonard, 2010; Spielhagen, 2006; Usiskin, 1995; Usiskin, 2004). Consequently, authors of position papers have promoted early algebra thinking problem solving (Katz & Barton, 2007; Notari-Syverson & Sadler, 2008; Stevens, Blanton, Knuth, Isler, & Gardiner, 2015) and various researchers have concluded that algebra should be presented in elementary grades to prepare students for the rigors of algebra later on (Blanton & Kaput, 2005b; Carraher, Martinez, & Schliemann, 2008; Carraher, Schliemann, & Brizuela, 2000; Carraher, Schliemann, Brizuela, & Earnest, 2006; Smith, 1996; Van Amerom, 2003). Since algebra is essential to learn, multiple international, national, and state organizations have developed and implemented testing programs to assess how well students understand it.

Algebra is assessed on international assessments such as the Programme for International Student Assessment (PISA Test in Mathematics), Trends in International Mathematics and Science Study (TIMSS), the *current* national Algebra I Common Core examinations (2014-Present), and the previous state exam programs including the *old* Mathematics A (2002-2008)

and *recent* Integrated Algebra (2008-2015). The National Council of Teachers of Mathematics has the following position: “Algebra is not confined to a course or set of courses in the school curriculum; rather, it is a strand that unfolds across a pre-K–12 curriculum. Students experience the algebra strand as a way of thinking and valuing structure that develops across grades” (“Algebra as a strand of school math”, 2014). New York State Board of Regents directly cites the NCTM process standards, and also promotes the teaching of algebraic thinking across grade levels. New York State students have to pass a mathematics exam with algebraic concepts before they can graduate high school.

After the No Child Left Behind Act of 2001 was passed to the current day, there were three different sets of state algebra standards used in New York State. Each set of state algebra standards had their own corresponding high stakes state algebra exam. These standards and their corresponding state algebra exams are referred to as the *old* Math A (2002-2008), *recent* Integrated Algebra (2008-2015), and *current* Algebra I Common Core (2014-Present). The old and recent standards and exams were developed in New York State while the current standards were developed by the national Common Core organization, and the exams were developed from those standards. From 2002-2018, there were traditionally three administrations of the exams in June, August, and January. Most students would take the June administration of the exam in either ninth grade (*recent* and *current*) or tenth grade (*old*). Advanced students could take it one year before, and students who struggled in mathematics could repeatedly take it before they graduated from high school. This was the first algebra Regents exam that they would take in New York State. Students are required to pass one Regents exam, and typically, they would pass the state algebra exam as other state mathematics exams required the mastery of this state algebra exam. Although the *old*, *recent*, and *current* standards in New York State included

different topics, they all mentioned that representations in mathematics were important to master along with applying the math to real world contexts. To understand the similarities and differences of each New York State exam, the readability, real world context, cognitive complexity, representations, and the relationships among representations were labeled on each New York State algebra exam item and then analyzed to determine patterns.

The *old*, *recent*, and *current* exams all had two sections: multiple choice and constructed response. Multiple choice items had one stem which was either a question or statement that the students needed to understand and answer. These items had four responses or possibilities of the answer with one of the responses being the correct response for the stem. Constructed response items could have one stem and one response, but they may have multiple stems and responses. Thus, constructed response items may consist of multiple item parts because of having multiple stems. A constructed response item may have two stems in the item, or in other words, the item has two item parts that need to be answered in order to receive full credit for the item.

Readability

Readability, or the ability to read and understand the wording of an item, is an important part of algebra exam items. Different words and phrases can signal different types of algebraic processes. Students who struggle with reading may also struggle with algebra items that include a significant portion of written directions. In this study, the wording was studied without the algebraic notation or notation on figures such as graphs or in tables. The level of readability has been measured with many different formulas over time. Readability formulas typically work in two ways: either a mathematical calculation using the number of syllables, words, and sentences such as the Flesch Reading Ease Formula (Kincaid, Fisburne, Rogers, & Chissom, 1975) and Flesch-Kincaid grade level (1979) or using a dictionary of familiar words that are leveled by

grade level combined with a calculation of the syllables, words, and sentences such as the Lexile framework or ATOS framework (Lennon & Burdick, 2004). Fortunately, there are calculators online that calculate these grade levels. By using both types of formulas along with counting words and sentence length, patterns of readability can be determined among exam items.

Real World Context

Items on mathematics exams are often written in a real world context to promote problem solving beyond the test. Context is separated from other representations. Visual, written, numeric, and symbolic representations illustrate context for the student. For instance, there may be a table of years of experience and salaries of workers along with a description of how the salary data was gathered. Also, the students may be required to find the line of best fit which would be symbolic, and then use the formula to find a potential salary at a certain number of years of experience. All of these representations are simultaneously working together. Other real world context problems may only include one type of representation. This shows how mathematics would be used later on in a future career and/or life setting for the student. It helps students make the connections between the mathematics that they are learning and the world around them.

If students are familiar with the real world context, then it may be easier to use different methods to solve the problem. Word problems that include terminology in an unfamiliar real world context may confuse students. For instance, although a problem about a silo may be a real world situation and make sense to students who live in a rural area, it may not be familiar to a student who lives in a city. When the final answer is found, students who understand the concept of a silo may be able to check their work based on the appropriateness of the answer. Unfortunately, a student who does not understand this concept would have to rely on solving the

problem without the real world connection and without the appropriate vocabulary to decode the problem situation.

Some parts of algebra, and even the subject of algebra itself, have previously been incorrectly classified as mathematics without connections to the real world context. For example, people in the world may not be asked to solve a two-step linear equation such as "$2x + 4 = 10$" in everyday life. This often leads to disconnect in how mathematics is integrated into daily life, which can demotivate students. However, the idea of working backwards to find initial values can be found in a variety of real world contexts such as in personal finance where a person may need to figure the beginning balance on a credit card. A heavy reliance on the procedural components of algebra may distract from the underlying conceptual connections to the real world. Instead of students learning about how different steps in a single algebraic problem situation can be generalized among various topics in mathematics, students may learn each set of processes as independent and devoid of connections to more familiar contextual situations. Knowledge of context and relevant vocabulary may encourage students to develop their own approach in solving problems in non-algorithmic ways or the highest level of algebraic reasoning that students can demonstrate.

Cognitive Complexity

Some items on the state algebra exam may require higher levels of cognitive reasoning in order to solve the problem. Working memory, inhibition, and attentional switching are cognitive processes that greatly affect how students process and interpret mathematical processes. There is a difference between problems that can be answered by the memorization versus mathematics that need to be answered in complex, non-algorithmic ways. Non-algorithmic, creative problem solving is an essential part of mathematics instruction and learning as students who will use

mathematics in their career fields will need to solve problems that may not have a developed procedure. In presentations, students may need to justify their newly developed procedure or their solution to a problem to a group of people. This type of reasoning ability is developed with more cognitively complex reading strategies and cognitive processes such as working memory and attentional switching (Fuchs, et al., 2010). The *current* state algebra standards in New York discussed how these standards would include more opportunities for non-algorithmic thinking.

Representations and Multiple Representations

In previous research, I counted the number of times different signal verbs appeared in the *current* Algebra I Common Core exam (Simpson, Frech, Garbade, & Schaefer, 2018). The most common signal verb among the current exam items was the word "represent". This aligns with the Common Core's initiative to promote deep conceptual understanding by encouraging students to solve items on examinations using multiple representations ("Regents exams – Mathematics: Multiple representations", 2013). Instead of just being required to "solve", students are being required to show how an algebraic concept is connected among its applicable representations.

There are many different types of representations that are used in the mathematics classroom and have been a part of mathematics educational research. From Bruner's (1966) enactive (objects or actions), iconic (pictures), or symbolic (symbols or words) mathematical representations to the Lesh, Post, and Behr (1987) math representation framework of visual, physical, symbolic, contextual, and verbal (or also known as written) representations (Lesh, Landau, & 1983; Lesh, Post & Behr, 1987), using different types of representations in mathematics has been important in the teaching and learning of mathematics (National Council of Teachers of Mathematics, 2000; National Council of Teachers of Mathematics, 2014). NCTM

(2000) discussed how development and organization of representations aids in the organization of mathematical thought and how representations extend mathematical thinking.

Representations are defined in various ways in the different standard systems and frameworks. NCTM has used the Lesh, Post, and Behr (1987) framework to describe representations as physical, verbal (or written), visual, symbolic, and contextual. Other standards systems seem to also follow the same framework. TIMSS described representations as tables, graphs, words, and symbols. PISA described representations as "…graphs, tables, diagrams, pictures, equations, formulae, and concrete materials" (PISA, 2015, p. 11). The New York State exams have also similarly defined representations. In the *old* standards (Math A), representation was mentioned as a specific topic area (Modeling/Multiple Representation) for the exam. In the *recent* and *current* New York State standards, representation was mentioned as a process standard that should be integrated across the entire curriculum so that students understood representations as a part of everyday topics.

Students should create representations for mathematical communication, appropriately select and translate between representations, and model mathematical situations (National Council of Teachers of Mathematics, 2000). Students need to not only be fluent with different types of representations, but be able to compare, contrast, and evaluate them for effectiveness. This is not only a part of NCTM's process standards, but a part of the PISA and TIMMS testing programs, national Common Core standards, and New York State's *old*, *recent*, and *current* standards.

Even though mathematical representations in Algebra are universally deemed important by multiple international, national, and state organizations, there is not a significant amount of analysis that has occurred combining standards and representations. At a New York State level,

knowing that certain standards have typically used a specific type of representation or representations can help teachers prepare students for these representations in class instruction. When students are in the process of learning algebra, multiple representations has the potential benefit of building long lasting conceptual meaning. For example, the symbolic algebraic linear equation can be manipulated in its current form to derive specific points on the line. However, this symbolic algebraic equation (i.e., $y = 2x + 5$) can be translated into a set of points, table of values, and visual graph representation. The symbolic algebraic equation can also represent a real world situation. When a symbolic algebraic equation is associated with these other representations, students can see the pattern of how slope is determined through these different representations. With iterations of the concept of slope in its various representational forms, students can understand the concept of slope rather than the "recipe" of how it is found without the conceptual basis. Understanding the topic of slope fully helps students understand concepts like the average rate of change which is when the slope concept is applied to nonlinear functions.

Representations often do not work singly; instead, they are considered in relation to other representations. These representation relationships may affect how a representation is perceived and how much knowledge the student needs to know about the representation. Some items on exams may require a representation treatment or when the manipulation occurs within one type of representation. For example, solving an algebraic equation to find an unknown variable would be a representation treatment. Another type of representation relationship would be the representation conversion. This occurs when one representation leads to another representation. For example, if the reader has to find a solution for a graphed system of equations, this would be an example of a representation conversion as they are looking at the graph of the system of equations and finding the point that represents the solution. Finally, representation synergy

occurs when two or more representations work together simultaneously. This often occurs in problems where the reader has to make sense of the answer in a context and will need to go between a graph and the algebraic symbols in order to explain his or her answer.

Problem Statement

Mathematical standards and topic areas are often used to map out the school year. The *current* standards even encourage teachers to use a certain number of days for instruction for a particular unit. State algebra items have been previously labeled with their corresponding standard along with the *recent* and *current* standards (Schaefer, 2019b). With the amount of allotted time, teachers focus on covering the material of the curriculum, and may not see other patterns that may be affecting the conceptual understanding of their students. This study aims to provide teachers with a more holistic view of state algebra exam item parts by looking at their readability, context, cognitive complexity, representations, and relationships between or among the representations. By analyzing each state algebra item, trends can be found across *recent* and *current* standards as well as by state algebra exam type (*old*, *recent*, and *current*).

Significance of the Study

Students struggle with mathematics for many different reasons, and it is not always about the content learned in the mathematics class. Since the curriculum is standards-driven, teachers focus on students mastering the standards rather than looking more at the components of the item parts. A student who does not understand how to use and/or interpret a representation may struggle on many different topic areas, and may struggle with any item referencing this type of representation. Students who have language disabilities and/or are speakers of other languages may have difficulty decoding the language of the problem. The real world context could enhance the problem solving abilities of the students, or it could distract from the actual solving of the

item. Students may understand the mathematics, but struggle with reading the item because of its readability level. Finally, some standards on average may be more cognitively complex. For the student who has difficulty in mathematics, although they may learn the whole curriculum, there may be ways to differentiate the curriculum so that they can access the curriculum at their own individualized pace and can draw comparisons between different topic areas.

In order to obtain a holistic view of algebra exam items, each item or item part was coded with readability calculations, presence of "real world" context, cognitive complexity, math representation type, and the representations relationships. Items are comprised of stem(s) and response(s). While the multiple choice items are comprised of a stem and four responses, one of which being the correct answer, constructed response items may have multiple stems and responses. Each of these constructed response items that had multiple parts were considered to have multiple item parts. First, the item's readability was calculated using three readability formulas: Flesch-Kincaid Reading Level, Flesch Reading Ease Level, and ATOS Level which were found using Microsoft Word's readability check and the ATOS online calculator. Also, the word count, average word length, average sentence length, and average vocabulary level were determined by using the ATOS online calculator. Second, items as well as their corresponding item parts were coded with either the presence or absence of a "real world" context. Third, the cognitive complexity of each individual item part was determined based on the cognitive complexity rating scale. Fourth, the item part's *stem* and *response(s)* was coded by the categories and subcategories applicable to the math representation(s). Fifth, item parts were coded with the type of relationship among representations.

Research Questions

The research questions can be divided into the five different coding schemas: readability, "real world" contexts, complexity, representations, and representation relationships:

Schema 1: Readability

1. What is the average grade level and average level of reading ease for each exam (*old, recent,* or *current)* and item part type (constructed response, multiple choice)?

Schema 2: Context

2. Do certain exams (*old*, *recent*, or *current*) rely more on a "real word" context? Which item part types (constructed response, multiple choice) rely the most on using "real world" contexts?

Item Readability and Context

3. Is there a correlation between item readability and presence of context?

Schema 3: Complexity

4. What is the average complexity of the items labeled by exam (*old*, *recent*, and *current*)? What is the average complexity of items labeled by item part types (constructed response, multiple choice)?

Item Readability, Context, and Complexity

5. Are there correlations among readability, context, and complexity?

Schema 4: Representations

6. How many representations appear on average for each standard and overall for each exam (*old*, *recent*, or *current*)? What types of mathematical representations appear most frequently by item part types (constructed response, multiple choice)?

Schema 5: Representation Relationships

7. How do representations interact with one another in the *stem* and *response* of the item? Do certain types of relationships occur with representations occur on certain exams (*old*, *recent*, or *current*) or by item part types (constructed response, multiple choice)?
8. What types of representation relationships occur most frequently according to exam (*old*, *recent*, or *current*) and item part types (constructed response, multiple choice)?

In this research, I will be referring to multiple terms and phrases that are defined below.

Definition of Terms

Algebra standards refer to the set of objectives adopted by New York State in the state algebra course. In this research, there are three sets of standards considered: *old*, *recent,* and *current* standards. These three sets of standards were selected because the exams were administered after the No Child Left Behind Act of 2001. Each standards set is connected to corresponding curriculum and exam in this time period. When a different set of standards were developed, the teachers would need to implement the latest curriculum and prepare students for the latest exams.

New York State coding refers to the labeling of standards for each exam item by the New York State Board of Regents. The *old* exams were labeled with the topic of the *old* standards, the *recent* exams were labeled with the topic of the *recent* standards, and the *current* standards were labeled with the topic and subtopic of the *current* standards.

JMAP coding refers to the labeling of standards for each exam item by JMAP.org, a respected website used by New York state teachers for preparing students for mathematics Regents exams in New York State. It is currently the only website that has personally classified

state algebra exam items by standards and specific topic areas. The JMAP.org website was created in 2003 by Steven Watson and Steve Sibol (Sibol & Watson, 2019). These first-year mathematics teachers at Thomas Jefferson High School in Brooklyn were surprised at the difference between the Math A test and their teaching practices; so, they decided to analyze the tests and create free mathematics resources for teachers. Steve Watson previously studied the progression of topic areas in mathematics from 1866 to the present and provided labeling of test items on the *old* (Math A) and *recent* (Integrated Algebra) exam by the recent standards (Watson, 2010). JMAP also labeled test items on the *current* (Algebra I Common Core) exam by the current standards with the strand, topic, and subtopic areas as well as the standard number. The goal of JMAP has been to provide teachers and students with free materials based on its database of test items sorted by mathematical topic.

State algebra exams refer to the three exams developed by New York State based on the *old*, *recent*, and *current* standards. Each set of standards matches each exam. For instance, by implementing the *old* standards into instruction and curriculum practices, teachers would prepare the students for the *old* exam. The *recent* standards would not completely prepare students for the *old* or *current* exams.

Topics refer to the first and broadest level of classification. The *old*, *recent*, and *current* standards all had topic areas. The New York State answer keys for *old*, *recent*, and *current* exams had their items classified by topic areas.

Subtopics refer to the second level of classification for the standard that would divide each topic. This type of classification was only present on the *current* exam answer keys.

Old exams refer to the exam items written with the Math A standards. The *old* standards were developed by New York State previous to the 1997 Mathematics A Pilot Exam. The first

administration of the exam was June 2002 and the last released administration of the exam was June 2008. The *recent* standards were developed in New York State and presented to the New York State Board of Regents on March 15, 2005. The first administration of the exam was on June 2008 and the last released administration of the exam was June 2015. There were two other special administration exams, but since these exams were not released to the public, they were not a part of the database. In the *recent* exams, there was one item omitted on the June 2013 exam. This exam item was omitted due to a "lack of specificity" and all students were given credit on this item. Therefore, only 857 items were gradable items on the *recent* exam.

Current exam items refer to the exam items written with the Algebra I Common Core standards. The *current* standards were developed from the national Common Core standards. In April 2009, the development of the Common Core standards commenced as a multiple state initiative and by June 2010, there was a finalized version of the Common Core standards. New York State adopted these standards on July 19, 2010. The first administration of the exam was in June 2014, and the Common Core standards are still the basis of state algebra exams today. *Old, recent,* and *current* state algebra exams were broken down multiple choice and constructed response items. They were also labeled by various topic areas. *Current* state algebra exam items were also labeled by various subtopic areas.

Table 1

Item Counts from Old, Recent, and Current State Algebra Examinations by Item Type and Value

Exam	*Item Type and Value*					*Totals*
	MC* 2 points	CR 2 Points	CR 3 Points	CR 4 Points	CR 6 Points	
Old						
June 2002 – June 2008	554	110	48	52	0	764
Recent						
June 2008 – June 2015	653	72	66	66	0	857
Current						
June 2014 – August 2018**	336	112	0	56	14	518
Totals	1543	294	114	174	14	2139

*MC stands for multiple-choice and CR stands for constructed-response.

** The *current* exams are still offered. The number of items reflects the time period of June 2014 to August 2018.

Table 2

List of Topic Areas for Old, Recent, *and* Current *State Algebra Exams.*

Old June 2002 – June 2008	*Recent* June 2008 – June 2015	*Current* June 2014 – Present
Mathematical Reasoning	Number Sense and Operations	Algebra
Number and Numeration	Algebra	Functions
Operations	Geometry	Number and Quantity
Modeling/Multiple Representations	Measurement	Statistics
Measurement	Statistics and Probability	
Uncertainty		
Patterns/Functions		

Table 3

List of Current *Subtopics, by Topic Area, Noted on Exam Items Found on the* Current *State Algebra Exams*

Algebra	*Numbers and Quantity*	*Functions*	*Statistics*
Arithmetic with Polynomials and Rational Expressions Creating Equations Reasoning with Equations and Inequalities Seeing Structure in Expressions	The Real Number System Quantities	Building Functions Interpreting Functions	Interpreting Categorical and Quantitative Data Linear, Quadratic, and Exponential Models

Exam items refer to the gradable multiple choice and constructed response items found on the *old*, *recent*, and *current* exams. The *stem* of the item comes first and it is the statement of the problem usually indicated by either a question or phrase. *Stems* could include various types of representations and wording to articulate the question. A correct *response* follows after the stem and it is the answer based on information from the *stem* and/or other *responses*. For multiple choice items, the *responses* would be the four answer choices and the student would need to select the answer to the *stem* based on the information given in the *stem*. For constructed response items, the *response* would be the written answer to the *stem*. Sample *responses* for the constructed response items were found on JMAP's website for the *old* exams and on the sample (or model) response set for the *recent* and *current* exams.

Exam item parts refer to the exam broken down into its individual parts for each exam item. While multiple choice items have one part, constructed response items may ask for two different answers as a part of the grading for that item. These separate, but often related answers were considered as a part of the item. For instance, on the August 2018 *current* New York state

algebra exam, there were 37 gradable items but 45 parts of the items that would need to be answered by the learner.

Readability refers to a numerical estimate of the ease of understanding the written representation of the item. By using Microsoft Words' readability calculator and an online readability calculator, an estimate of the grade level and reading ease may be determined ("ATOS text analyzer", 2020). The calculated *grade level* would be comparable to the grades in school (kindergarten through college level). The *reading ease* would range from very easy to understand text to very difficult text. Other methods of readability such as word count, average word length, average sentence length, and average vocabulary level were also considered as measures of readability. Appendix A further describes readability.

Context refers to the real world situation of the item if applicable. Some items may use a real life scenario in the mathematics item. Appendix B fines the *context* in schema 2.

Complexity refers to the level of thinking required to answer the algebra item. This thinking ranges from memorization to doing mathematics. Appendix C further defines *complexity*.

Mathematical representations are the multiple ways in which abstract concepts in mathematics can be represented. *Mathematical representations* can be broken down into categories and subcategories. The four categories in my dissertation are *visual*, *symbolic*, *numeric*, and *written*. The *visual* category can be broken down into five subcategories: graphs, charts, geometric figures, diagrams, and pictures. The *symbolic* category can be broken down into six subcategories: algebra, sets, intervals, probability and statistics, geometric notation, and sequences and series. The *numeric* category can be broken down into three subcategories: numerals, points, and tables and matrices. The *written* category is not broken down into any

subcategories. Appendix D further defines these categories and subcategories with examples from *old*, *recent*, and *current* state algebra exams.

The *relationships between* or *among representation(s)* refer to the ways in which similar or different representations may interact with one another. In a state algebra exam item, a single mathematics concept may be illustrated in the same and/or different representations. Appendix E further illustrates relationships between or among representations on state algebra exam items.

Overview of Dissertation

Chapter 2 covered a review of the literature on standards research, readability, context, cognitive complexity, mathematical representation, and relationships among representations along with a theoretical framework of semiotics. Chapter 3 described the methodology of the study. State algebra test items were previously classified by each of their corresponding state standards, then by the *recent* standards, and the *current* standards. These state math items were classified by readability, the presence or absence of "real world" context, cognitive complexity, the types of representation (visual, written, numeric, and symbolic), and the overall relationships between the representations in terms of the stem and response. Chapter 4 was a discussion of the analysis of research based on the research questions. Chapter 5 included a summary of the findings, an acknowledgement of the limitations in the study, and a discussion of recommendations.

Chapter 2 – Review of Literature

This chapter is divided into six major sections: the history of the standards and accountability movements, research on readability of mathematics, context, complexity, mathematical representations, and relationships among representations, the limited nature of the corpus of research literature for each of the four previous sections, and the semiotic theoretical framing for the study. The first section describes the beginnings of using high stakes testing with students in the United States and then focuses on New York State mathematics Regents exams and the future of a Common Core exam. The second section reviews the psychology of cognitive load in mathematics. It focuses on a modified Lesh, Post, and Behr (1987) framework to illustrate how representations have been studied separately. The third section reviews how readability has been assessed in mathematics. Then, the fourth section, research on context is described. The fifth section describes how cognitive complexity has been measured and applied in mathematics teaching. The sixth section describes research on mathematical representations. The seventh section describes the limitations of the current research on mathematical standards and the limitations of the research on readability, context, cognitive complexity, cognitive load, mathematical representations, and relationships among representations. Finally, the eighth section presents the theoretical framework of mathematical semiotics and describes the idea of conversions between registers.

History of Standards and Accountability Movements

In 1963, the Carroll model of school learning focused on the alignment between student aptitude and achievement. For students to achieve, there needed to be quality instructional events and an alignment between classroom instruction and assessment (Carroll, 1963). The alignment between instruction, assessment, and the overall quality of teachers were further discussed in *A*

Nation At Risk (National Commission on Excellence in Education (NCEE, 1983). After this publication, there were three separate time periods: the standards movement of the 1980s and 1990s, the accountability-based reforms of the 2000s, and the current era of alignment and accountability (Coburn, Hill & Spillane, 2016). All of these periods of time greatly affected algebra assessments nationally, and in particular, New York State.

In the 1980s and 1990s, there were standards created by states and professional organizations such as the National Council of Teachers of Mathematics (1989) publication of *Curriculum and Evaluation Standards for School Mathematics.* This document encouraged a progression toward algebraic thinking in the middle and high school grade levels. During this time, researchers began to explore the alignment between standards and assessments across states. For example, Webb (1999) found that assessment items in multiple states were not reflective of the full range of the states' standards. Instead, algebra assessments may only address procedural understanding rather than both procedural and conceptual understanding which would lead to a higher level of the depth of knowledge of the student. Formative assessment techniques were also developed to address the standards in the classroom (Van Den Heuvel-Panhuizen, 1996). This included meaningful and informative problem solving situations in all levels of mathematics, and a focus on creating realistic problems that the students could solve.

The accountability-based reforms of the 2000s increased the number of tests in subjects and grade levels to help decrease achievement gaps between schools in the same district and schools across a single state. The National Council of Mathematics published the *Principles and Standards in School Mathematics* in 2000 (National Council of Teachers of Mathematics, 2000) which encouraged schools to build high quality mathematics programs. This document included the Lesh, Post, and Behr (1987) representational forms of mathematics and mentioned them as

important components of mathematics instruction and assessment. Examinations became "high stakes" according to the No Child Left Behind Act of 2001, and teaching became more focused on testing as it affected teachers, students, administrators, and school funding (No Child Left Behind Act [NCLB], 2002). The first year of high school math, or often times, the first year of algebra, became the high stakes test that would determine if a student was able to graduate high school.

The No Child Left Behind Act encouraged more rigorous standards and assessments for students but did not always clarify the learning expectations in all state documents (Reys et al., 2006) which can cause teachers to spend more time analyzing standards (Boardman & Woodruff, 2004) and relying on textbooks (Grouws et al., 2013). International and state assessments from NCLB greatly impacted instruction more than before such as adding algebra earlier into middle school topics to align with TIMSS assessments (Schmidt, Wang, & McKnight, 2005; Jacobs, et al., 2006). By using national data, researchers found that students were sometimes scoring higher in algebra (Reys et al., 2003) but there were also mixed effects in middle school achievement when considering the NAEP scores from 1990 to 2009 (Lee & Reeves, 2012).

Although there was more focus on assessment in the era of accountability, researchers focused on changing the instructional process from a "traditional approach" such as student investigations of mathematical content (Remillard & Bryans, 2004), mastery-based learning (Zimmerman & Dibenedetto, 2008), cognitively demanding middle school math (Silver et al., 2007), and alternative assessment practices (Lianghuo, 2004; McCafferty et al., 2002). Some teachers still wanted more traditional elements of the classroom and explicit objectives (Lloyd & Behm, 2005; Gabriele & Joram, 2007). This reform curriculum was more conceptually based and promoted student understanding of mathematics (Cai & Moyer, 2010; Cai, Nie, & Moyer,

2010), as traditional curriculum caused gaps in mathematical understanding (Sfard, 1991). Researchers at the time started to see that students who learned standards-based curriculum outperformed students taught with traditional approaches on assessments (Riordan & Noice, 2001). Reform curriculum required training and support (Smith, 2000; Handal & Herrington, 2003; Drake & Sherin, 2006; Charalambous & Philippou, 2010) to align with end-of-year assessments (Fan & Zhu, 2008; Kurz, Elliott, Wehby, & Smithson, 2010). Teachers also needed training on high level cognitive instructional tasks (Boston & Smith, 2009; Baumert et al., 2010) for tasks such as student self-assessment of solutions (Andrade, 2010).

In the current era of accountability and alignment, there has been a vision for a national high stakes assessment that all states would take in order to promote high levels of mathematics for all students. The United States Department of Education has aligned standards from multiple states when considering the Common Core standards. Bandeira de Mello (2011) found that in the years 2005 to 2009, most states' standards (12 out of 39) were below the basic level including New York State. In 2013, New York States' NAEP equivalent score was in the Proficient range (Bandeira de Mello, Bohrnstedt, Blankenship & Sherman, 2015), but in 2015, New York State was excluded from the analysis because of the changes to the Common Core curriculum (Bandeira de Mello, Rahman, & Park, 2018).

To promote more reform based methods and ultimately the combination of both procedural and conceptual learning methods, the Common Core mathematics standards were developed to emphasize students' demonstration of their understanding (Porter, McMaken, Hwang, & Yang, 2011) and aimed to build connections between the process of arithmetic structure and algebra (Wu, 2014). However, this change of standards did not come without criticism from teachers and researchers. Although the Common Core claimed to eliminate topics

to give more time to conceptual understanding, Germain-Williams (2014) found that the Common Core texts did not reduce the number of topics or number of problems presented in the chapters. For New York state and in future national assessments, the standards, curriculum, and assessment would need to be aligned in order to prepare students not only for the culminating exam, but also for future mathematics courses and future mathematical thinking.

In the future, national standards movements like the Common Core Standards, whose foundations are rooted in previous NCTM standards, aim to have shared expectations among school districts across the United States, more focused standards, and a more efficient and higher quality testing process with states using shared assessments (Porter et al., 2011). However, logistically, algebra is still not being taken by all students, and not all students are prepared to take this course (Remillard, Baker, Steele, Hoe, & Traynor, 2017) especially because of previous levels of elementary math ability (Venenciano & Heck, 2016; Siegler et al., 2012). Students may have missing background information that can prevent them from accessing algebra as comprehensiveness and balance of content can be different from school to school (Herman, Webb, & Zuniga, 2007). The requirement to take algebra may cause students to drop out of high school (Plunk, Tate, Bierut, & Grucza, 2014). More states would need to make algebra a requirement for graduation in order for a national assessment to occur.

In a national assessment, there needs to be a careful consideration of the standards (Schoenfeld, 2015) so that the standards could influence the assessment. Standards and objectives from different states may not be clear to all reviewers involved in a coding process (Webb, 1999; Webb & Herman, 2006). This has been found even in New York State (Fasano, 2014). Enacting official curriculum requires multiple levels of interpretation and implementation

which explains why curriculum may vary across systems of education (Henningsen & Stein, 1997; Remillard & Heck, 2014).

Although mathematical representation has been mentioned as an important part of instruction by the TIMMS (2015), PISA (2015), NCTM (2014), Common Core Framework (2011), and current adoption of the New York State Common Core, there has not been direct connection between standards on these assessments and the types of representations that appear on the assessments.

New York State.

The era of alignment and accountability truly affected New York State mathematics curriculum and the state algebra exams. State algebra mathematics exams from 2002 to the present were developed after the No Child Left Behind Act in the accountability time period during the current era of both alignment and accountability. Since the mathematics Regents examinations administered in New York State are released to the public and are readily available on the New York State Education Department (NYSED) website, the alignment of the test items to standards can be measured. During the *old* standards (Math A) and *recent* standards (Integrated Algebra), New York State only released the topic area of the standard while the *current* standards (Algebra I Common Core) released the topic, subtopic, and standard areas. Other websites for New York State algebra assessments only address the *current* standards and/or offer paid services to prepare students for the various mathematics Regents exams in New York. There are many websites that discuss algebra in general, but the topic areas do not exactly match New York's Regents assessments. Out of all New York mathematics websites, JMAP.org (JMAP) is the most comprehensive resource for mathematics teachers as they have analyzed

standards and topic areas on assessments from 1866 to present day as a part of Dr. Watson's previous research (Sibol & Watson, 2019).

Previously, I looked at New York State's Regents algebra exams from 2002 to present to see how the exams have changed overall. Previously, I labeled all state algebra exam items with the *recent* and *current* standards to see trends in topic areas (Schaefer, 2019a) and have informed New York State teachers in professional development about how to use this information to reuse *old* and *recent* items in helping students understand the basics of the mathematical topics (Schaefer, 2019b; Schaefer, 2019c). These ideas led into this research to find out more about the exams than the standard, and to see if there were major changes in readability, "real world" context, complexity, representations, and representational relationships.

Summary.

From the Carroll model of school learning (1963) to present day, there is a logical progression of why the Common Core standards were developed. Its overall objective is to provide students from all over the United States with high quality learning opportunities. The Common Core standards have made several claims about its curriculum. It has stated that multiple representations will be used and students will need to know the relationships among similar and different representations. The reading level of the Common Core materials will be more challenging than prior assessments in both English language arts and math, and the mathematics items will be more cognitively complex that prior assessments based on previous standards. The following sections will describe the readability, context, cognitive complexity, representations, and representation relationships along with the semiotic framing of the study.

Research on Readability in Mathematics

History and research of readability formulas.

Readability and topic areas affect student engagement and performance in mathematics (Walkington, Clinton, Ritter, & Nathan, 2015). It can be used as the quantitative part of text complexity analysis as described by the Common Core ("New research on text complexity", 2017). There have been many different formulas created to measure the readability of texts. One of the most famous formulas was created by Flesch. Flesch created a reading ease formula in 1948. It was later adapted by with Kincaid in 1979 for use in the Navy (Kincaid, et al., 1975; Flesch, 1979). Later on, Flesch (1979) created a reading ease score. It has been used to understand the readability of many different texts, but in particular, it has been used to analyzed readability for Certified Nursing Assistant exams (Neuhoff, Feeser, Sutherland, & Hovatter, 2016) and the *current* New York State algebra 1 exams (Spoth, 2016).

After the development of the Flesch reading ease score, there has been a variety of researchers that have used this formula to understand the differences in readability in textbooks and other reading resources. Johnson (1957) used a modified Flesch formula that counted a word for each digit of a large number similar to Tinker (1954). He found that informative material in mathematics textbooks was on average more difficult than explanatory, problem material, or explanatory and informative material together.

Different formulas after the Flesch formula included counting the number of unfamiliar words. The Dale-Chall (1948) formula was developed by Edgar Dale and Jeanne Chall and the Spache (1953) formula was developed for primary grade levels. These formulas are different than the Flesch formulas because they rely on a dictionary of familiar words. Heddens and Smith (1964) measured five textbooks in grades 1-6 for readability using the Spache formula for the

grades 1-3 texts and the Dale-Chall formula for grades 4-6. They also found that these math texts were above grade level. Smith (1965) used the Spache and Dale Chall to analyze elementary mathematics materials. He modified the Spache formula by counting the words every time they appeared in the textbook. After these studies came out, Kane (1970) questioned how Dale-Chall, Flesch, and Spache formulas can be applied to mathematics when they were not specifically meant for mathematics textbooks (Kane, 1970). Another similar formula to the Dale-Chall and Spache formulas was the McLaughlin-SMOG readability formula (1969) as it was created based on the difficulty of words. It has been previously used with the readability of New York State Regents Physics exams (Zawacki, et al., 2003).

Readability tests often require at least 100 words in order to analyze the item problem. With the growing criticism of using readability tests for analyzing assessment, Homan, Hewitt, & Linder (1994) devised a formula specifically for measuring multiple choice items that may only have one sentence in the item (Helwig, Rozek-Tedesco, Tindal, Heath, & Almond, 1999; Hewitt & Homan, 2004) which was then used to measure the readability of New York State United States History and Government exams (Morton, 2015). However, this formula requires that unfamiliar words are identified. In Morton's (2015) thesis, he found that it took 20-40 minutes to calculate the readability of an item. Other readability tests using unfamiliar words have calculators online that find the level of readability more quickly.

Two modern types of measuring text readability include the ATOS formula and Lexile formula. These formulas also take into account the level of difficulty of the words. By studying the Lexile of each mathematics item on the Texas Assessment of Knowledge and Skills, Lamb (2010) found that the reading levels affected the student performance on the exam. However, the Lexile measure is not typically used with the readability of mathematics. Instead, Lexile has

created a measure called the Quantile that would measure the familiarity of the skills and concepts of mathematics ("What is a quantile measure", 2019). The Harris-Jacobson 2 formula was suggested when studying mathematics terminology (Paul, Nibbelink, & Hoover, 1986) which later evolved into the ATOS formula (Milone, 2014). The ATOS formula includes a graded vocabulary list in its formula and has been endorsed by the Core Standards website ("New research on text complexity", 2017).

One of the major changes in the Common Core was an increase in text complexity as the levels of reading complexity were purported to have declined over the last fifty years. Although the Common Core has stated that reading texts had to become more complex because of the recent decline in text complexity levels, Gamson, Lu, and Eckert (2013) found that the reading levels of third and sixth grade texts have either increased or stayed the same for the last seventy five years.

Linguistic factors for readability research.

Other studies have used basic word count and count of linguistic structures to measure readability. Word count was used to determine readability. For instance, the number of words seemed to affect how eighth grade students performed on word problems as more words led to fewer students succeeding on the problem (Lepik, 1990). Other researchers looked at the linguistic properties of the word such as with PISA and TIMSS assessments (Walkington, Clinton, and Shivraj, 2018) Kansas general assessments (Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006), sixth grade mathematics tests versus reading fluency (Helwig, et al., 1999), fifth, sixth, seventh, and eighth grade mathematics textbooks (Çetinkaya, Yenmez, Çelik, & Özpinar, 2018), algebra problems in an online cognitive tutoring mathematics program for middle and high school students (Walkington, Clinton, & Howell, 2015), in *current* New York State Algebra

1 examination problems (Spoth, 2016), and situational problem solving tasks in ninth grade algebra (Leiss, Schukajlow, Blum, Messner, & Pekrun, 2010). New York State mathematics assessments were also analyzed by the number of difficult words that appeared on the assessment by rating the vocabulary of the New York State middle school assessments (Czekanski, 2018) and New York State Algebra 1 assessments into tiers (Spoth, 2016). One strategy to help students do better on exams was to simplify the vocabulary so that students would understand what it means. On the New York State Living Environment exam, Becker (2007) found that students achieved more on exams with simplified vocabulary than exams without the simplification.

Combining readability tests.

Many researchers used a combination of readability tests. For fifth and sixth grade tests, Walker, Zhang, and Surber (2008) found that reading ability affects math performance on an assessment. To measure readability, they used the Flesch-Kincaid, Powers, and Dale-Chall index. King and Burge (2015) used five different readability tests with the mathematics, science, and reading clusters of the PISA assessments. They used the Flesch-Kincaid, New Dale-Chall, New Fog Count, SMOG, and the Spache Revised. They found that the New Dale-Chall test was the most appropriate test for the data, and that the reading level of the test was similar to the 15 year old age level of the students taking the test.

For my research, I have decided to select the Flesch reading ease (1979), the Flesch-Kincaid (1975) reading grade level and ATOS level (2014). All of these have the ability to be calculated online. The ATOS (2014) level calculator also gives the word count, average word length, average sentence length, and average vocabulary level. I decided to select these formulas as they come from different time periods and they have all been endorsed by Core Standards. In

some instances, I will be only analyzing a sentence of text with an item. Readability, in this case, will be used more as a comparison measure across standards and types of exams rather than to draw conclusions about the reading levels of individual items.

Benefits of measuring readability.

Students who are learning English in high school may struggle with the linguistic properties of texts on all subject assessments (van Lier & Walqui, 2012; Moschkovich, 2012) and simplifying the language on standardized assessments such as the NAEP exam can help English language learners (Abedi & Lord, 2001). Target vocabulary and obscure information can confuse students on assessments. In the past, students in South Africa struggled with the reading level of the text on mathematics assessments which in turn affected their academic achievement (Prins & Ulijn, 1998). By studying brain scans, it was apparent that there are definite complex relationships in the brain among languages and mathematics for students from China, with more brain activation when calculations were done in English versus Mandarin Chinese (Wang, Lin, Kuhl, & Hirsch, 2007). In first through third grade mathematics, the wording of an assessment item certainly affected the comprehension of the item more with applied problem solving than with general calculation (Hickendorff, 2013). Caponera, Sestito, and Russo (2016) found that there was a correlation in achievement with reading and mathematics when analyzing the PIRLS and TIMSS scores of grade four Italian students.

Using measures of readability may give more insight into the linguistic complexity of the *old, recent,* and *current* exam items. This would be essential to know as the reading level of the assessment can truly affect the way in which students interpret and comprehend algebraic ideas. In research question 1, these average grade levels, reading ease, and other measures of readability were determined for each exam administration of the *old*, *recent*, and *current* exams.

Also, constructed response and multiple choice item parts were measured with different measures of readability. This may indicate that there is a difference in the reading level of constructed response and multiple choice items.

Contextual research.

Problem solving is a major part of mathematics, and students need to learn how to link problem solving processes to appropriate real world contexts (Ruthven, 1994; Boston, 2012). Contextual simulation prepares students more for the problem solving process (Raub, Shukor, Arshad, & Rosli, 2015; Suvya, Sabandar, & Kusumah, 2013; Yu, Fan, & Lin, 2014). Brenner and colleagues (1997) explored the connection between problems and mental representations. They described how mathematics units should address multiple representations of context, meaningful problem-solving context, problem solving process, and guided discovery.

Researchers have highlighted how cultural and social situations are an important part of understanding mathematical representations (Font, Godino, & Gallardo, 2013; Berg, 2013), but it has not been extensively studied in connection to the types of representations that students use when solving problems (Nistal, Dooren, Clarebout, Elen, & Verschaffel, 2009) Leiss et al. (2010) found that there was a strong positive correlation between the situation model and the level of "real life" modeling. The problem solving contexts may even be associated with political and/or social events (Gjoci, 2017).

The inclusion of context has been an important part of algebraic instruction in the classroom. Students should become owners of the context of the word problem in order to accurately solve word problems in the future (Van Den Heuvel-Panhuizen, 2005). Math chat, or also known as the contextual learning approach (Surya, Sabandar, Kusumah, Darhim, 2013) can also determine problem solving processes as students compare their processes to one another

(Schreiber, 2013). Similarly, when students work collaboratively on problems, students can benefit in learning procedures from others (Care & Griffin, 2014). Learning different problem solving techniques can help students learn how to approach word problems and be able to construct their own methods of solving them (Bonner, 2006) such as self-regulated learning strategies (Puteh & Ibrahim, 2010).

Factors affecting context.

Word problems may not always appropriately illustrate a "real life" situation (Drake, Wake, & Noyes, 2012; Germain-Williams, 2014). Drake et al. (2012) found that the "authentic problem" solving contexts were problems that were truly routine problems with an added human element. Since the context frames the problem, if there does not seem to be a connection between context and the mathematics used to solve, students may not understand how to approach a solution. This could explain why some students are unable to solve word problems: if the context seems to be unrelated to the mathematical concept, it may distract students from solving the problem. Students who do not have the background knowledge about the context will also struggle with solving word problems that are in unfamiliar contexts.

Unfortunately, students who struggle with these unfamiliar contexts are often the students who are the most underserved. Previously based on Matthew 25:39 in the Holy Bible, the Matthew Effect occurs in reading comprehension with struggling students: when students struggle in the primary grades with their education, their gap widens as the years progress onward. The Matthew Effect was first described by social scientist Robert Merton (1968) as a way of describing how certain scientists receive more publicity for their works than others. Walberg and Tsai (1983) noted that young adults' levels of science achievement depended significantly on their educational background, their ethnic and socioeconomic groups, and their

family background. Similarly, when analyzing academic location and low versus high academically rigorous curriculum, Kerckhoff and Glennie (1999) found that academic location along with the level of the academic rigor in the curriculum affected the future level of higher education. Duff, Tomblin, and Catts (2015) also found the Matthew Effect in reading, as there were significant differences in vocabulary growth rates between average and above average readers when students progressed from fourth grade to tenth grade. After analyzing empirical research on the Matthew Effect with Reading, Pfost, Hattie, Dörfler, and Artelt (2014) indicated that this effect occurred with composite reading score and vocabulary on achievement tests.

By measuring the percentage of exam items that include context (marked as a 1) or do not include context (marked as a 0), trends can be determined among *old*, *recent*, and *current* exams. This could greatly affect the teaching of algebra as context can both help and hinder a student's response in an exam item. In research question 2, it was determined how much of a "real world" context is included on *old*, *recent*, and *current* exams and if constructed response or multiple choice items include more "real world" context than one another. In order to describe the "real world" situation in an item, it could possibly require more words or a higher grade level of reading as compared to an item that does not include the "real world" context. In research question 3, it was determined if there was a correlation between the readability level of an item and its presence (or absence) of context.

Research on Complexity of Mathematics

Cognitive complexity refers to the level of thinking required to answer the algebra item which can range from memorization to doing mathematics. Complexity has been measured in many different ways through the years (Holmes, 2012). The framework of the Diagnostic Assessment of Mathematics and Science (DTAMS) was compared to Bloom's original

taxonomy (1956), Skemp's (1976) framework, Bloom's revised taxonomy (Anderson, Krathwohl, & Bloom, 2001), Hiebert and Carpenter's (1992) procedural and conceptual understanding, Webb's (2005) depth of understanding, and Porter's (2002) Cognitive Complexities. All of these frameworks promoted higher order thinking abilities as the students' goal were to independent problem solving.

Classroom tasks in algebra 1 were often divided into "lower order thinking" tasks and "higher order thinking tasks" (Park, 2011). By using Bloom's taxonomy, teachers saw knowledge and comprehension problems as lower order thinking, application problems as lower or higher order thinking, and analysis, synthesis, and evaluation as higher order thinking (Thompson, 2008). Other researchers combined different frameworks into a matrix. Hess, Jones, Carlock, and Walkup (2009) combined Bloom's Taxonomy levels and Webb's Depth-of-Knowledge Levels to create the cognitive rigor matrix.

The Stein, Grover, and Henningsen (1996) cognitive demands of tasks model was derived from mathematics instruction in classrooms. The cognitive demands of tasks were separated into non-mathematical, memorization, procedures without connections, procedures with connections, doing mathematics, and other types of tasks. It was later adapted to a four part complexity framework of memorization, procedures without connections, procedures with connections, and doing mathematics (Stein & Smith, 1998). This framework has been used with videotaping lessons and analyzing teacher's complexity of task (Charalambous, 2008; Major & Mangope, 2012; Polly, 2012; Martin, 2018) and with analyzing textbooks and curriculum (Cai, Moyer, Nie, & Wang, 2009; Delaney, Charalambous, Hsu, & Mesa, 2007; Jones & Tarr, 2007; Pepin & Haggarty, 2001; Ubuz, Erbaş, Çetinkaya, & Özgeldi, 2010). Park (2011) labeled items in algebra 1 textbooks with this framework, but also added a level 5 which included a mathematical

justification of procedures. For my purposes, an item with mathematical justification of procedures would be labeled with written representation along with the level of complexity of the task. A more recent study divided algebra complexity into eight parts, but upon further analysis, they all fit into the original Stein and Smith (1998) framework (Pointon & Sangwin, 2003). Additionally, Lian and Yew (2012) divided the answers to tasks in similar ways as the Stein and Smith (1998), as students understand the most when they can apply it to new, nonroutine novel tasks. The Stein and Smith (1998) framework was selected for this study as it has been used by many different types of educational research in the past. It is complex enough to capture many of the qualities that would make an item seem more difficult, yet simple enough to assign a number to the item and distinguish it as more difficult than another.

In instruction and assessment, mathematical complexity has been measured in several different ways especially in both traditional and reform textbooks. Traditional textbooks may have more opportunities for students to engage in multiple representations (Sood & Jitendra, 2007). Several studies measured representations in reform textbooks such as an increase in complexity in a reform oriented seventh grade curriculum (Cho & Kwon, 2012) as traditional algebra 1 textbooks can have more low-level tasks than reform textbooks (Park, 2011).

Examination programs such as the NAEP, GSCE, and various word problem assignments have divided their problems into categories of complexity. NAEP has rated their exam items by low, moderate, and high levels from 2005 to the present (Walkington et al., 2018) while Fisher-Hoch and Hughes (1996) identified nineteen sources of difficulty from the mathematics GSCE examinations in Great Britain.

Standards based curriculum and their culminating examination programs have emphasized that it is important for students to solve high level, complex tasks. In research

question 4, I assigned scores to each item part from 0, the lowest level of complexity to 3, the highest level of complexity to see if cognitive complexity has relatively increased, decreased, or stayed the same across *old*, *recent*, and *current* exam administrations. Also, I analyzed if there have been changes in cognitive complexity in the *old*, *recent*, and *current* exams' constructed response and multiple choice sections. In research question 5, I established if there were correlations among readability, "real world" context, and complexity.

Research about Mathematical Representations

Bruner (1966) described mathematical representations as being enactive (objects or actions), iconic (pictures), or symbolic (symbols or words). These three types of representations were studied with traditional and reform-based textbooks (Sood & Jitendra, 2007). In the 1980s (Lesh, Landau, & Hamilton 1983; Lesh, Post & Behr, 1987), this research was expanded to five areas of math representation: physical, visual, symbolic, verbal (or also known as written), and contextual representations and was later cited by NCTM (NCTM, 2014).

In mathematics, Dehaene (1992) created a model of numerical processing called the triple code model. The triple code of numbers include the verbal (or in this dissertation, the written) frame, the visual form, and the magnitude representation (Dehaene, 1992; Dehaene & Cohen, 1995; Dehaene & Akhavein, 1995). Recently, Bossé, Adu-Gyamfi, and Cheetham (2011) divided mathematical representations into verbal (or written), symbolic, tabular, and graphical. Recently, Dyrvold (2016) wrote that the four different semiotic resources in mathematical task text were "...natural language, mathematical notation, schematic images, and pictorial images" (p. 29) which again is very similar to the Lesh, Post, and Behr (1987) framework The representation schema that will be used for this dissertation were taken from the Lesh, Post, and Behr (1987)

framework and then refined; the broad categories include *visual*, *symbolic*, and *verbal* with an added *numeric* category.

Representations have been studied along with executive functioning in the brain. Different types of executive functions include updating or the ability to hold onto information during the problem solving process, inhibition or the suppression of information, and shifting or the ability to go between multiple types of representations (Bull & Lee, 2014). These executive functions truly affect the development of all academic skills, but in particular, mathematics. In particular, working memory and updating processes affect students' solving of algebraic word problems (Lee, Ng, & Bull, 2017).

Physical and *contextual* representations were not included in this schema, with *physical* representation being eliminated completely from this research and *contextual* representation being a part of a different schema. *Physical* representations are the hands-on objects, often common items in the classroom or commercial manipulatives that are used to represent mathematical concepts, and therefore, were not considered in this dissertation due to the nature of the data being state assessment items. *Contextual* instances are not representations per se, so this category was coded independently as being present (a contextualized application was evident) or not. Since "real world" context was not considered as a representation in the Bossé, Adu-Gyamfi, and Cheetham (2011) as well as Dyrvold's (2016) studies of the PISA test and 9th grade Swedish exam of algebraic representations, I separated the concept of context into its own schema after representations.

Types of representations.

In my dissertation, the representational schema refers to the types of representations present in the state algebra items. These representation categories are *visual*, *symbolic*, *numeric*,

and *written*. *Visual*, *symbolic*, and *numeric* representations categories have corresponding subcategories that divide the category further. The category of *written* representation does not have any subcategories. Each of the categories and subcategories are described below. A table with examples of a part of an item labeled with its category and subcategory can be found in Appendix D.*Visual* representations involve pictorial or other inscriptions for which shape and spacing have mathematical meanings. They correspond to the Dyrvold's (2016) schematic images and pictorial images and to the visual category in the Lesh, Post, and Behr (1987) framework. Visual tasks were also analyzed in calculus by Bofah and Hannula (2011). For this representation category, I also put five subcategories that are used for *visual* representations: graphs, charts, geometric figures, diagrams, and pictures that represent the mathematical problem. Graphs are plots of single points on one-dimension lines, points and curves in two-dimensional planes that may or may not be functions, or other types of points, curves, and surfaces in three dimensional spaces that may or may not be functions. Charts represent comparative quantities such as pie charts, dot plots, bar charts, and histograms. Geometric figures are polygons, polyhedrons, lines, and/or shapes with geometric relationships. This includes circles, squares, triangular prisms, and parallel lines cut by a transversal. Diagrams represent mathematical relationships and patterns. A common diagram shows how the elements in the function's domain map to elements of a functions' range. Pictures are icons that illustrate situations that would be seen in life that do not in and of themselves represent mathematical concepts or relationships such as hats or spinners.

Symbolic representations are the specific mathematical markings and notations that denote mathematical ideas, mathematical relationships, and structural components of mathematics. They primarily correspond to the Dyrvold's (2016) mathematical notations and to

the symbolic category in the Lesh, Post, and Behr (1987) framework. They also correspond to symbolic tasks as described by Bofah and Hannula (2011) in their analysis of calculus tasks. The subcategories for the symbolic representation category include: algebraic, sets, intervals, probability/statistics, geometric notation, and sequences and series. Algebraic uses variables in expressions, equations, or inequalities that often contain arithmetic operator symbols, exponents, and numerals such as coefficients, variables, and exponents. Sets are a group(s) of elements written in set notation or as a list using brackets. A set may be written so as to include the entire collection of elements, or it may be written in where only the rule for creating the entire set is given. Intervals are a subset of real numbers written with square brackets, parentheses, or equivalent symbols. Probability/Statistics uses letters and specific markings in mathematics that refer to the likelihood of events and/or the analysis of numerical data such as factorials, permutations, combinations, and symbols for measures of central tendency. Geometry is a subcategory that refers to the use of letters and specific markings in mathematics that denote geometric figures and relationships such as triangles, perpendicular and parallel lines, and angle measures.

Numeric representations include quantities and corresponding quantities that are expressed with numbers. This category was also studied by Bofah and Hannula (2011) with calculus problems as using strictly numbers. The subcategories are numerals, points, and tables and matrices. Numerals correspond to the use of numerals and specific letters in mathematics that stand for specific rational, irrational or complex numbers such as e, π, and $2 + 3i$. This subcategory is only assigned when numerals appear alone or when the quantification is the primary focus but units such as 55 miles per hour or six oranges are included. This subcategory is not assigned when numerals appear as part of a passage of text, are used on a number line, are

used in points, or are used with any other notation. Points are corresponding quantities that are typically ordered pairs or triples written in the traditional n-tuple form with parentheses. An example would be the point (3, 4) without its visual representation on the coordinate plane. Tables and matrices are sets of ordered n-tuples that are arranged in rectangular arrays using rows and columns to depict corresponding relationships.

Written representations are a written explanation or description involving words and phrases that represent the mathematical ideas; this category corresponds with Dyrvold's natural language and to the verbal category in the Lesh, Post, and Behr (1987) framework. *Written* representations may include numerals or numbers, but unless they are pertinent to the mathematics at hand, they are not included. For example, a problem such as *Two girls travel x miles north, turn and travel y miles east. How far did they travel?* would not receive this coding since two is not mathematically pertinent information.

Context is not technically a representation in and of itself (Bossé, Adu-Gyamfi, and Cheetham, 2011; Dyrvold, 2016). Instead, it is the "real" world applications that bring in outside settings or scenarios into the problem. It is where the problem and its meaning take place (Halliday, 1978; Morgan, 2006).

Different mathematical representations have been studied separately as have conversions from one representation type to another. In the next section, I first address the cognitive load benefits of using representations in mathematics. Then, I describe the different types of representations and how they have been studied in educational research. Context research is presented after the representations. Finally, I discuss how representations have interacted with one another.

Cognitive load benefits in representations.

Appropriate representations can help with the idea of computational offloading, or reducing the cognitive load needed to solve mathematics (Ainsworth, 2006) which therefore optimizes working memory (de Jong, 2009). Cognitive load can be divided into intrinsic, extraneous, and germane cognitive load (de Jong, 2010). Intrinsic load is the load of the content, extraneous load is the load of the presentation methods, and the germane load is the load of the learning processes within the learner. If the attention of the student is split separately into each mode of representation, it can lead to a high cognitive load with too much repetition of teacher explanations as extraneous load or limiting student explanation as germane load (Renkl, Hilbert, & Schworm, 2009). However, presenting multiple representations together may lead to a lower cognitive load if connections are made explicitly among the multiple representations (Mousavi, Low, & Sweller, 1995) which leads to more successful learning (Horz & Schnotz, 2009). Cognitive load was decreased when similar elements of a subject are learned in succession (Sweller, 1994) and students are trained before the instruction in more difficult topics by focusing on similarities (Renkl et al., 2009). Also, cognitive load decreased with repeated practice as a task becomes more automatic (Mousavi, Low, & Sweller, 1995).

Cognitive load has been further studied in instructional practices by using both worked and modeling examples. Von Gog and Rummel (2010) compared how worked and modeling examples help learners reflect on their own problem solving practices. Students who study worked examples may remember more about problem solving than with solving isomorphic problems (von Gog & Kester, 2012). Effective worked examples include multiple modalities in presentation (Atkinson, Derry, Renkl, & Wortham, 2000). In graphical models, if the user

controls the graphical representation, it can reduce overload of information (Moritz, Meyerhoff, Meyer-Dernbercher, & Schwan, 2018).

Visual representations research.

Visual representations are an essential part of understanding mathematics and can be used to connecting systems of representations together. Visual representations can be broken down into graphs, charts, geometric figures, diagrams, and pictures.

Graphs are an essential part of algebra. Students must have strong graph sense (Friel, Curcio, & Bright, 2001). Teachers must emphasize to students that graphs of functions have strong connections to other representations such as the algebraic formula (Dubinsky & Wilson, 2013) and the concepts of positive and negative real numbers (Leinhardt, Zaslavsky, & Stein, 1990). In particular, understanding a number line correlates with mathematical competence (Schneider et al., 2018). Identification of graph tasks tend to be easier than any analysis of the graph tasks (Olande, 2014). In graphs, shapes can be either static, based on the learned rules of graphing, or emergent, based on the complete understanding of how the graph is formed and how it relates to the axis (Moore & Thompson, 2015; Moore, 2016).

Although the word "chart" is often interchanged with graph and both have x and y axes, in this dissertation, charts are used only for comparison purposes across the independent variable or variables. Students must understand the underlying principles of bar and line charts in order to be successful in graphing tasks which includes top-down and bottom-up processes (Shah & Freedman, 2011). Students must be able to make conclusions from the charts based on the data available. Shah and Hoeffner (2002) emphasized how different graphs and charts are used for different tasks, and how the designer of the chart must keep in mind the density of the image so that the user is able to read it. In the elementary school, there should be an emphasis on chart

literacy as it can be used in mathematics, science, and history classes in the high school (Doig & Groves, 1999).

Geometric figures can be difficult for students to identify and use. The Van Hiele (1984) levels of geometric understanding have been used to assess a student's knowledge of geometry, as previous researchers have found that elementary students have difficulty even identifying basic geometric figures or a level 0 visualization task (Carroll, 1998). The recognition of basic geometric figures is essential in order to be able analyze these figures for commonalities (Crowley, 1987). By having students manipulate geometric figures in the Van Hiele (1984) method, students have improved in the geometric reasoning abilities (Guven, 2012; Mistretta, 2000). Teachers must be careful that they do not establish incorrect assumptions such as a formula (such as the Pythagorean Theorem) should always be applied to right triangles (Panaoura & Gagatsis, 2009).

Diagrams are the figures and images that involve mathematical concepts but are not tables or graphs. In order to analyze a visual image, there are certain properties of the page and specialized marks that convey meanings (Tversky, 2011). The individual must comprehend and apply these rules to understand the image. Different types of diagrams have been used to illustrate algebraic process such as using Picture Algebra to understand above grade level mathematics problems (Koedinger & Terao, 2002). Seeing visual patterns in diagrams can help students have more of a flexibility in understanding mathematics (Barbosa & Vale, 2015; Barbosa, Vale, & Palhares, 2009; NCTM, 2000). Diagrams can help students problem-solve more successfully by bridging their visual representation with symbolic representations (Steele & Johanning, 2004).

Pictures are defined as representing non-mathematical concepts. While diagrams that include mathematical concepts assist students in their comprehension of the item, decorative images that are truly not connected to the deeper mathematical content do not assist students in comprehension (Dewolf, Van Dooren, Hermens, Frouke, & Verschaffel, 2015; Elia & Philippou, 2004). Students will distinguish between important and unimportant details in images by understanding what part of the image would be essential in terms of solving a problem (Wang & Adesope, 2016).

Technology has helped with understanding visual representations. For example, students have used interactive diagrams on computers that have led to better writing of mathematical ideas (Ottmar, Landy, & Goldstone, 2012; Yerushalmy, 2005) and better learning performance (Yung & Paas, 2015). However, most of the research that includes visual representations does not focus solely on this type of representation; rather, it is more common for visual representation to be connected to written and/or other types of representation in research studies.

Symbols representations research.

Symbols are fundamental to mathematical communication and affect mathematical performance in secondary and higher education settings (Bardini & Peirce, 2015). Students can often guess at the meanings of symbols based on their knowledge of other symbol systems which may lead to errors (MacGregor & Stacey, 1997).

Algebraic symbols have been studied in many different research studies. Using a trial and error process may lead to an ability to understand the algebraic symbols of mathematics (Radford, 2006). Students connected symbol systems together better when two sides of the equations have clear visual relationships such as with the distributive property (Kirshner & Awtry, 2004). When symbols are combined with text in the word problem solving process, the

spatial proximity of the words and overall alignment of symbols and text can affect a students' problem solving abilities (Landy, Brookes, & Smout, 2014; Landy & Goldstone, 2007). When students correctly write symbols in equations, it does not mean that they will naturally use this equation to solve a problem (Swafford & Langrall, 2000). Instead, there may not be strong connections between the symbols and the overall context of the problem. Set and interval notation has been studied in relation to the algebra classroom (Vamvakoussi & Vosniadou, 2010), but the complexity of that notation has not been studied by itself.

On the *current* exam, effective function notation is integral to demonstrating understanding of functions. On the *current* August 2015 exam item #25 in Figure 1, students needed to write a function based on a given table of values.

25 Each day Toni records the height of a plant for her science lab. Her data are shown in the table below.

Day (n)	1	2	3	4	5
Height (cm)	3.0	4.5	6.0	7.5	9.0

The plant continues to grow at a constant daily rate. Write an equation to represent $h(n)$, the height of the plant on the nth day.

Figure 1. *Current* Exam, August 2015, #25Reprinted from Regents High School Examination Algebra I Model Response Kit, by New York State Board of Regents, August 12, 2015, retrieved from https://www.nysedregents.org/algebraone/815/algone82015-mrs.pdf. Copyright 2015 by The University of the State of New York.

Students would receive the full credit of 2 points for $h(n) = 1.5n + 5$ as this is effective notation based on the day or n and height or h(n). They would only receive 1 point if the student wrote $h(n) = 1.5x + 5$ as the function is not directly referring to n, the number of days.

Probability and statistics notations have been difficult for students in terms of understanding the symbols and the ways in which symbols may be referred to. Students need to develop a statistical literacy by knowing the symbols before they are able to access the content of the class (Garfield & Ben-Zvi, 2008). Students make significant errors in connecting symbols in statistics with concepts (Garfield & Ben-Zvi, 2007; Kim, Fukawa-Connelly, & Cook, 2012) such as identifying the mean notation as a number, not as an operation (Mevarech, 1983). From class to class, they may struggle with what to call the mean symbol as some people may call it *mu-hat* and others may call it *x-bar* (Rumsey, 2002). Connecting these symbols to real life data can improve learning (Bradstreet, 1996) along with using computer tools in understanding statistical concepts because students can see how data affects the model of variability (Lehrer, Kim, & Schauble, 2007).

The final two subcategories of geometry and sequence/series did not have research dedicated to their symbols. Geometric notation has not been studied outside of its usage with its corresponding geometric figures. Sequences and series are a relatively new topic to state algebra exams, and therefore, have not been studied in the middle school, high school, or early college settings.

Numeric representations research.

Seeger (2008) emphasized that the core number system is abstract and numbers can apply to a variety of objects. Number systems in various cultures can be added and subtracted in order to compare the numbers or combine the numbers. Students have difficulties with number types

(Fischbein, Jehiam, & Cohen, 1995) and with using decimals and fractions (Calhoon, Emerson, Flores, & Houchins, 2007; Hiebert & Wearne, 1983; Tian & Siegler, 2017) especially where numbers are expressed in intervals (Vamvakoussi & Vosniadou, 2010). Godino, Font, Wilhelmi, and Lurduy (2011) studied how six year olds worked with place value. They found that the students were not able to conceptually differentiate between the tens place and the units place. This led the researchers to see the systems of rules that exist within place and how first grades worked with the text of the problems.

Point notation was either studied with visual graphs or with numeric tables. Tables establish relationships by organizing data into rows and columns. Students may have particular difficulty with reading tables as those skills may not be directly taught in classrooms. This is why researchers have encouraged the teaching of functional relationships by tables in primary grades (Blanton & Kaput, 2005a; Greenes, Chang & Ben-Chaim, 2007). Primary school students do not always make connections between columns in tables (Schielmann, Carraher, & Brizuela, 2001), but they may use them successfully to understand additive relationships (Brizuela & Roth, 2002). In middle school textbooks, ordered pairs are often manipulated, appreciated, or calculated and include another symbolic or a visual representation such as a Cartesian plane (Mesa, 2004). Sharma (2006) found that high school students particularly struggled with interpreting tables in statistics. Tables in statistics do not typically follow the same rules as algebraic tables. Additionally, there is often an absence of teaching the rules of set theory in the high school (Jones, 2006) and students can struggle with the idea of an infinite set (Jones, 2006; Hazzan & Zazkis, 2005).

Written representations research.

Wilson (2011) recommended that multimodal representations should be an integral part of reading and writing instruction in the content areas. Compared to other subject areas, writing in mathematics often involves a set of rules and procedures that lead to a single correct solution. The students in mathematics would need to understand the rules of mathematics and literacy and be able to work with them together in instruction and assessments (Aiken, 1972; Draper & Seibert, 2004; Wilson, 2011) and understand the meanings of different keywords (Phonapichat & Wongwanich, 2013). During class instruction, students would need to be able to take appropriate notes to capture verbal discussions, as effective note-taking is an important part of the mathematics learning process (Dündar, 2016). For instance, formulas can be considered like stories as students read through problems and then determine an appropriate procedure based on the wording of the problem (Radford, 2002).

Mathematics includes cohesive devices such as vocabulary words, connectors, and words that refer to how to read the text (Morgan, 1996). Students must be able to understand linguistic elements in the mathematical text and connect to its global meaning. Overall, by writing about mathematics, mathematical writing helps students to be a part of the greater mathematics community (Bicer, Perihan, & Lee, 2018; Burton & Morgan, 2000; Clarke, Waywood, & Stephens, 1993).

Although teaching writing in the secondary mathematics classroom can take time, teachers noted there was some effect on mathematics achievement by using writing in the mathematics classroom (Teuscher, Kulinna, & Crooker, 2015; Powell, Hebert, Cohen, Casa, & Firmender, 2017). There was even more of a pronounced effect when the process of writing was intertwined with the process of doing mathematics (Kuzle, 2013) which produced metacognitive

behaviors essential to the understanding of the mathematical concept (Pugalee, 2001). By assessing writing in the mathematics classroom, Steele (2005) found that it improved problem solving practices and the ability to reason through a problem.

In assessment, students are responsible for reading and interpreting written information. After studying various methods of analyzing reading and math ability, Österholm and Bergqvist (2012) recommended to do a principal component analysis. Using student results, teachers found the demand on reading ability for both mathematics and reading items. Morgan and Sfard (2016) studied the language complexity of a mathematics assessment. They considered the average number of logical connectives per sentences, length of sentence, and the length of the compound nominal group. They also created an idea called the grain size of the problem which is the number of steps a student would need to do to solve the problem. The number of logical connectives decreased from 1980 to 2011, the number of words per sentence decreased when comparing a test from 1987 to a test from 2011, and there were fewer compound nominal groups from 1957 to 2011. Additionally, mathematics vocabulary use over the years was similar, but there more complex vocabulary was used on more recent exams compared to the pre-reform curriculum of 1980 (Morgan & Tang, 2016). Additionally, teaching particular vocabulary strategies can lead to increased knowledge of mathematical concepts (Monroe & Pendergrass, 1997; Freitag, 1997; Bossé & Faulconer, 2008). However, when measuring written information, Bergqvist, Dyrvold, and Osterholm (2012) found that there are connections between vocabulary and task difficulty, but it may be because of the number of words in a task rather than the different types of words.

In research question 6, I counted the number of representations present in each exam item part. Also, after labeling all item parts with the sixteen types of representations, I determined

which representations were prevalent on the *old*, *recent*, and *current* exam administrations, as well as if these representations were more prevalent on the constructed response or multiple choice item parts.

Relationships among representations.

Three types of relationships occur among multiple representations in the state algebra exams. *Representation treatments* involves staying in the same representation but moving through different versions of that representation. For example, when two linear equations are compared to see if they are the same equation, this would be a treatment. *Representation conversions* occur when there is a movement from one representation to another representation. This could be when the learner needs to connect a graph in the stem with a symbolic algebraic inequality in the response. Finally, *representation synergy* requires the use of multiple representations in order to obtain a correct response. This could occur when there are many different representations such finding the largest maximum of a parabola in a multiple choice item stem, and then having to choose the largest maximum among different representations of it: two graphs of parabolas, a table of a parabola, and a symbolic equation of a parabola. Visual and written representations were often studied together, so they are presented below. Researchers have also studied the multiple connections that can exist in mathematics, and they are also described below.

Visual and written representations.

The connections between written and visual representations were considered by researchers. When written and visual representations are combined, students may understand more than with one representation alone, and multiple visual representations combined with text may help students more than one single visual representation with text (Rau, 2016). A visual

representation can establish relationships between the symbolic form and written word form of the problem (Banerjee & Subramaniam, 2012; Swee & Lee, 2009). In the instructional process, teachers rely on spoken metaphors that relate to the orientation of the image and connections among different parts of the image and thought process (Font, Bolite, & Acevedo, 2010).

The combination of visuals and text more fully understand the topic area. Arcari (2003) mentioned how visual displays help mathematics students understand the mathematical story along with cause and effect relationships. Quinnell (2004) similarly highlighted the importance of using written and visual word association diagrams in teaching in order to help students read visual images. Also, she found that close reading was important as students would be able to understand the entirety of the problem. In the classroom, students can learn more from visual and the written word at the same time. For instance, the action of handwriting on the blackboard combined with lecture assists students in doing and thinking about mathematics (Greiffenhagen, 2014; MacKenzie & Barany, 2014).

Connections of multiple representations.

In the triple code model, Dehaene and Cohen (1995) stated that numbers include an internal transcoding, or the ability to transform an input code into an output code. Various subject areas have connected different types of representations including physics where performance on an assessment was compared to the type representation used in question including diagram, written word, symbolic, and graphical problems (Meltzer, 2005). In algebra, Panasuk (2010) created a three-phase ranking framework for multiple representations of algebra. She measured how students used written word, visual, and symbolic representations, and then ranked the students based on their understanding of the topic. Students who used multiple representations (symbols, written word, image, and table) were able to answer problems better

than those who depended on one representation alone (Nizaruddin & Murrianto, 2017). Students who work together on multiple representation problems interact more with one another than with single representation problems (Pape & Tchoschanov, 2001). The MERLO or Meaning Equivalence Reusable Learning Objects approach allows students to look at a subject from multiple representations. Teachers develop a target statement and a boundary of meaning where the author knows how the answers to the problem share the meaning equivalence (Arzarello, et al., 2015). There are two major scores on the MERLO assessment: a recognition score and a production score. The recognition score measures how well the students determined the mismatch among representations while the production score measures how well the learner described why the mismatch occurred (Etkind, Kenett, & Shafrir, 2010).

Research findings have suggested that there is a correct ordering of the presentation of representations: instance-based representations or visuals such as tables and point-wise graphs should be presented before holistic representations such as symbolic equations and written word expressions (Nathan, Stephens, Masarik, Alibali, & Koedinger, 2002). Bossé, Adu-Gyamfi, & Cheetham (2011) mentioned how teachers felt that translations between representations can range in difficulty. Easier translations included verbal to table, symbolic to graph, symbolic to table, table to graph, and graph to table. More difficult translations included written word to symbol, table to symbol, and graph to symbol while the most difficult were symbolic to written word, table to verbal, and graph to written word.

Benefits of multiple representations.

Stylianides and Stylianides (2007) found that many authors highlighted the ability of mathematical transfer, or the ability to take previously learned mathematical concepts and apply them to new, novel problem situations (Sweller, 1994). Students who have integrated worked

examples with multiple representations were able to perform better than students with split-source examples (Mwangi & Sweller, 1998). When written word, visual, and symbolic elements are presented at once, there is more likelihood for students to understand the relationships between these representations (Moreno & Durán, 2004). Overall, students who can move fluidly between different representations perform better on assessments of mathematical topics as compared to students who are unable to see the connections between different representations of mathematics (Lepak, Wernet, & Ayieko, 2018; Daryaee, Shahvarani, Tehranian, Lotfi, & Rostamy-Malkhalifeh, 2018).

When preparing students for the rigors of algebra, multiple representations help students. In preparatory courses like pre-algebra, multiple representations can lead to a stronger conceptualization of symbolic notation (Moseley & Brenner, 1997). Students who traditionally struggle with mathematics benefit from seeing the concept in its multiple representations. For example, students with disabilities benefit from both traditional and reform-based practices that highlight the use of multiple representations in mathematics (Hudson, Miller, & Butler, 2006).

Factors affecting multiple representations.

Previous studies highlight how students struggle connecting representations. Students may struggle in translating representations from one form to another which greatly affects their learning of mathematics (Lesh, Post, & Behr, 1987). There is often a lack of mediating elements, and students may need to be more trained in linking representations (Amit & Fried, 2005; Panasuk & Beyranevand). Students who used a "trial and error" process in solving problems may also have difficulty in connecting multiple representations of mathematics (Rossi Becker & Rivera, 2005). Other students may transfer to a different representation when the initial representation does not explain their solution called a representational disfluency (Bieda &

Nathan, 2009). Learners needed significant evidence to connect graphic and algebraic representations (Amit & Fried, 2005; Maffia & Mariotti, 2015). In terms of students with mathematical disabilities, they may struggle with all types of number codes in assessments (Moll, Göbel, & Snowling, 2015).

Students may have different mental models when solving mathematical problems. The mental schema from one mathematics problem may transfer over to another problem (Stacey & MacGregor, 1995) which can cause errors in conceptual understanding. Although students had many extensive years in using algebra in solving problems, few students communicated with algebraic representations (Neria & Amit, 2004).

Although it is important to transfer between various representations, teachers can struggle with this cognitively difficult task (Bal, 2014). Teachers may need to learn how different mathematical representations work together in order to teach their students about these connections (Vale & Cabrita, 2008). Teachers struggled with the link between graphic to numerical representations (Cunningham, 2005; Kiliç, 2017) and the link from graphs and algebraic representations to oral representations (Bal, 2015). Visual representations of algebraic concepts may not always be correctly connected with analytic and graphic representations (Hitt, 1998; van den Bosch, Espin, Chung, & Saab, 2017).

For my purposes, in research question 7, I used the most prominent representations found in research question 6 and analyzed their interactions in stem and response for *old*, *recent*, and *current* exam administration as well as with constructed response and multiple choice item parts. In research question 8, after labeling item parts with *representational treatment*, *representational conversion*, or *representation synergy*, I found out if different representations occurred more

frequently on *old*, *recent*, or *current* exams, and if these representations occurred more on constructed response or multiple choice item parts.

Limitations of Research: Standards, Mathematical Representations, Context, Readability, and Complexity

There has not been a true holistic approach to studying mathematics exams. All of these areas have been studied separately before. There has been research on standards, but standards and topic areas have not been connected to types and/or relationships of representations, context, readability, or complexity. When exam items are studied by standards or topic areas, teachers may focus just on the topic without thinking about its inherent connections to other topic areas. If a student is really struggling with the concept of plotting points on a graph, this does not just affect a linear functions unit. Instead, it affects many different units such as quadratics, polynomials, and statistics units. The student may need to see a variety of functions graphed side by side in order to organize and understand this process. Having the items separated by representation would make this type of instruction easier for teachers.

Also, context and readability have been studied together, but the presence of context can really affect different representations both in positive and negative ways. Having a real world connection to an equation could make writing the equation easier if the student has a strong knowledge of the context. For example, if adults and students are charged different amounts at the movie theater, it may be easier to understand how to write that equation than if it was solely in abstract form. Conversely, students who do not have that experience may do better on an abstract form of the equation versus the real world situation. When context and complexity are studied together, I can see if real world word problems are typically more complex than non-real world problems.

There may be a misconception that more representations equate to more complexity of item. This may not be the case all of the time. Instead, more representations of the same problem may decrease cognitive load and build more meaningful connections among the representations. Connecting these representations together may lead to more flexibility in problem solving later on.

Through this research, I was able to see if the different exams and their corresponding constructed response and multiple choice item parts have higher or lower readability levels, "real world" context, higher or lower complexity levels, certain types of representations, certain stem-response relationships, and different relationships between representations. Instead of reteaching a unit, I may be more focused on teaching strategies for representations, reading word problems, and using items of lower complexity first and then using increasingly more difficult problems to build up to the highest level of complexity.

Theoretical Framework

The study of semiotics was founded by Pierce who stated that there were three semiotic elements: the sign, the object, and the interpretant. The *object* may be dynamic or static. As the learner finds out more about the object, it is a dynamic object as it changes with the learner's perspective. Finally, the real object stays the same as it has been agreed on by many different people and into various sign-vehicles (Sáenz-Ludlow & Kadunz, 2016), which is similar to how a faithful representation that can be manipulated but always maintains its own identity as an icon (Carter, 2018). The *sign* is created because of the nature of the object to represent the object or to communicate with others about the object. Signs can be used for representation, mediation, and a dialetical materialist view where signs are neither representational or used for mediation (Radford, 2014; Bussi & Marriotti, 2008). Transformations of signs can occur, but the overall

meaning must stay the same (Ernest, 2008a). Finally, the *interpretant* is the person's view of the object. The collateral experience of the interpreter of the object affects the view of its context (Ernest, 2008a; Morgan, 2006; Silva & Almeida, 2018). In the Piercean approach, signs of an object are perceived (firstness), then reacted to (secondness), and then reflected upon by the interpretant (thirdness). For example, consider the quantity of seven. If we consider that particular quantity to be the object, there are many signs that are used for the object. Some of them include seven dots, the numeral 7, the word "seven" and the word "siete." Now think about how the interpretant comes into play. The last two signs for the quantity are viewed are impacted greatly by whether a person speaks English, Spanish or neither. This is similar to Dehaene's triple code model (1992) as the verbal (or in this dissertation, the written) frame would be the word seven, the visual form would be the numeral 7 or seven dots, and the magnitude representation may be its place on the number line in relation to zero.

There are systems of signs. Vickers, Faith, and Rossiter (2013) recommended that the user looks at the "membership" of each sign system to know how to use the sign. Lemke (2003) described how every sign has a presentational, orientational, and organizational meaning. *Presentational* meanings are the relationships between the participants and their processes. *Orientational* meanings are the connection between the presentational content and the community of potential audiences. Finally, *organizational* meanings are the connections between the entire context and this part of the context or between different parts of the content (Lemke, 2003). Thus, by understanding Piercean semiotics, it can be used to encourage people to produce multiple tasks in many different domains (Smith, 1997). Figure 2 illustrates the different types of sign meanings.

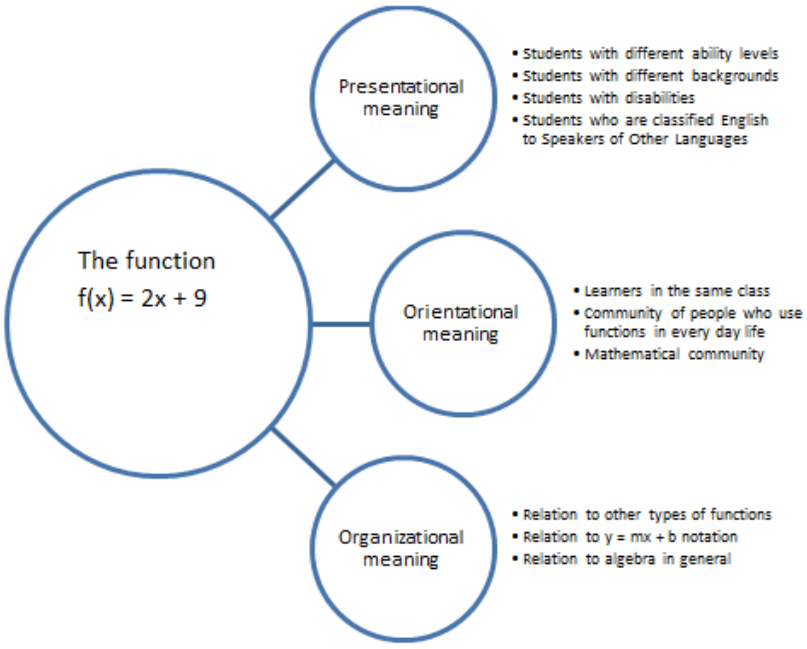


Figure 2. The sign meanings of the function f(x) = 2x + 9.

In mathematics, Tarasenkova (2013) described the sign-symbolic means in mathematics, and divided the means into verbal and nonverbal categories. The verbal means was the wording, terminology, numbers, and symbols of mathematics while the nonverbal means was the visual diagrams, tables, and images relating to mathematics. Also, there is sign and symbol activity that occurs in mathematics such as substitution, encoding, decoding, recoding, schematization, and modeling.

Objects, subjects, and the act of symbolizing can be interconnected (Radford, 2000). Signs can be similar to representations in the representational approach as signs are similar, cause one another, or can be referenced by each other (Iori, 2017). Multiple systems of representation can work together in order to create meaning (Duval, 2006) by complementing, constraining, or constructing one another (Ainsworth, 1999; Ainsworth, 2006).

Conversions occur when registers change, but the objects do not change (Duval, 2006). For example, seeing "three added to seven times some amount" in print and writing "7x + 3"

would be a conversion. Relating representation depends on the familiarity with the representation, the familiarity with the mathematics domain, age, and other individual differences (Ainsworth, 2006). Teachers may not always connect equivalent representations because they have difficulty in using the technical terms of mathematics and/or they have difficulty understanding how non-mathematics can affect mathematics education (Iori, 2018).

Different systems of representations have been discussed as the idea of semiotic circuits that bring together language, image, and symbolism (O'Halloran, 2010). Within a system of representation, there can be a particular metalanguage where all interpretants need to have the same level of background knowledge in order to access the material as indicated with both readability (RQ1) and "real world" context (RQ2) (Ernest, 2008b; Radford, Bardini, & Sabena, 2005). These connections of these systems allow learners to see a resource in many different ways (Morgan & Alshwaikh, 2009).

The semiotic theory of model was described as where an object called a model is developed from an original object sign to establish a relationship between the objects (Kralemann & Lattmann, 2013). One idea related to the synergy between different registers being used for the same object is the semiotic bundle. While a semiotic set is a set of signs, modes, and relationships, a semiotic bundle is the collection of these sets and it describes the relationships between the sets in the bundle (Arzarello, 2006; Maffia & Sabena, 2015). This type of work has its beginnings in neuroscience as these multiple modes come together to express broader meanings (Presmeg, Radford, Roth, & Kadunz, 2016). Semiotic bundles have been traditionally expressed the connections between gestures, symbols, and written language (Arzarello, 2006; Maffia & Sabena, 2015; Sabena, 2018). Amieda & Silva (2018) analyzed semiotic bundles in calculus textbooks where they stated that the symbolic equation of the

derivative, understanding the idea of slope of the tangent line to the graph, and the value as a rate of change all would be a part of a semiotic bundle. When students are asked to interpret a function for its real world meaning, it requires multiple sets of information form the world around them and the mathematics world. For example, on the *current* June 2015 exam, the student would need to understand the parts of the function as it pertains to the proposed real world situation. At the same time, they would need to know how the linear function may appear on a graph as well as understand what each symbol means in the context of algebra. Figure 3 shows an example of the semiotic bundle.

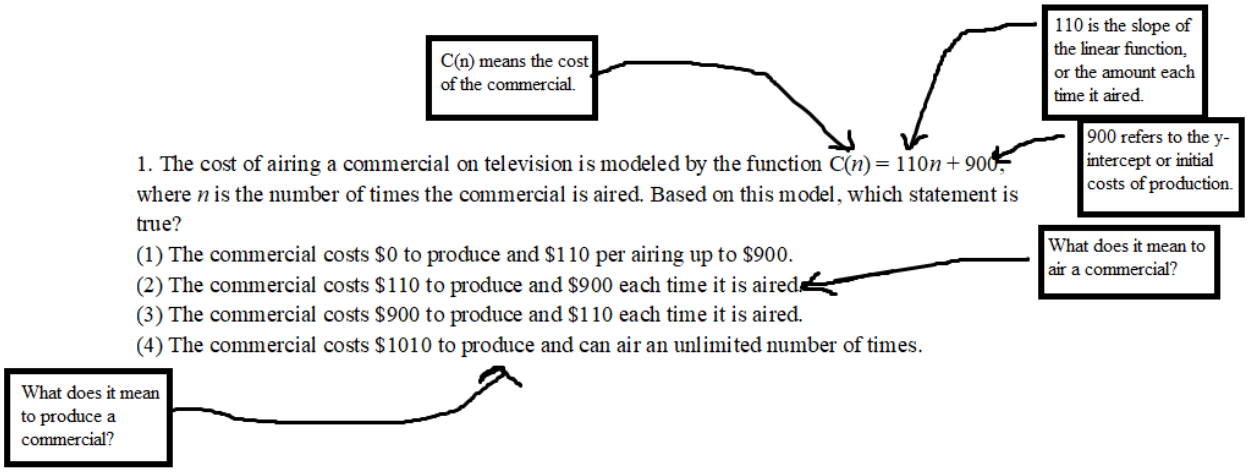


Figure 3. Example of the semiotic bundle on the *current* June 2015 exam, item 1.

A multimodal semiotic chain occurs when there are different representations and there are subsequent links leading to each representation (Maffia & Sabena, 2015; McGee & Moore-Russo, 2014; Radford, Bardini, & Sabena, 2005; Presmeg, 2006). Ceuppens, Deprez, Dehaene, and De Cock (2018) clearly illustrated how semiotic chains worked in physics and algebra function problems, and this is illustrated in figure 4.

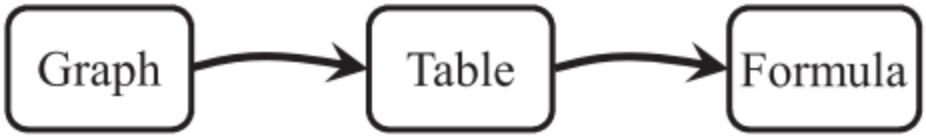


Figure 4. Example of the semiotic chain.

From "Design and validation of a test for representational fluency of 9th grade students in physics and mathematics: The case of linear functions" by S. Ceuppens, J. Deprez, W. Dehaene, & M. De Cock, 2018, *Physical Review Physics Education Research, 14*, p. 15. Copyright 2018 by the American Physical Society.

An example of a semiotic chain could be graphing an absolute value equation such as one part of item 25 on the *current* June 2015 exam. First, the student reads the written word description of the problem to understand that they need to graph this function. Then, the student concentrates on the symbolic algebra equation, recalling the absolute value symbol. From this equation, the student selects applicable x values. If the student understands the idea of transformations of functions, it will be easier to find x values that show the entirety of the absolute value function. With each x value, the student has to find the y value. Then, the x and y value are considered together in point notation. Then, the student may plot all of the points using the rules of the coordinate plane. This process will continue until the entire absolute value function is plotted. Figure 5 shows this example of the semiotic chain. All of these models highlight the need to establish connections between different sets of modes.

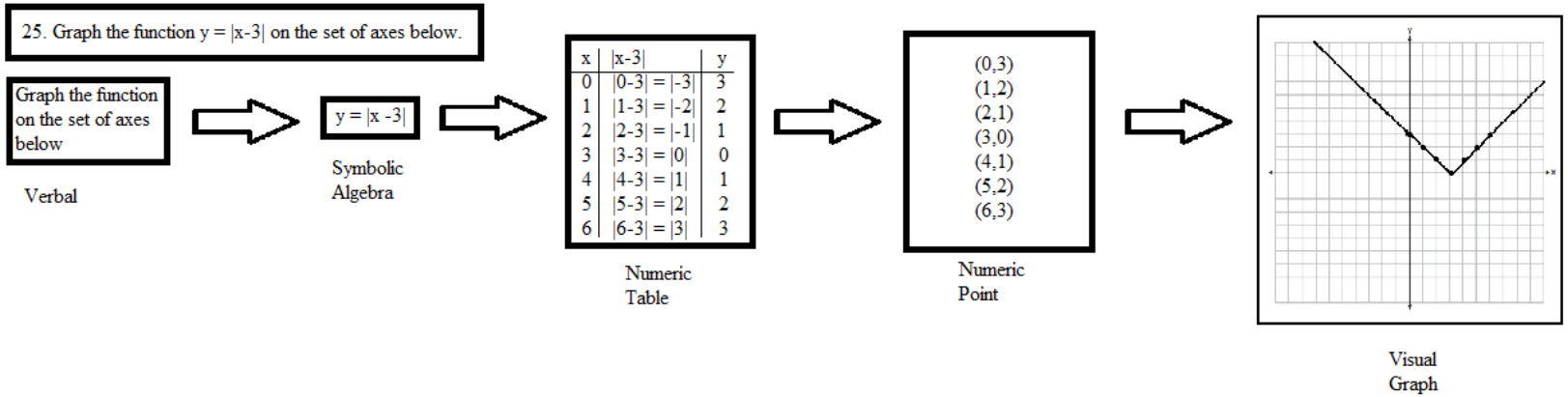


Figure 5. Example of the semiotic chain on the *Current* June 2015 exam, item 24.

The process of mathematical thinking.

Algebraic concepts can be incorporated into many different areas of mathematics later such as in geometry, statistics, and calculus classes (Ferrini-Mundy, Floden, & McCrory, 2005). There is a zone of emergence of algebraic thinking where factual, contextual, and standard algebraic thinking should take place (Radford, 2010). *Factual* algebraic thinking may include understanding generalized algorithms and interpreting the relationships between symbols and visual representations. *Contextual* algebraic thinking was more metaphorical as students would understand the formula. Finally, *standard* algebraic thinking was when students would understand that formulas are related among other signs. Mathematical theory has to be recontexualized into the school mathematics topic (Ernest, 2006) as mathematical practice can be divided into institutional and personal meanings (Godino, Batanero, & Font, 2007).

Representations are thoughtfully chosen in mathematics. In one way, a representation in mathematics is universal, but at the same time, its meaning can be different for every person (Otte, 2011). Heritage and Niemi (2006) developed a framework for teaching and assessing representations in the mathematics classroom. First, the teacher would analyze the context and make sure that the students understand the mathematics content and language. Then, the teacher would plan the instruction using the representations in the classroom. Next, there is a cycle of content development and creation of representations by the students. Afterwards, the teacher would interpret the representations of the students and determine the next step in the instructional process. Similarly, Ernest (2006) described how the semiotic systems affect learners, as they have to be able to identify the task, accept it, and then be able to use their own body of knowledge to appropriately represent the answer.

Relating semiotics to the assessment process.

On the *old, recent,* and *current* exams, students were required to answer both constructed response and multiple choice item parts. Within each of these item parts, there were parts that students need to read and understand. In the Piercean approach, these words were the signs that described the underlying mathematical objects, and the learner needed to interpret these signs correctly in order to receive full credit in a response that demonstrates the appropriate algebraic thinking. By using Dehaene's triple code model, in one item, a verbal description would be used to construct a visual graph and the axes of this graph would indicate the magnitude in the stem.

Readability or the use of words as signs can be on a spectrum of understanding for a learner from easily understood to completely confusing. Although more difficult vocabulary and descriptions can be very specific and pinpoint the direction of an exam item for members of the mathematics community, it can affect the reading of the item if the learner does not understand the conceptual basis of the word and how the word is organized along with other related signs. When items are written with a contextual basis, it can be confusing for students who have little knowledge of the context. Items with context may also hinder students who have difficulties in reading or students who have difficulty with focusing on an item with more words based on their disability. When items are more cognitively complex and require a mastery of the mathematics, this may require more in-depth knowledge of each of the representations on the exams and how they can be used with one another. Each item could be a semiotic bundle of signs that are required to be used together in order to select the correct response.

Representations can either help or hinder the learner. Students who are fluent at the procedure of graphing a line may also be able to graph a parabola correctly as well. This may help the student be able to find the vertex of a parabola by using the maximum found in the

graph, but it does not always mean that the student understand how the vertex would be found in other representations like a table of values or as a symbolic algebraic function.

There were always different signs being used within one item, and each of these signs could be interpreted differently based on the experiences of the learner. In all item parts, the learner had to understand the process of how to transform one or more signs in the stem to sign or multiple signs in the response. For representational treatments, the learner would stay within that representation. For representation conversions, the movement would be from the one representation of a mathematical object (for example, a visual-graph) to another representation of the same mathematical object (for example, the corresponding symbolic-algebraic equation).

Summary

This chapter reviewed the history of the standards movement and in particular, the New York State algebra exams. It then covered a review of literature of representations, context, relationships among representations, readability, and complexity. This type of holistic research may assist teachers in preparing students for state algebra exams in the future. Using the theory of semiotics, the semiotic bundle and chain were discussed in relation to mathematics teaching.

Chapter 3 – Methods

This chapter contains four major sections outlining how algebra exam items were collected and analyzed. The first section described the data set of the New York state algebra standards and exam items. The second section described the research instruments used for coding the algebra exam items. The third section outlined the research methodology and rationale for selecting these research methods. This section is divided into coding by representation, context, relationships of representations, readability, and cognitive complexity. Finally, the fourth section described how the codes will be collected and how they were analyzed to find trends in algebra state exam items.

Study Design: Data Set of New York State Regents Standards and Examinations

Since 1866, there have been exams in New York State that assess students' mathematical abilities. Currently, in New York State, all students need to pass a state mandated exam for a high school diploma, and the majority of students will attempt to pass the Algebra exam. For each exam, there is a corresponding set of standards that prepare students for the academic rigor of the exam. The state mandated exams in New York State are called Regents examinations, and currently, they are given in mathematics, science, social studies, and English. There are two levels of New York state diplomas: a Regents diploma and a Regents diploma with advanced designation. Students have been required to pass one exam in high school mathematics for the Regents diploma and all of the exams in high school mathematics for the Regents diploma with advanced designation. Students traditionally attempted to pass the lowest level Regents exam that focused on algebra.

Since 2002, there have been three versions of state algebra exams: Math A (referred to as the *old* standards), Integrated Algebra (referred to as the *recent* standards), and Algebra I

Common Core (referred to as the *current* standards). For the *old* standards, the corresponding exam was most commonly taken in the middle of the 10th grade year, but it could also be taken in the middle of the 9th grade year for advanced students or middle of 11th grade year for remedial mathematics students. For the *recent* and *current* standards, their corresponding exams were most commonly taken at the end of the ninth grade year, but it could also be taken at the end of an eighth grade year for advanced mathematics students or at the end of tenth or eleventh grade year for students who struggle with mathematics. Other exams offered in the time period required knowledge of the state algebra curriculum, or did not lead to a Regents diploma. Since all students who graduated with a Regents diploma had to take a state algebra exam, it was appropriate to analyze only state algebra exam items.

In the *old*, *recent*, and *current* standards systems, teachers use the standards map as a guide in order to figure out how to teach their students. In order to cover the entire curriculum, it is important to make sure that the standards are being met. However, standards only mention broad areas of study, and do not necessarily mention the types of mathematical representations and transfers of representations, level of readability, or the level of complexity. Without this knowledge, teachers may have difficulty creating their own curriculum and assessments that would meet the expectations of the set of standards. If teachers do not know how the topic may be represented on an exam, then they may not adequately prepare their students for this concept. Students and teachers may also rely solely on a single type of representation such as graphing a line and spend the majority of time teaching the procedure rather than focusing on the conceptual basis of *why* they are creating the line, and showing how that connects to tables, symbolic equations, and real world problems if they have not adopted reform based teaching practices and have instead continued teaching in a traditional method. Students with disabilities may have

difficulty reading and interpreting some types of representations which may span across topic areas. Students who are English language learners may struggle with confusing language structures. To promote algebra for all, teachers need to have an awareness of representations, context, readability, and complexity of items in addition to the amount of topic and standard coverage on the exams.

The total number of items on NYS Regents exams from June 2002 to August 2018 was 2,176. Items on the *old* Mathematics A examination from June 2002 to January 2009 (n = 764) and items on the *recent* Integrated Algebra from June 2008 to June 2015 examinations (n = 858) were either multiple-choice items worth two points or constructed-response items worth two, three, or four points. The *current* Algebra I Common Core examination administered from June 2014 to June 2018 (n = 554) eliminated the three-point items and created a six-point item. Algebra I Common Core examinations consist of multiple-choice items worth two points or constructed-response items worth two, four, or six points (see Table 1 in the introduction).

Each of the *old* Math A exams from June 2002, August 2002, January 2003, and June 2003 consisted of 35 items. However, each of the *old* Math A exams from January 2004 to June 2009, had 39 items. For all of the *recent* Integrated Algebra examinations, there were 39 items. Finally, the *current* Algebra I Common Core examinations have consisted of 37 items. Table 4 further outlined the item count for the exams.

Table 4

Number of Item Counts for Old, Recent, *and* Current *Regents Examinations*

Item Type* and Value	*Old* June 2002 - June 2003	*Old* January 2004 - June 2009	*Recent*	*Current*
MC (2 points)	20	30	30	24
CR (2 points)	5	5	3	8
CR (3 points)	5	2	3	0
CR (4 points)	5	2	3	4
CR (5 points)	0	0	0	0
CR (6 points)	0	0	0	1
Totals	35	39	39	37

*MC stands for multiple-choice and CR stands for constructed-response.

Previous Critical Reading and Standards Research

Previously, I was involved in a Broome-Tioga BOCES study group that focused on signal verbs on the current Regents exams in Algebra I, Geometry, AP examination in Biology, and 8th Grade New York State Science examinations (Simpson, Frech, Garbade, & Schaefer, 2018). Different types of signal verbs were found in the stem of the item. We wanted to see if these verbs were similar across disciplines. After analyzing this data, our group of teachers developed specific reading strategies that could be incorporated in mathematics and science classrooms.

After analyzing Algebra I, I found that the most common signal verb was the word "represent". When I noticed that the word represent was most commonly found in Algebra I, I began to think about how one of the major emphases of the common core was to help students connect multiple representations.

In my research component, I first used the topic area and if available, subtopic area coding that existed on the New York State Regents Exams (Schaefer, 2019a). For *old* and *recent* exams, they were coded by topic area, and for *current* exams, they were also coded by subtopic area. After previously coding all exams by their own topic and subtopic areas (for *current* only), I then coded all exams by *recent* and *current* topics and subtopics using the coding by JMAP.org as interrater reliability (Schaefer, 2019b). Table 5 below shows how these exams were coded.

Table 5

Topics and Subtopic Areas of Old, Recent, *and* Current *Items*

Old Items (n = 764)	*Recent* Items (n = 857)	*Current* Items (n = 518)
Mathematical Reasoning Number and Numeration Operations Modeling/Multiple Representations Measurement Uncertainty Patterns/Functions	Number Sense and Operations Algebra Geometry Measurement Statistics and Probability	Algebra • Arithmetic with Polynomials and Rational Expressions • Creating Equations • Reasoning with Equations and Inequalities • Seeing Structure in Expressions Functions • Building Functions • Interpreting Functions • Linear, Quadratic, and Exponential Models Number and Quantity • The Real Number System • Quantities Statistics • Interpreting Categorical and Quantitative Data

Once I had completed coding all exams by *recent* and *current* standards, it occurred to me that it would be possible to narrowly focus on topic and subtopic areas as a learner and an educator. Although each subtopic area may have its differences, there are representations such as

the visual, symbolic, numeric, and written representations that are woven through each subtopic area. Students may struggle in certain subtopic areas in algebra because they have difficulties in reading, understanding and visualizing the "real world" context, applying their knowledge to novel and complex situations, interpreting various representations, and/or making connections among the representations.

Interrater Reliability

In the first round of coding, I coded around 11% of the total items: old items (78, two complete exams), recent items (78 items, two complete exams), and current items (74, two complete exams) with the representation types, absence or presence of context, relationships of representations, readability, and cognitive complexity coding. I selected two exams of each administration at random. I selected another researcher with expertise in multiple representations in mathematics, real world context in mathematics, readability, and item complexity levels who coded the exams to find interrater reliability.

The unit of analysis in each exam was an item part. For the multiple choice items, there was only one item part. In the constructed response items, there were often multiple item parts. Essentially, students needed to answer all applicable item parts in order to receive full credit for the item. For each unit of analysis different elements were assigned different codes. For example, the item part was assigned a label for the absence or presence of context and the levelcognitive demand complexity (i.e., 0, 1, 2, 3). The stem and responses for a unit were separately assigned one or more of the 15 representational categories and subcategories (i.e., V-G, V-C, V-F, V-D, V-P, S-A, S-S, S-I, S-P, S-G, S-Q, N-N, N-P, N-T, W). The stem-response pair was assigned codes for: presence of a contextual situation; relationship of representations (i.e., RT, RC, RS); and. One exam was randomly chosen; it was the June 2018 current exam. The exam had 32

shorter items and an additional 5 multi-part items with a total of 11 parts; therefore, this exam had 43 total units that were coded independently by two coders. Table 6 outlines the elements of each unit of analysis that were coded, the number of elements for the exam, the types of codes assigned to each element, the number of possible codes assigned, the total number of element-code pairs, and the agreement of the two coders.

Table 6

Coder Agreement for Stems and Responses

Element of Unit Coded	Number of Elements	Type of Code Assigned	Number of Possible Codes Assigned	Total Number of Element-Code Pairs	Coder Agreement
Unit Stem	43	Representational Categories & Subcategories	15	43 x 15 = 645	645/645
Unit Response	43	Representational Categories & Subcategories	15	43 x 15 = 645	644/645
Unit Stem-Response Pair	43	Contextual Situation	1	43	43/43
Unit Stem-Response Pair	43	Relationship of Representations	3	129	129/129
Unit Stem-Response Pair	43	Cognitive Demand Complexity	4	172	171/172

Since agreement between the two coders was above 99% for each coding category, only one coder was used to code the remainder of the data. When confronted with an element that was difficult to code on the remainder of the data, this coder consulted with the second coder on which code to use; this happened two times. Note that the high inter-coder agreement is likely

due to the fact that the two coders worked together to create the coding schemes by discussing the coding in 11 meetings in addition to exchanging 38 emails back and forth related to the coding prior to actual coding in order to refine all the coding schemes. During this communication exchange, the two coders frequently created examples to check their agreement on the coding schemes.

Research Instruments

There are four different schemas that I used to code the items on the *old, recent*, and *current* state algebra exam items. An overview of all of the coding schemas is shown in Table 7 below.

Table 7

Overall Coding Schema

RQ	Coding Schema	Calculation or Frequency	Unit of Analysis
	Preliminary Item Parts Count	• Item Parts Count	By exam
	Preliminary Standards	• Topics and Subtopics Count	By exam
1	Schema 1: Readability	• Flesch Reading Ease • Flesch-Kincaid Grade Level • ATOS Level • Average Word Length • Average Sentence Length • Average Vocabulary Level • Word Count	By exam By constructed response and multiple choice
2	Schema 2: Context	• Presence or absence of context	By exam By constructed response and multiple choice
3	Schema 1: Readability Schema 2: Context	• ATOS Level • Presence or absence of context	By exam items
4	Schema 3: Complexity	• Cognitive Complexity Level	By exam By constructed response and multiple choice
5	Schema 1: Readability Schema 2: Context Schema 3: Complexity	• ATOS Level • Presence or absence of context • Cognitive Complexity Level	By exam items
6	Schema 4: Representations	• Representations Framework	By exam By constructed response and multiple choice
7	Schema 4: Representations	• Representations Framework Stem-to-Response Relationships	By exam By constructed response and multiple choice
8	Schema 5: Representation Relationships	• Representation Relationships Framework	By exam By constructed response and multiple choice

Preliminary Research

Before the research was conducted, a count of the part of items (or also labeled as "item parts") was calculated for all 20 *old,* 22 *recent,* and 14 *current* exams. Also, the percentages of the topic areas for the *old* and *recent* exams along with the topics and subtopic areas of the *current* exams were noted.

Schema 1: Readability

To answer RQ 1, each item was be coded with applicable readability grade levels and measures of reading ease. Each of the *old*, *recent,* and *current* exams was transcribed by JMAP in a doc format. Visuals and symbolic equations were deleted and they did not affect the readability score similar to how Walkington et al. (2018) measured readability. The Flesch reading ease score and the Flesch-Kincaid reading grade level will be computed for each exam administration (i.e., June 2018) and each exam item on the *old*, *recent,* and *current* exams. This was automatically calculated as a readability statistic in Microsoft Word. Then, each item was run through the ATOS for Text Analyzer to find the ATOS level, average vocabulary level, average word length, average sentence length, average vocabulary level, and word count. Appendix A described the steps in finding readability for the exam items.

Readability was determined by a readability score based on the entire exam. It was also determined by items on the constructed response and multiple choice. To answer the exam part of the research question, the unit of analysis was the exam administrations: the readability tests were run on all 20 old exams, 22 recent exams, and 14 current exams to determine the medians of each exam (*old*, *recent*, and *current*). For the constructed response/multiple choice part of the research question, since the unit of analysis was exam items, the readability tests were run on *old*

(n = 204), *recent* (n = 198), and *current* constructed response items (n = 184) as well as *old* (n = 560), *recent* (n = 653), and *current* multiple choice items (n = 336).

Schema 2: Context

To answer RQ 2, the presence or absence of "real world" context was found. Each item was coded for either the presence or absence of a real-world context in the problem. If the item has a real world context present, it was coded with a 1 in the context column. If the item does not have a real world context present, it was coded with a 0 in the context column. Appendix B describes real world context by highlighting examples from the state algebra exams.

The percentage of "real world" context was determined by a relative context percentage for exams as well as separate relative context percentages based on the constructed response and multiple choice item parts. To answer the exam part of the research question, the unit of analysis was the exam administrations: the relative context percentage was found for all 20 old exams, 22 recent exams, and 14 current exams to determine the medians of each exam (*old*, *recent*, and *current*). For the constructed response/multiple choice part of the research question, since the unit of analysis was exam item parts, the relative context percentage for constructed response and multiple choice was run on *old* (n = 221), *recent* (n = 276), and *current* constructed response item parts (n = 288) as well as *old* (n = 560), *recent* (n = 653), and *current* multiple choice item parts (n = 336).

Item Reliability and Context

To answer RQ 3, a relationship was determined between the readability of each exam administration and the relative "real world" context. A scatter plot was plotted for each exam administration using the ATOS reading level's per exams as the point's input and the relative context percentage on all exam items as the point's output.

Schema 3: Complexity

To answer RQ 4, the complexity code is based off of a modified Stein and colleagues (1996, 1998) framework. Item parts will be coded by four levels of cognitive demand: Level 0 (very low), Level 1 (low), Level 2 (high), and Level 3 (very high). Appendix C organizes the levels along with relevant examples of each for coding. Each item in the multiple choice and in the short answer sections will be labeled with the applicable level of cognitive demand.

Cognitive complexity was determined by a relative cognitive complexity percentage for exams as well as separate relative cognitive complexity percentages based on the constructed response and multiple choice item parts. To answer the exam part of the research question, the unit of analysis was the exam administrations: the relative complexity percentage was found for all 20 old exams, 22 recent exams, and 14 current exams to determine the medians of each exam (*old*, *recent*, and *current*). For the constructed response/multiple choice part of the research question, since the unit of analysis was exam item parts, the relative context percentage for constructed response and multiple choice was run on *old* (n = 221), *recent* (n = 276), and *current* constructed response item parts (n = 288) as well as *old* (n = 560), *recent* (n = 653), and *current* multiple choice item parts (n = 336).

Item Readability, Context, and Complexity

To answer RQ 5, relationships among the readability of each exam administration, the relative "real world" context, and cognitive complexity were found by using a calculation of Spearman's rho as data was not normally distributed (Spearman, 1904). This would determine if there were correlations among item readability, context, and complexity.

Schema 4: Representations

To answer RQ 6 and 7, representation types were identified in the stem and response of each item. The representation type instrument is based off of a modified Lesh, Post, and Behr (1987) framework. The stem and response of each item will be coded by visual, numeric, symbolic, and written word categories and if present, their corresponding subcategories (see Appendix D).Table 8 below outlines the categories and sub-categories for stems and responses.

Table 8

Representation Types for Stems and Responses of State Algebra Exam Items

Category	Subcategory
V. Visual	V-G Graphs
	V-C Charts
	V-F Geometric Figures
	V-D Diagrams
	V-P Pictures
S. Symbolic	S-A Algebraic
	S-S Sets
	S-I Intervals
	S-P Probability and Statistics
	S-G Geometric notation
	S-Q Sequences and Series
N. Numeric	N-N Numerals
	N-P Points
	N-T Tables and Matrices
W. Written	-

For each item, the stem of the item will be coded with any applicable category and their corresponding subcategory. For example, an item with a symbolic algebraic equation in the stem will be coded with S-A. If there was also a graph in the stem, the stem will be coded with the previous code of S-A and the code of V-G. Then, the response of the item will be coded with any applicable category and their corresponding subcategories. For multiple choice items, the responses are the four answer choices, and they will be coded with the applicable categories and/or subcategories. For constructed response items, the responses will be the second suggested response found in the "Scoring Key and Rating Guide" of the exam for the *recent* and *current* exams, and the completed exam by JMAP for the *old* exam.

For the stem and response of the item, the number of representations was counted based on each exam. The count of the representations of the stem will be added to the number of representations of the response to find a total number of representations in the item.

After calculating the medians for all representations, four prominent representations among all exams (*old*, *recent*, and *current*) were selected. To answer the exam part of the research question, the unit of analysis was the exam administrations: the relative representation percentage for the four prominent representations was found for all 20 old exams, 22 recent exams, and 14 current exams to determine the medians of each exam (*old*, *recent*, and *current*). For the constructed response/multiple choice part of the research question, since the unit of analysis was exam item parts, the relative context percentage for constructed response and multiple choice was run on *old* (n = 221), *recent* (n = 276), and *current* constructed response item parts (n = 288) as well as *old* (n = 560), *recent* (n = 653), and *current* multiple choice item parts (n = 336).

To understand the percentages of how representations in the stem interacted with representations in the response for research question 7, a 4x4 stem-to-response matrix for *old*, *recent*, and *current* exams was created. After determining the relative stem-to-response percentages for all 20 old exams, 22 recent exams, and 14 current exams, 16 stem-to-response medians were calculated for each exam (*old*, *recent*, and *current*). To illustrate these median relationships more effectively, Sankey diagrams were created to show the amount of connections among stem-to-response on the *old*, *recent*, and *current* exams.

Schema 5: Representational Relationships

To answer RQ 8, each stem and response of each item part was coded for one relationship among representations. Then, a relative representation relationship percentage was calculated for the 20 *old*, 22 *recent*, and 14 *current* exams. The relationships among the representations in the problem are *treatments (RT)*, *conversions (RC)*, and *synergy (RS)*. A item might be coded as with both *synergy (RS)* and *conversion (RC)* if an item stem gives requisite information as written word and symbolic representations *(RS)* and then the response that is required is visual *(RC)*. Appendix E describes each of these relationships between representations and highlights applicable examples from the state algebra exams.

For the constructed response/multiple choice part of the research question, since the unit of analysis was exam item parts, the relative context percentage for constructed response and multiple choice was run on *old* (n = 221), *recent* (n = 276), and *current* constructed response item parts (n = 288) as well as *old* (n = 560), *recent* (n = 653), and *current* multiple choice item parts (n = 336). Relative representation relationship percentage medians were calculated for constructed response and multiple choice item parts in the *old*, *recent*, and *current* exams.

Research Methodology and Rationale

This study is both quantitative descriptive and quantitative correlational. Descriptive research is important in studying state algebra items because teachers can see if there are trends or patterns in the way topics are presented on the exams. I chose to categorize state algebra items by more than only the standard because there is benefit to understanding how state algebra exam items are typically written. Teachers will have more of an awareness of the types of representations and relationships among representations. Also, teachers would be able to see the types of "real world" contexts that have been asked on previous exams. Seeing these patterns will empower teachers to build their own personalized curriculum and assessments that still meet the expectations of the Common Core curriculum. Students would be able to see which representations are used on exams. Also, students would be more able to prepare for the reading and context required to answer items on the *current* exam.

Once I have this data, I will be able to see if there are any noticeable correlations among readability, context, complexity, representations, and relationships of representations. Although this analysis is only based on the subset of New York State algebra exam items, it may lead to significant conclusions about how these different components of the exam item may interact with one another.

Data Analysis

Previously, the exams were compared by recent and current standards. Different topics were found as major and important topics according to the recent and current standards. With this data, I would see if exams included more representation types than other exams, and if these patterns were more prevalent on the constructed response items versus the multiple choice items. I would see if there were correlations among readability, context, and complexity.

Once interrater reliability was found, I checked with the other coder to make sure that I was coding other items correctly. The entire data set was first coded by readability scores for each item and each exam (RQ 1). The items were grouped by constructed response and multiple choice, and a median of the readability scores was determined by exam and item type (constructed response or multiple choice). Next, the items were coded with either the presence or absence of context, and then a relative context percentage was calculated for each exam administration and then by constructed response/multiple choice (RQ 2). Correlations were calculated between readability and context (RQ 3). Next, the items were coded with the cognitive complexity level, and then a relative complexity percentage was calculated for each exam administration and then by constructed response/multiple choice (RQ 4). Correlations were calculated among readability, context, and complexity (RQ 5).

Next, the items were coded with the applicable representation(s). The number of representations and the average number of representations were calculated by exam and constructed response/multiple choice. Four prominent representations were found. Then, a relative representation percentage was calculated for each of the prominent representations for each exam administration and then by constructed response/multiple choice (RQ 6). Based on the four prominent representations, stem-to-response medians were calculated for each exam administration and for constructed response and multiple choice items (RQ 7). Finally, the item parts were coded with the representation relationship, and then a relative representation relationship percentage was calculated for each exam administration and then by constructed response/multiple choice (RQ 8).

Summary

This chapter describes the data set of the New York State algebra standards and exam items. Then, it describes how the schemas will be used for coding the algebra exam items while discussing the methodology and rationale for selecting these research methods. Finally, it describes how the codes will be analyzed to find trends in algebra state exam items.

Chapter 4 – Results

The purpose of this study was to determine any patterns in readability, context, complexity, and representations by looking at different versions of the New York State Algebra exam (*old*, *recent*, and *current*) and related standards (*recent* and *current*). This chapter is in order of the eight research questions posed. Before the discussion of the first research question, there was an analysis of the number of items on exams and the exams' item types, comparing the number of single-part items to the number of multi-part items per exam version. In the next section, the first research question was addressed as several readability tests were used to analyze *old*, *recent*, and *current* exam versions. The following section focused on the second research question where the presence or absence of context was analyzed by exam version. Then, to address the third research question, both the readability and the relative context were considered together in order to determine any relationship between those variables. In the next section, the fourth research question was considered as the cognitive complexity was analyzed by exam version. The following section addressed the fifth research question as correlations were determined among cognitive complexity, context, and readability. To address the sixth research question, the number and type of representations were determined by exam version. The seventh research question addressed the representation relationships among stem and response by exam version. Finally, the eighth research question was an analysis of the patterns of representations and representation treatments according to complexity. After each of the results related to each research question are presented, a general summary of results is provided as a conclusion to this chapter.

Preliminary Analysis of Items and Item Parts

The *old* exam had 35 items on each of first four administrations (that occurred from June 2002 to June 2003), but for the rest of the administrations, there were 39 items on each exam. The *recent* exam consistently had 39 items on each administration, and the *current* exam consistently had 37 items. However, upon examining all three versions of the exam (*old*, *recent*, and *current*), it was noted that was also variance in the number of constructed response items that contained multiple parts. Item parts include all of the individual parts that make up a constructed response item. For instance, while multiple-choice items are comprised of one part (i.e., a single stem and four different options), constructed response items may have one, two or more related parts that need to be answered in order to receive full credit for the item, resulting in one, two or more distinct answers for the item. Therefore, each exam administration may have more tasks for students to calculate in the parts of the item than the number of items present on the exams. After the number of parts was determined for each exam administration (June 2002, August 2003, etc.), it was noted that the number of item parts increased from the *old* exam version to the *recent* exam version then again to the *current* exam version. The median number of parts for *old* exams was 39, for *recent* exam item was 42, and for *current* exams was 45. This is illustrated in figure 6.

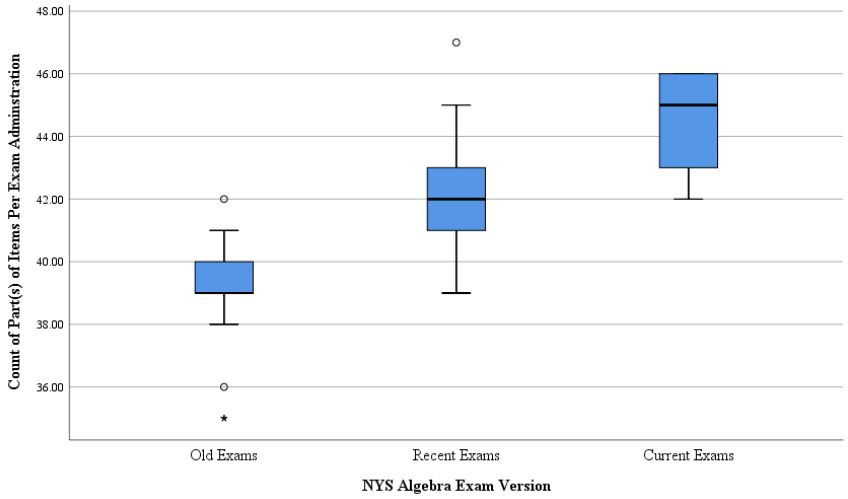


Figure 6. Box and whisker plots of number of item parts for *old*, *recent*, and *current* exams.

Preliminary Standards Research

From previous research, I labeled all items with their respective topics and subtopic areas, if applicable, from the standards that were in place at the time of the state algebra exam. If 25% or more of the exam items (for the *old*, *recent*, and *current* exams) were classified as addressing a topic area, that area was labeled as a **major** topic area. If the topic area had addressed 10% to 25% of the exam items, it was labeled as an **important** topic area. Finally, if the topic area addressed less than 10% of the exam items, it was labeled as a **minor** topic area.

***Old* exams.**

On the old state algebra exams, no **major** topics were identified. *Modeling/Multiple Representation*, *Operations*, *Measurement*, and *Patterns and Functions* were **important** topics with each topics accounting for 18-20% of the total items on each *old* exam. *Uncertainty*, *Number and Numeration*, and *Mathematical Reasoning* were **minor** topics with each topic accounting for 7-8% of the total items on each *old* exam. This information is displayed in Table 9.

Table 9

Counts and Relative Counts of Items on Old *State Algebra Exams by* Old *Topic Areas*

	Counts	Relative Counts
Mathematical Reasoning	52	7%
Number and Numeration	63	8%
Operation	148	19%
Modeling/Multiple Representation	155	20%
Measurement	148	19%
Uncertainty	62	8%
Patterns/Functions	136	18%
Totals	764	100%

***Recent* exams.**

When considering the items on *recent* state algebra exams, 54% of the exam items were in the *Algebra* topic area; so, it was deemed a **major** topic. Other **important** topics included *Geometry* as well as *Statistics and Probability* which each accounted for 16% of the total items on the exam. *Number Sense* and *Operations and Measurement* were **minor** topics including 6-8% of the total items for each standard category. This information is displayed in Table 10.

Table 10

Counts and Relative Counts of Items on Recent *State Algebra Exams by* Recent *Topic Areas*

Topic Area	Item Counts	Relative Counts
Algebra	461	54%
Geometry	135	16%
Statistics and Probability	138	16%
Number Sense and Operations	71	8%
Measurement	52	6%
Totals	857	100%

***Current* exams.**

For the *current* state algebra exams, the **major** topics were identified as *Algebra* and *Functions* with 49% and 37% of the total items on the exams, respectively. *Number and Quantity* and *Statistics* were found to be **minor** topic areas with 5% and 9% of the total exam items. This information is displayed in Table 11.

Table 11

Counts and Relative Counts of Items on Current *State Algebra Exams by* Current *Topic and Subtopic Areas*

Current Topic and Subtopic Area	Counts	Relative Counts
Algebra	**256**	**49%**
Arithmetic with Polynomials and Rational Expressions	27	5%
Creating Equations	66	12%
Reasoning with Equations and Inequalities	110	21%
Seeing Structure in Expressions	53	10%
Number and Quantity	**25**	**5%**
The Real Number System	14	2%
Quantities	11	2%
Functions	**191**	**37%**
Building Functions	30	6%
Interpreting Functions	117	23%
Linear, Quadratic, and Exponential Models	44	8%
Statistics	**46**	**9%**
Totals	**518**	**100%**

There were several important findings in terms of the standards topics. In terms of topic areas, *recent* and *current* topics are most similar to one another. However, on the *recent* state algebra exams, there were 91 standards divided among the five topics while on the *current* exams, there were governed 45 standards divided among the four topic areas with subtopic areas.

Old exams were governed by 32 standards divided among seven topic areas. Although there were fewer topic areas and standards with the *current* exam, there were more disagreements about the standard when looking at the coding by the NYSED Regents and JMAP.org coding (Schaefer, 2019) as standards seemed to overlap and not be as specific as the *recent* standards. As an improvement to previous versions of the state algebra exam, on the *current* exams, items were also labeled with appropriate subtopics as well. Previously (i.e. on *old* and *recent* exams), only coarser grained topics were labeled.

All three versions of the state algebra exam included algebra topic areas. However, these algebra topic areas were expressed in various ways in terms of their readability, context, complexity, representations, and relationships among representations.

Research Question 1: Readability for Exam Versions and Standards

Readability was measured by several readability tests: the Flesch Reading Ease score, the Flesch-Kincaid Grade Level, and ATOS level. By using the ATOS calculator, word count, average word length, average sentence length, and average vocabulary level were also determined. Each of the readability tests was run for all exam administrations from June 2002 to August 2018. Since the unit of analysis was each exam, the readability tests were run on all 20 old exams, 22 recent exams, and 14 current exams. All words on the entire exam (including all exam items but excluding directions, numbering of items, and algebraic expressions) were run through online calculators to find Flesch Reading Ease Score, Flesch-Kincaid Grade Level, ATOS level, word count, average word length, average sentence length, and average vocabulary level.

By exam version.

Median scores were determined for each of the three types of exams (i.e., *old, recent, current*) for each readability tests. The exam versions showed a decreasing Flesch reading ease score and increasing Flesch-Kincaid Grade and ATOS Levels, indicating that *recent* exams were more difficult to read then the *old* exams, and *current* exams were yet even more difficult to read than *recent* exams. The median of the average word length and vocabulary level only increased marginally across the three versions of the exam. In fact, the average sentence length decreased from the *old* to the *recent* exams, and increased from the *recent* to the *current* exam. Table 12 illustrates the changes in medians for each of the readability tests. Another point that deserves notice is that the word count on the *current* exams is much higher than either the *old* or *recent* versions of the exam.

Table 12

Readability Test Medians for Old, Recent, *and* Current *Exams*

Readability Tests	Median		
	Old	*Recent*	*Current*
Flesch Reading Ease Score	69.4	67.6	66.3
Flesch-Kincaid Grade Level	6.4	6.4	6.9
ATOS Level	7.0	7.1	7.5
Average Word Length	4.3	4.4	4.5
Average Sentence Length	11.9	11.3	12.3
Average Vocabulary Level	3.9	4.0	4.0
Word Count	940.5	932.5	1281.5

The Flesch Reading Ease Score decreased in median for each exam version, *old* (69.4), *recent* (67.6), and *current* (66.3), indicating that the reading difficulty has slightly increased from *old* to *current* exams. By using further statistical testing, there was a notable difference between *old* and *current* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in Flesch Reading Ease among the different exam versions, $\chi^2(2) = 6.906$, $p = 0.032$ with a mean rank score of 34.88 for *old* exams, 28.14 for *recent* exams, and 19.96 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams (z = 2.625, p = 0.026 (adjusted Bonferroni). There were no statistically significant differences between other groups. This was further illustrated in Figure 7.

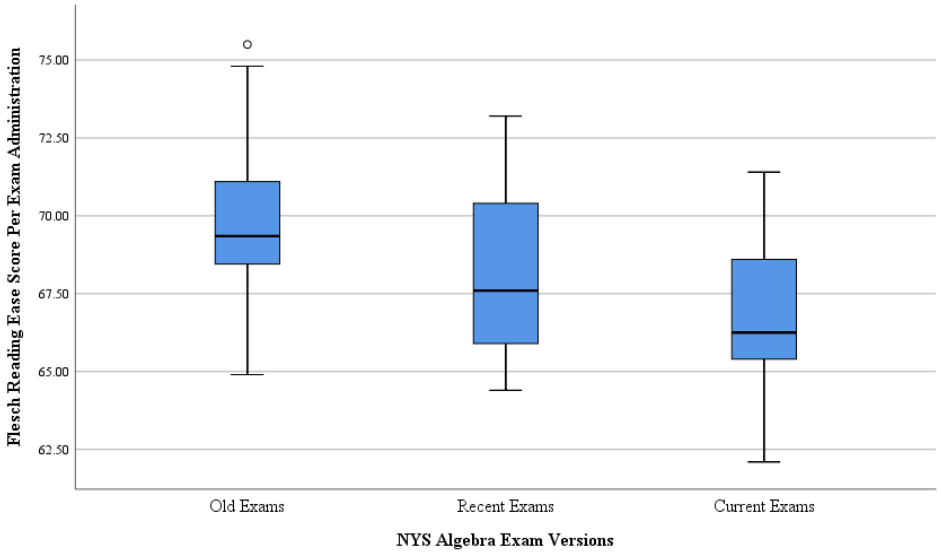


Figure 7. Box and whisker plots of Flesch Reading Ease for *old*, *recent*, and *current* exams.

As would be expected, as the Flesch Reading Ease scores decrease, the reading difficulty scores increase. The Flesch-Kincaid Reading Grade Level increased in median for each exam version: *old* (6.3), *recent* (6.4), and *current* (6.9), indicating that the reading difficulty has slightly increased from old to current exams. By using further statistical testing, there were statistically significant differences between *old* and *current* as well as *recent* and *current* exams.

A Kruskal-Wallis H test showed that there was a statistically significant difference in Flesch-Kincaid scores among the different exam versions, $\chi^2(2) = 15.019$, $p = 0.001$ with a mean rank score of 21.27 for *old* exams, 26.07 for *recent* exams, and 42.64 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = -3.769$, $p < 0.001$ ($p < 0.001$) (adjusted Bonferroni). There was a statistically significant difference among the *recent* and *current* exams ($z = -2.980$, $p = 0.009$). There were no statistically significant differences between other groups. This is illustrated in Figure 8.

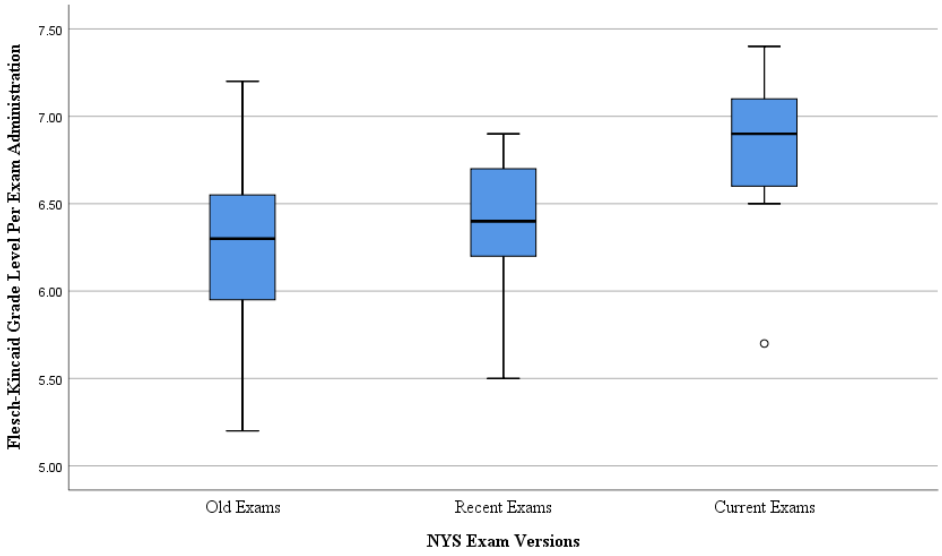


Figure 8. Box and whisker plots of Flesch-Kincaid Grade Level for *old*, *recent*, and *current* exams.

The ATOS Level increased in median for each exam version: *old* (7.0), *recent* (7.1) and *current* (7.5), indicating that the reading difficulty has increased from *old* to *current* exams. Figure X illustrates the increase in reading difficulty for ATOS level; note that this measure of reading difficulty is very similar to the findings for the Flesch-Kincaid Grade Level measure that was reported above. By using further statistical testing, there were statistically significant

differences between *old* and *current* as well as *recent* and *current* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in ATOS scores among the different exam versions, $\chi^2(2) = 13.515$, $p = 0.001$ with a mean rank score of 23.40 for *old* exams, 24.34 for *recent* exams, and 42.32 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = -3.338$, $p = 0.003$) (adjusted Bonferroni). There was a statistically significant difference among the *recent* and *current* exams ($z = -3.234$, $p = 0.004$). There were no statistically significant differences between other groups. The increase in ATOS level is indicated in Figure 9.

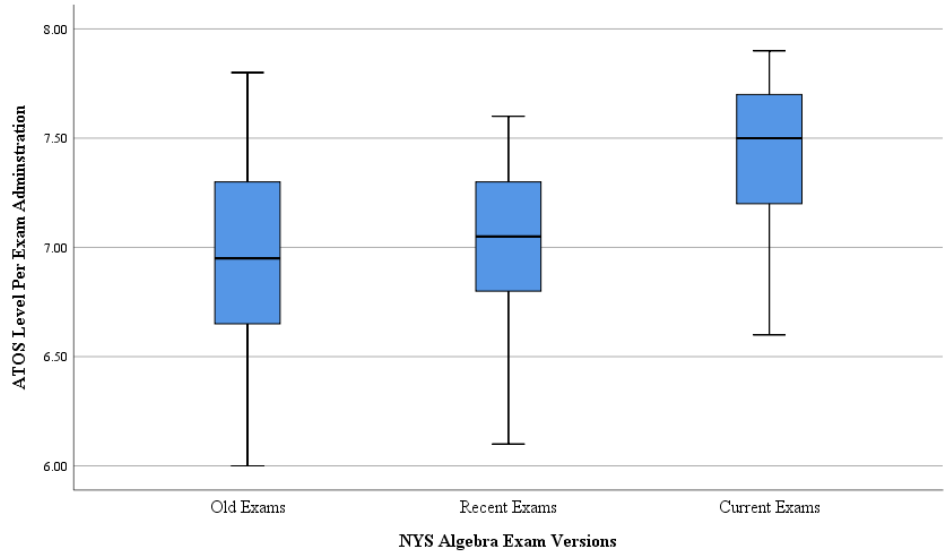


Figure 9. Box and whisker plots of ATOS Level for *old*, *recent*, and *current* exams.

The median Average Word Length increased for each exam version: *old* (4.3), *recent* (4.4) and *current* (4.5), indicating that the length of words has increased from *old* to *current*. By using further statistical testing, there were statistically significant differences between *old* and *recent* as well as *old* and *current* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in average word length scores among the different exam versions, $\chi^2(2) = 12.505$,$p = 0.002$ with a mean rank score of 19.30 for *old* exams, 31.16 for *recent* exams, and 37.46 for *current* exams. When running a Dunn-Bonferroni test, there was a

statistically significant difference between the *old* and *recent* exams (z = -11.859, p = 0.038) (adjusted Bonferroni). There was a statistically significant difference among the *old* and *current* exams (z = -18.164, p = 0.002). There were no statistically significant differences between other groups. This is displayed in figure 10.

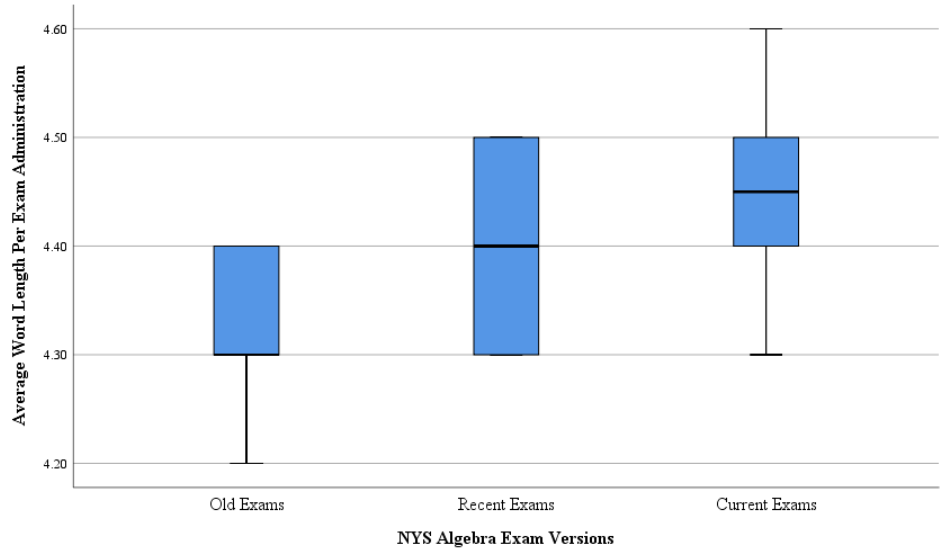


Figure 10. Box and whisker plots of Average Word Length for *old*, *recent*, and *current* exams.

The Average Sentence Length decreased in median for *old* (11.9) to *recent* (11.3), and then increased in median for *current* (12.3). By using further statistical testing, there were statistically significant differences between *recent* and *current* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in average sentence length scores among the different exam versions, $\chi^2(2) = 7.324$, p = 0.026 with a mean rank score of 28.93 for *old* exams, 22.41 for *recent* exams, and 37.46 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *recent* and *current* exams (z = -15.055, p = 0.021) (adjusted Bonferroni). This is illustrated in Figure 11.

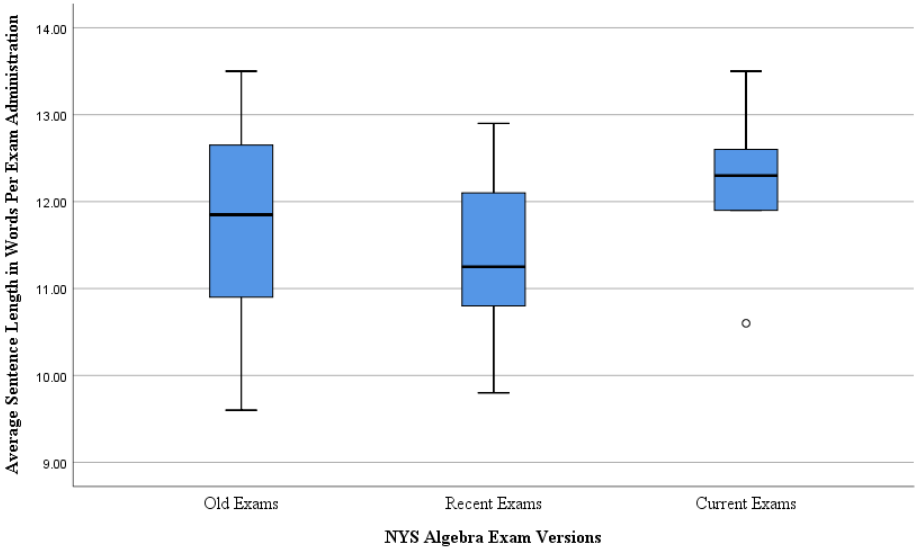


Figure 11. Box and whisker plots of Average Sentence Length in words for *old*, *recent*, and *current* exams.

The Average Vocabulary Level increased in median from *old* (3.9) to *recent* (4.0), and then stayed about the same for *current* (4.0). After further testing, there was no significant difference found in average vocabulary level among the three versions of the exam. A Kruskal-Wallis H test showed that there was no statistically significant difference in average vocabulary scores among the different exam versions, $\chi^2(2) = 3.632$, $p = 0.163$ with a mean rank score of 24.13 for *old* exams, 28.48 for *recent* exams, and 34.79 for *current* exams. This is displayed in figure 12.

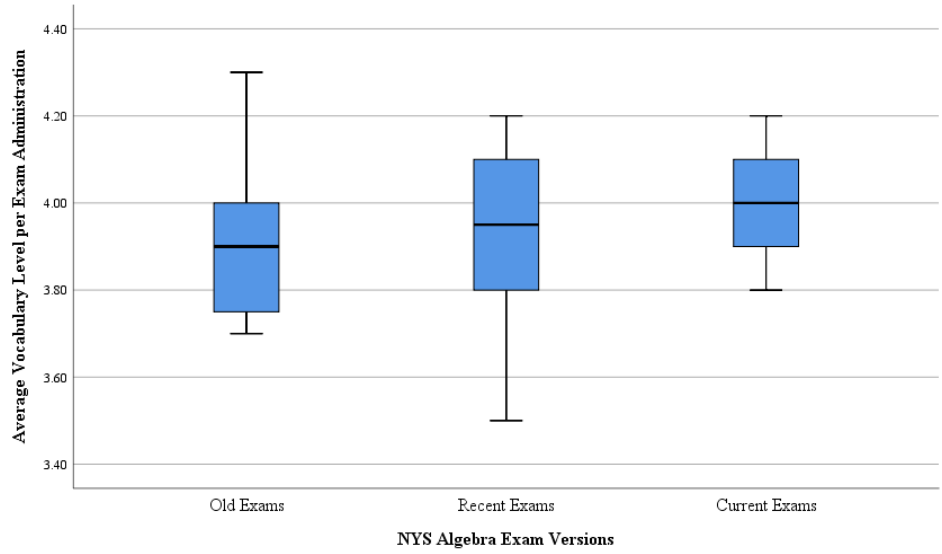


Figure 12. Box and whisker plots of Average Vocabulary Level for *old*, *recent*, and *current* exams.

Finally, the Average Word Count dramatically increased from *old* to the *current* exam version. The Average Word Count slightly decreased in median from *old* (940.5) to *recent* (932.5), and then increased for the *current* exam (1281.5). There was an increase in word count for all exam versions; it also displays information that even though there were more words on the *current* exam, it was also the exam that was most variable across administrations for the word count. By using further statistical testing, there were statistically significant differences between *old* and *current* as well as *recent* and *current* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in Word Count scores among the different exam versions, $\chi 2(2) = 29.145, p < 0.001$ with a mean rank score of 22.38 for *old* exams, 21.11 for *recent* exams, and 48.86 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = -27.744$, $p = 0.000$) (adjusted Bonferroni). There was a statistically significant difference among the *recent* and *current* exams ($z = -26.482$, $p < 0.001$). There were no statistically significant differences between other groups. This is shown in figure 13.

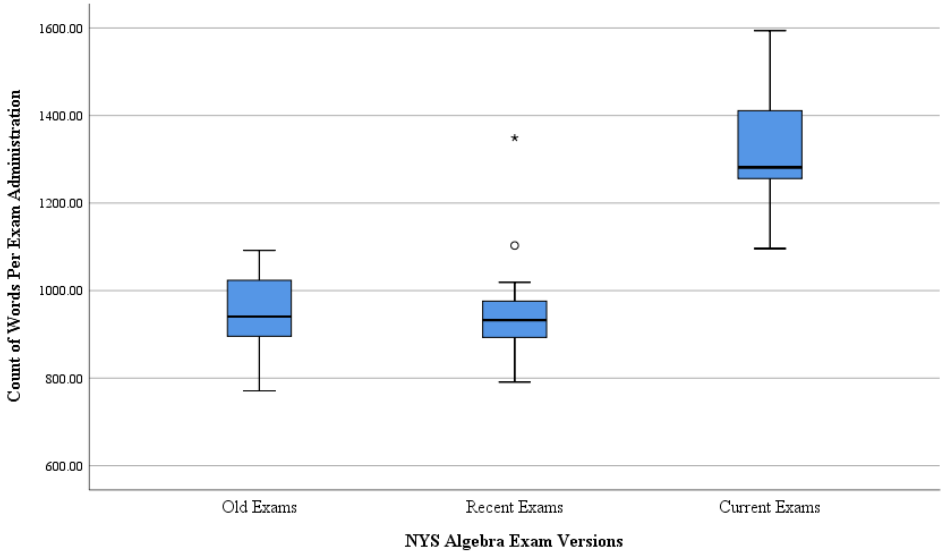


Figure 13. Box and whisker plots of Word Count for *old*, *recent*, and *current* exams.

By item type.

Medians were determined for the two item types (i.e., constructed response and multiple choice items) on the *old*, *recent*, and *current* version of the state algebra exam. Each of the readability tests was run on each individual item on all exam administrations from June 2002 to August 2018. Since the unit of analysis was an exam item, the readability tests were run on *old* constructed response items (n = 204), *recent* constructed response items (n = 198), and *current* constructed response items (n = 184) as well as *old* multiple choice items (n = 560), *recent* multiple choice items (n = 653), and *current* multiple choice items (n = 336). Each exam item excluding directions, numbering of items, and algebraic notation was calculated using online calculators to find Flesch Reading Ease Score, Flesch-Kincaid Grade Level, ATOS level, word count, average word length, average sentence length, and average vocabulary level.

Similar patterns emerged as before, but there was a more marked increase in readability difficulty on the multiple choice items when comparing *old* multiple choice items to *current* multiple choice items. For instance, Table 13 shows that the median word count of *old* and *recent* multiple choice items was half of the word count of the *current* multiple choice items.

Table 13

Readability Test Medians by Item Type for Old, Recent, *and* Current *Exams*

Item Type	Readability Tests	Median		
		Old	*Recent*	*Current*
Constructed Response	Flesch Reading Ease Score	64.7	61.8	65.8
	Flesch-Kincaid Grade Level	7.5	7.6	6.9
	ATOS Level	7.0	7.0	7.0
	Average Word Length	4.3	4.5	4.5
	Average Sentence Length	13.8	11.5	11.1
	Average Vocabulary Level	3.7	3.8	3.8
	Word Count	36.0	31.0	37.0
Multiple Choice	Flesch Reading Ease Score	68.0	68.7	66.1
	Flesch-Kincaid Grade Level	6.6	6.4	7.0
	ATOS Level	7.0	7.0	7.4
	Average Word Length	4.4	4.4	4.5
	Average Sentence Length	10.0	10.0	11.0
	Average Vocabulary Level	4.4	4.1	4.3
	Word Count	16.0	15.0	27.0

Additional testing.

For Flesch Reading Ease, there was a significant difference between recent and current exams on the constructed response, but there was no significant difference on the multiple choice items. A Kruskal-Wallis H test showed that there was a statistically significant difference in Flesch Reading Ease for constructed response among the different exam versions, $\chi^2(2) = 8.768$, $p = 0.012$ with a mean rank score of 286.10 for *old* exams, 272.00 for *recent* exams, and 321.97 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant

difference between the *recent* and *current* exams (z = -49.967, p = 0.012) (adjusted Bonferroni). A Kruskal-Wallis H test showed that there was no statistically significant difference in Flesch Reading Ease for multiple choice among the different exam versions, $\chi^2(2) = 5.493$, p = 0.064 with a mean rank score of 773.92 for *old* exams, 803.96 for *recent* exams, and 733.89 for *current* exams.

For Flesch-Kincaid Grade Levels, there was a statistically significant difference between *recent* and *current* exams for both constructed response and multiple choice. There was a statistically significant difference among *old* and *current* exams for constructed response as well. A Kruskal-Wallis H test showed that there was a statistically significant difference in Flesch-Kincaid Grade Level score for constructed response among the different exam versions, $\chi^2(2) = 10.614$, p = 0.005 with a mean rank score of 305.30 for *old* exams, 310.26 for *recent* exams, and 258.84 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams (z = 46.461 , p = 0.021) (adjusted Bonferroni). There was also a statistically significant difference between the *recent* and *current* exams (z = 51.425 , p = 0.009) (adjusted Bonferroni). A Kruskal-Wallis H test showed that there was a statistically significant difference in Flesch-Kincaid Grade Level score for multiple choice among the different exam versions, $\chi^2(2) = 7.486$, p = 0.024 with a mean rank score of 789.73 for *old* exams, 744.72 for *recent* exams, and 823.71 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *recent* and *current* exams (z = -78.990, p = 0.026) (adjusted Bonferroni).

For ATOS levels, there were no statistically significant differences in levels among the exams for constructed response, but there was a statistically significant difference between *old* and *current* as well as *recent* and *current* for multiple choice. A Kruskal-Wallis H test showed

that there was no statistically significant differences in ATOS level for constructed response among the different exam versions, $\chi^2(2) = 4.051$, $p = 0.132$ with a mean rank score of 273.48 for *old* exams, 302.64 for *recent* exams, and 302.79 for *current* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in ATOS level for multiple choice among the different exam versions, $\chi^2(2) = 17.664$, $p < 0.001$with a mean rank score of 769.56 for *old* exams, 740.99 for *recent* exams, and 864.66 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *recent* and *current* exams ($z = -123.672$, $p < 0.001$) (adjusted Bonferroni). When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = -95.107$, $p = 0.006$) (adjusted Bonferroni).

For average word length, there was a significant difference between *old* and *recent* exams for constructed response items only. There were no significant differences among exams for multiple choice items. A Kruskal-Wallis H test showed that there was a statistically significant difference in average word length for constructed response among the different exam versions, $\chi^2(2) = 11.886$, $p = 0.003$ with a mean rank score of 261.68 for *old* exams, 318.82 for *recent* exams, and 298.41 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *recent* exams ($z = -57.134$, $p = 0.002$) (adjusted Bonferroni). A Kruskal-Wallis H test showed that there was no statistically significant difference in average word length for multiple choice among the different exam versions, $\chi^2(2) = 5.711$, $p = 0.058$ with a mean rank score of 760.44 for *old* exams, 766.65 for *recent* exams, and 829.52 for *current* exams.

For average sentence length, there was a statistically significant difference between *old* and *current* exams for both constructed response and multiple choice items. There was also a

statistically significant difference between *old* and *recent* constructed response items. A Kruskal-Wallis H test showed that there was a statistically significant difference in average sentence length for constructed response among the different exam versions, $\chi^2(2) = 18.512$, $p = 0.000$ with a mean rank score of 332.73 for *old* exams, 277.76 for *recent* exams, and 263.45 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = 69.285$, $p < 0.001$) (adjusted Bonferroni). There was a statistically significant difference between the *old* and *recent* exams ($z = 54.973$, $p = 0.003$) (adjusted Bonferroni). A Kruskal-Wallis H test showed that there was a statistically significant difference in average sentence length for multiple choice among the different exam versions, $\chi^2(2) = 6.823$, $p = 0.033$ with a mean rank score of 758.92 for *old* exams, 765.50 for *recent* exams, and 834.32 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = -75.403$, $p = 0.045$) (adjusted Bonferroni).

For average vocabulary level, there was a statistically significant difference between *old* and *recent* exams for both constructed response and multiple choice items. There was also a statistically significant difference between the *old* and *current* exams for constructed response items. A Kruskal-Wallis H test showed that there was a statistically significant difference in average vocabulary level for constructed response among the different exam versions, $\chi^2(2) = 10.294$, $p = 0.006$ with a mean rank score of 261.95 for *old* exams, 308.21 for *recent* exams, and 309.64 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *recent* exams ($z = -46.261$, $p = 0.018$) (adjusted Bonferroni). When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = -47.689$, $p = 0.017$) (adjusted Bonferroni). A

Kruskal-Wallis H test showed that there was a statistically significant difference in average vocabulary level for multiple choice among the different exam versions, $\chi^2(2) = 8.148$, $p = 0.017$ with a mean rank score of 816.86 for *old* exams, 743.34 for *recent* exams, and 781.21 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *recent* exams ($z = 73.522$, $p = 0.013$) (adjusted Bonferroni).

For word count, there were no statistically significant differences among exam versions for constructed response items, but there were statistically significant differences between *old* and *current* as well as *recent* and *current* exam items for multiple choice items. A Kruskal-Wallis H test showed that there was no statistically significant difference in word count for constructed response among the different exam versions, $\chi^2(2) = 3.597$, $p = 0.166$ with a mean rank score of 294.24 for old exams, 275.90 for *recent* exams, and 308.61 for *current* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in word count for multiple choice among the different exam versions, $\chi^2(2) = 40.654$, $p < 0.001$ with a mean rank score of 737.21 for *old* exams, 742.23 for *recent* exams, and 916.13 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *recent* and *current* exams ($z = -178.920$, $p < 0.001$) (adjusted Bonferroni). When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = -173.901$, $p < 0.001$) (adjusted Bonferroni).

Research Question 2: "Real World" Context for Exam Versions and Standards

Real world context was measured on each item of an exam by either the presence (1) or absence (0) of "real world" or applied situational context in the item part. The presence or absence of context was determined for all exam administrations from June 2002 to August 2018, and a relative context percentage (i.e., a calculation of the percentage of items parts that included

an applied situation or “real world” scenario compared to all item parts) for each exam administration was calculated. Since the unit of analysis was each exam, a relative context percentage was determined for all 20 *old*, 22 *recent*, and 14 *current* exams.

By exam version.

Overall, there has been an increase in medians for the relative context percentage with a slight decrease from *old* (32.5%) to *recent* (31.5%) and a substantial increase from *recent* to *current* (49%). By running further statistical testing, there were statistically significant differences between *old* and *current* as well as *recent* and *current* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in relative context among the different exam versions, $\chi^2(2) = 22.869$, $p < 0.001$ with a mean rank score of 21.78 for *old* exams, 23.16 for *recent* exams, and 46.50 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = -24.725$, $p < 0.001$) (adjusted Bonferroni). When running a Dunn-Bonferroni test, there was a statistically significant difference between the *recent* and *current* exams ($z = -23.341$, $p < 0.001$) (adjusted Bonferroni).Figure 14 illustrates the increase in relative context percentage by exam for the three exam versions.

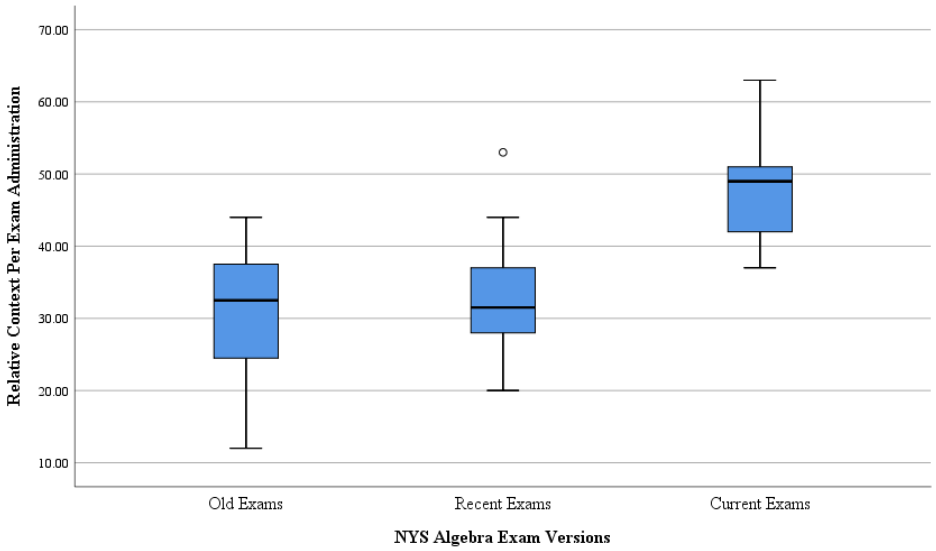


Figure 14. Box and whisker plots of relative context percentage for *old*, *recent*, and *current* exams.

By item type.

Medians were determined for the two item types (i.e., constructed response and multiple choice items) on the *old*, *recent*, and *current* version of the state algebra exam. A relative context percentage was calculated for all exam administrations from June 2002 to August 2018. Since the unit of analysis was an exam item part, the readability tests were run on *old* constructed response items (n = 204), *recent* constructed response items (n = 198), and *current* constructed response items (n = 184) as well as *old* multiple choice items (n = 560), *recent* multiple choice items (n = 659), and *current* multiple choice items (n = 336). When looking more specifically at constructed response and multiple choice items, there was a slight dip in the relative context percentage for constructed response item parts with *old* (52.5%), *recent* (46.1%), and *current* (48.9%). Through further statistical testing, there was no statistical differences among the exams. A Kruskal-Wallis H test showed that there was no statistically significant difference in relative context among the different exam versions, $\chi^2(2) = 0.294$, $p = 0.876$ with a mean rank score of

28.00 for *old* exams, 27.73 for *recent* exams, and 30.43 for *current* exams. However, there is an increase in relative context percentage for the exams' multiple-choice items: *old* (19.5%), *recent* (26.2%), and *current* (39.2%). Through further statistical testing, there were statistical differences between *old* and *current* as well as *recent* and *current* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in relative context among the different exam versions, $\chi^2(2) = 29.237$, $p < 0.001$ with a mean rank score of 16.15 for *old* exams, 28.11 for *recent* exams, and 46.75 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = -30.600$, $p < 0.001$) (adjusted Bonferroni). When running a Dunn-Bonferroni test, there was a statistically significant difference between the *recent* and *current* exams ($z = -18.636$, $p = 0.002$) (adjusted Bonferroni).Table 14 outlines the differences in relative context percentage. Figures 15 and 16 below display the box and whiskers plots of these same results.

Table 14

Relative Context Percentage by Item Type (Constructed Response, Multiple Choice) for Old, Recent, *and* Current *Exams*

Item Type	Presence/Absence of Context	*Old %*	*Recent %*	*Current %*
Constructed Response	Context	19.5	26.2	39.2
	No Context	80.5	73.8	60.8
Multiple Choice	Context	52.5	46.1	48.9
	No Context	47.5	53.9	51.1

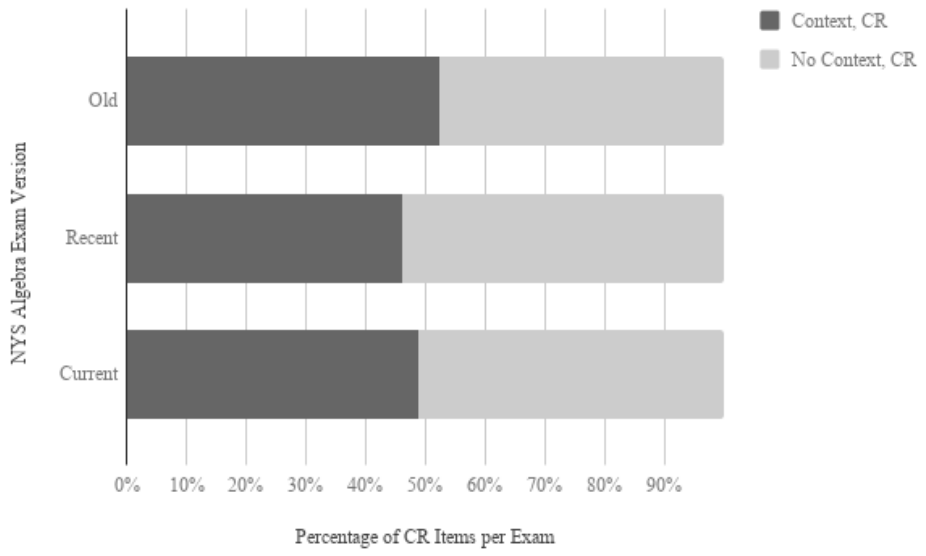


Figure 15. Stacked bar graph of the Relative Context Percentage by item type (constructed response) for *old*, *recent*, and *current* exams.

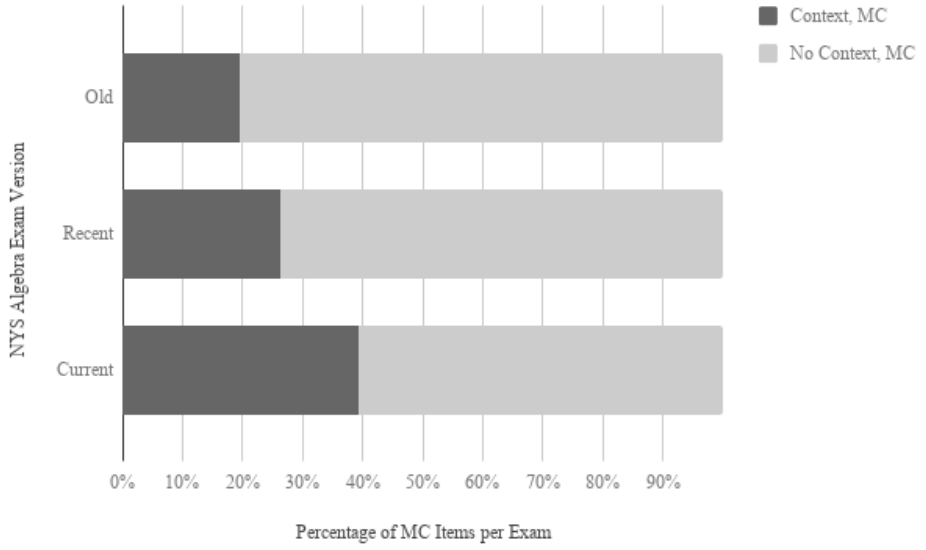


Figure 16. Stacked bar graph of the Relative Context Percentage by item type (multiple choice) for *old*, *recent*, and *current* exams.

Research Question 3: Relationship between Item Readability and Context

To determine if there was a relationship between the readability of each exam administration and the relative "real world" context, a scatter plot was plotted for each exam administration using the ATOS reading level's per exams as the point's input and the relative context percentage on all exam items as the point's output. A coefficient of determination was found (Wright, 1921). There was no significant relationship between the readability and relative real world context percentage when the unit of analysis was an exam as displayed in Figure X, $r^2 = 0.216$. This relationship is displayed in figure 17.

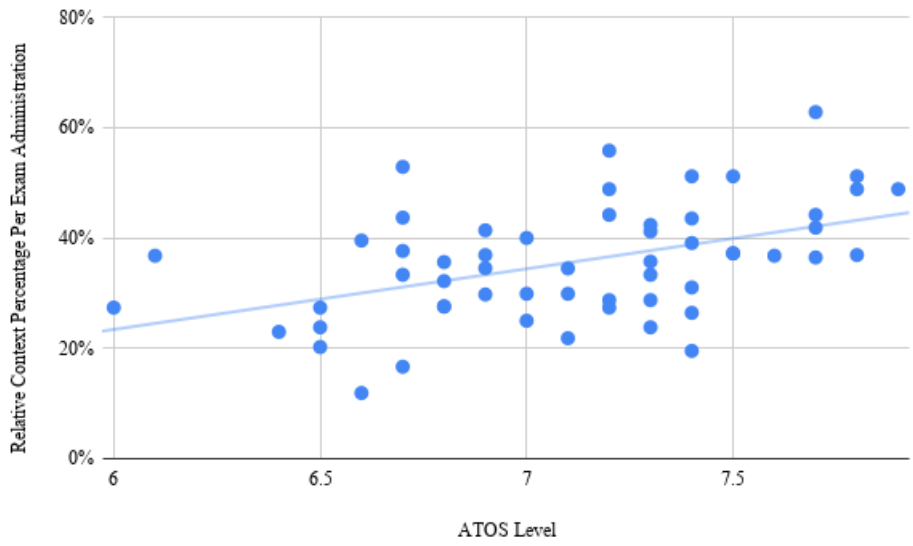


Figure 17. Scatterplot of ATOS Readability Scores versus Relative Context Percentage for *old*, *recent*, and *current* exams.

Next, the same consideration was made using more granularity, where the unit of analysis was an individual exam item. Each item on each exam was assigned an ATOS score and identified as having the presence or absence of context. When a Spearman's rho was calculated, there was a significant weak positive relationship between the ATOS score and context presence ($r_s(2139) = 0.039$, $p < 0.05$).

Research Question 4: Cognitive Complexity for Exam Versions and Standards

A cognitive complexity level (i.e., 0, 1, 2, 3) was assigned for each exam item as well as for each item part, for multi-part exam items. Levels 0 and 1 were labeled as "low complexity" items while levels 2 and 3 were labeled as "high complexity" items. Cognitive complexity was determined on all exam administrations from June 2002 to August 2018, and a relative complexity percentage for each exam administration was calculated. Since the unit of analysis was an exam, the percentage of items with high cognitive complexity was determined for all 20 *old*, 22 *recent*, and 14 *current* exams.

By exam version.

Overall, there was no significant change in the percentage of items with high cognitive complexity (i.e., those with a rating of 2 or 3); there was a decrease from *old* (78.8%) to *recent* (70.3%) and an increase from *recent* to *current* (77.8%). However, there was an increase in the interquartile range for the percentage of items with high relative cognitive percentages: *old* (IQR = 76.3-84.9), *recent* (IQR = 60.8-75.6), and *current* (IQR = 65.8-81.6) as can be noted in Figure 18 below. By further statistical testing, there was a significant different in high relative complexity items between *old* and *recent* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in high relative complexity among the different exam versions, $\chi^2(2) = 11.219$, $p = 0.004$ with a mean rank score of 37.10 for *old* exams, 20.25 for *recent* exams, and 29.18 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *recent* exams ($z = 16.850$, $p = 0.002$) (adjusted Bonferroni).

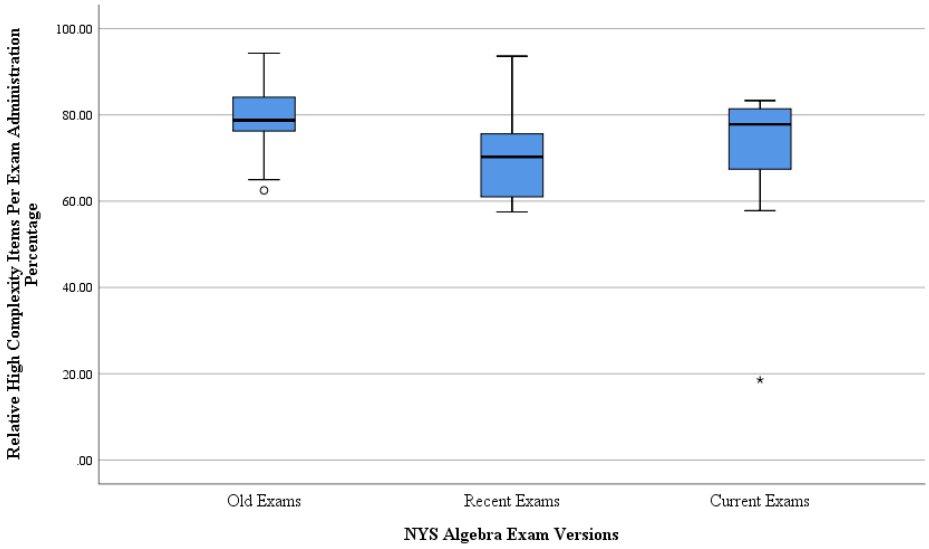


Figure 18. Box and whisker plots of percentage of items with high Cognitive Complexity for *old*, *recent*, and *current* exams.

Correspondingly, there was no significant change in the percentage of items with low cognitive complexity (those with a rating of 0 or 1) with an increase from *old* (21.3%) to *recent* (29.8%) and a decrease from *recent* to *current* (22.2%). However, there was an increase in the interquartile range for the percentage of items with low cognitive complexity with *old* (IQR = 15.1-23.8), *recent* (IQR = 24.4-39.3), and *current* (IQR = 18.4-34.3) as noted in figure 19 below. Through further testing, a statistically significant difference was found for low relative cognitive complexity items between the *old* and *recent* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in high relative complexity among the different exam versions, $\chi^2(2) = 11.219$, $p = 0.004$ with a mean rank score of 37.10 for *old* exams, 20.25 for *recent* exams, and 29.18 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *recent* exams ($z = 16.850$, $p = 0.002$) (adjusted Bonferroni).

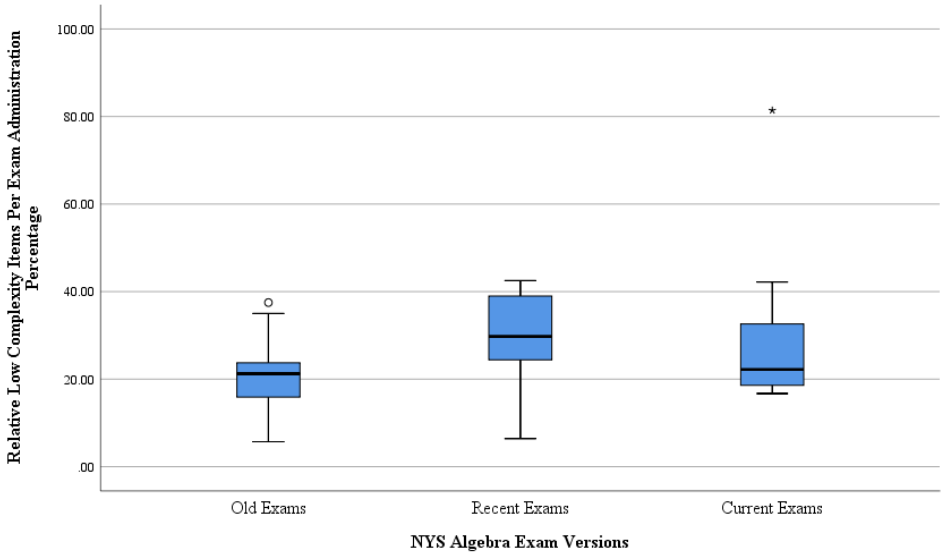


Figure 19. Box and whisker plots of percentage of items with low complexity for *old*, *recent*, and *current* exams.

By item type.

A score for cognitive complexity was assigned for each exam item part to determine trends in complexity for constructed response and multiple choice item parts. Since the unit of analysis was each exam item part, the relative cognitive complexity percentage was calculated for item parts: *old* constructed response (n = 221), *recent* constructed response (n = 276), and *current* constructed response (n = 288) as well as *old* multiple choice (n = 560), *recent* multiple choice (n = 653), and *current* multiple choice (n = 336).

There were higher percentages of high cognitive complexity on *old* exams for both constructed response (90%) and multiple choice (74.5%) item parts than on *current* exams constructed response (75%) and multiple choice (67.3%) item parts respectively. Through further testing, a statistically significant difference occurred between *old* and *current* exams for both constructed response and multiple choice items. A Kruskal-Wallis H test showed that there was a statistically significant difference in high relative complexity for constructed response item parts

among the different exam versions, $\chi^2(2) = 7.280$, $p = 0.026$ with a mean rank score of 36.03 for *old* exams, 26.02 for *recent* exams, and 21.64 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = 14.382$, $p = 0.033$) (adjusted Bonferroni). A Kruskal-Wallis H test showed that there was a statistically significant difference in high relative complexity for multiple choice items among the different exam versions, $\chi^2(2) = 8.855$, $p = 0.012$ with a mean rank score of 36.15 for *old* exams, 21.20 for *recent* exams, and 29.04 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = 14.945$, $p = 0.009$) (adjusted Bonferroni).

There were higher percentages of low cognitive complexity on *current* exams for constructed response (25%) and multiple choice (32.7%) item parts than on *old* exams constructed response (10%) and multiple choice (25.5%) item parts respectively. Through further testing, there was a statistically significant difference between *old* and *current* constructed response and multiple choice items. A Kruskal-Wallis H test showed that there was a statistically significant difference in high relative complexity for constructed response item parts among the different exam versions, $\chi^2(2) = 7.280$, $p = 0.026$ with a mean rank score of 36.03 for *old* exams, 26.02 for *recent* exams, and 21.64 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = 14.382$, $p = 0.033$) (adjusted Bonferroni). A Kruskal-Wallis H test showed that there was a statistically significant difference in high relative complexity for multiple choice items among the different exam versions, $\chi^2(2) = 8.855$, $p = 0.012$ with a mean rank score of 36.15 for *old* exams, 21.20 for *recent* exams, and 29.04 for *current* exams. When running a Dunn-Bonferroni test, there was a

statistically significant difference between the *old* and *current* exams (z = 14.945, p = 0.009) (adjusted Bonferroni). Table 15, Figure 20, and Figure 21 illustrate this.

Table 15

Cognitive Complexity by Item Part Type (Constructed Response, Multiple Choice) for Old, Recent, *and* Current *Exams*

Item Type	Cognitive Complexity	% of Item Parts		
		Old	*Recent*	*Current*
Constructed Response	Low	10.0	17.8	25.0
	High	90.0	82.2	75.0
Multiple Choice	Low	25.5	35.4	32.7
	High	74.5	64.6	67.3

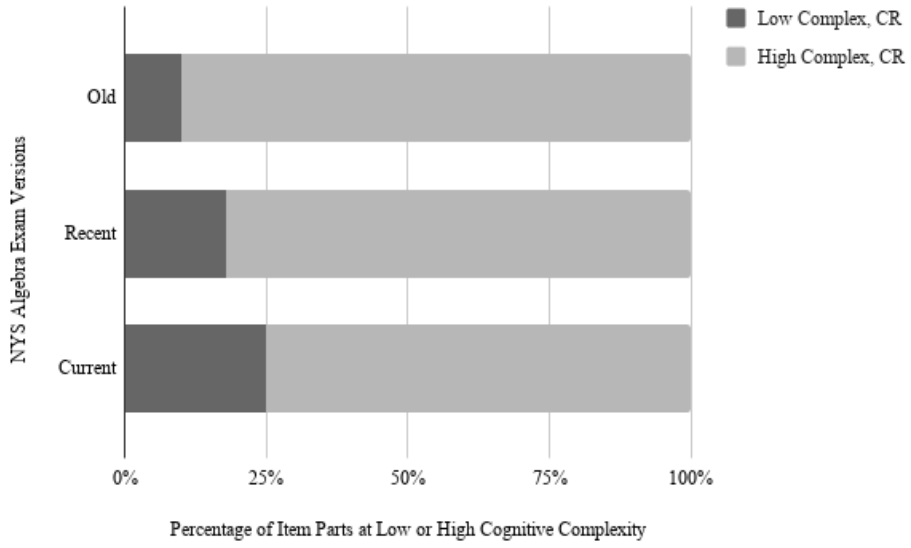


Figure 20. Stacked bar graph of cognitive complexity (for constructed response items) for *old*, *recent*, and *current* exams.

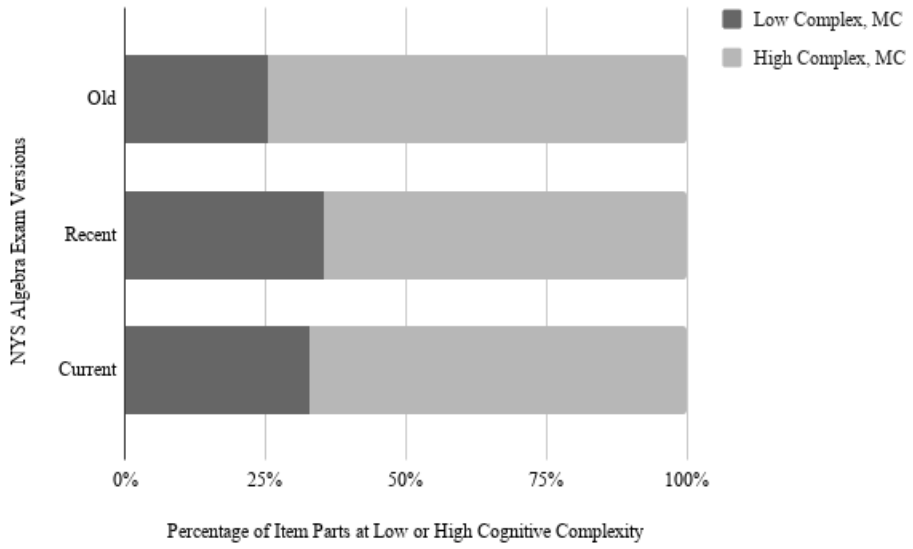


Figure 21. Stacked bar graph of cognitive complexity (for multiple choice items) for *old*, *recent*, and *current* exams.

Research Question 5: Relationships among Readability, Context, and Cognitive Complexity

To determine if there was a relationship between the readability of each exam administration, the relative "real world" context, and cognitive complexity, Spearman's rho was calculated (Spearman, 1904). Using Spearman's rho, a significant moderate positive correlation between cognitive complexity and context ($r_s(2139) = 0.303$, $p < 0.01$) was determined (Corder & Foreman, 2009). A significant very weak positive correlation between cognitive complexity and ATOS level readability ($r_s(2139) = 0.101$, $p < 0.01$) was also determined. Taken together this suggests that item parts that involved real world context had slightly higher levels of cognitive complexity and a slightly higher ATOS reading level.

Research Question 6: Type and Number of Representations for Exam Versions

For the type and number of representations, each item part was assigned one or more applicable representations. Since each item part has a stem and corresponding response, both were used as units of analysis. Representations were assigned for 781 *old* item part stems and 781 *old* item part responses, 929 *recent* item part stems and 929 *recent* item part responses, and 624 *current* item part stems and 624 *current* item part responses. To determine exam trends, a representation percentage based on the total number of item parts was determined for the stem and response of each item part. This was determined for the stem and response of each item part that appears on all 20 *old*, 22 *recent*, and 14 *current* exams. To determine item type trends, a relative representation percentage was determined based on the constructed response and multiple choice response of these items to see if a certain type of item part showed more of an increase on a constructed response item versus a multiple choice item. For the constructed response and multiple choice item parts, the unit of analysis was item parts: *old* constructed response (n = 221), *recent* constructed response (n = 276), and *current* constructed response (n = 288) as well as *old* multiple choice (n = 560), *recent* multiple choice (n = 653), and *current* multiple choice (n = 336).

Number of representations by exam version.

Each representation had two levels of categorization: a representational category of visual, symbolic, numeric, or written, and a more specific representational subcategory (for the visual, symbolic and numeric categories). For instance, a graph would be labeled with visual-graph (V-G), while the equation for the graph would be labeled with symbolic-algebraic (S-A).

For example, on the August 2018 *current* exam, item 26, displayed below in Figure 22 and Figure 23, was a constructed response item with two item stems. The responses were noted

in Table 16 and Table 17. To measure the two stems and the two responses, the expected response was taken from the Model Response Set on the New York State Education Department Regents website. Using the representations framework, for 26a, visual-graph (V-G), numeric table (N-T), symbolic set (S-S), and symbolic-algebraic (S-A) were determined for both the stem and response. The students needed to understand these representations in order to correctly answer 26a. For 26a stem, the written code (W) was determined. The written code (W) was not determined for the response of 26a as the student could have circled the previous representations as an acceptable response for this item part. For 26b, visual-graph (V-G), numeric table (N-T), symbolic-set (S-S), symbolic-algebraic (S-A), and written (W) were determined for both stem and response.

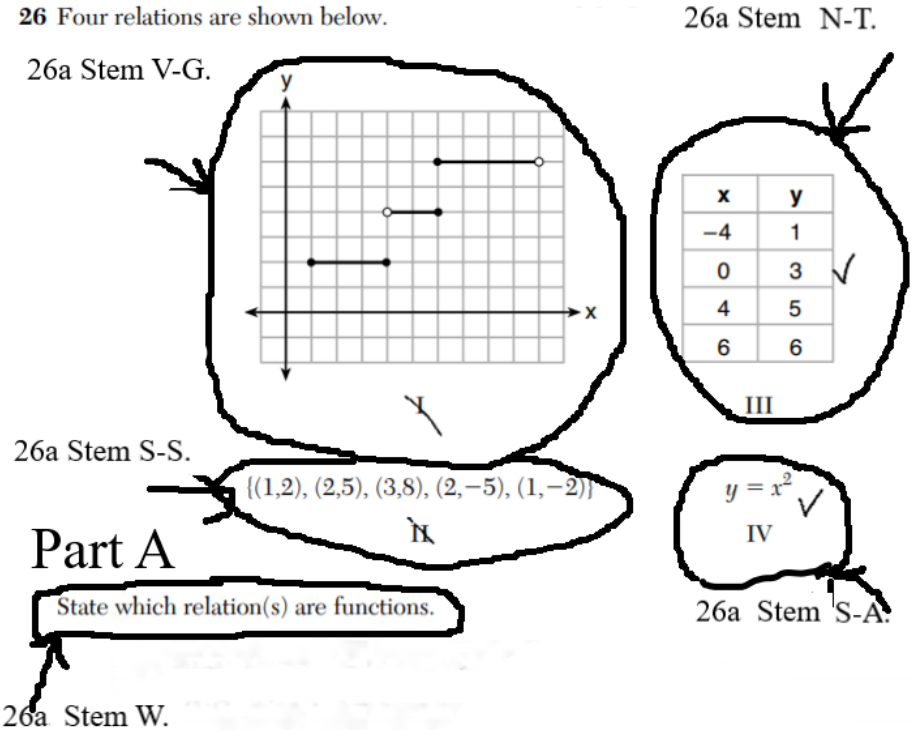
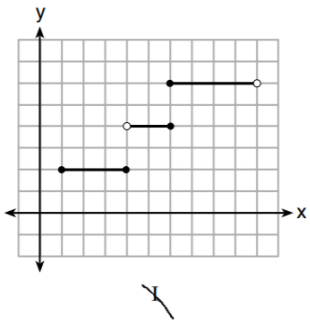
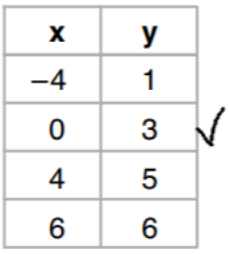


Figure 22. Example of the labeling of representations on the *current* August 2018 exam, item part 26a.

Table 16

Current *August 2018, item part 26a responses*

Coding	Response
V-G.	 I There is an indication that this is not a function.
S-S.	$\{(1,2), (2,5), (3,8), (2,-5), (1,-2)\}$ II There is an indication that this is not a function.
N-T.	 III There is an indication here that it is a function along with a marking of this number below in the response area.
S-A.	$y = x^2$ ✓ IV There is an indication here that it is a function along with a marking of this number below in the response area.

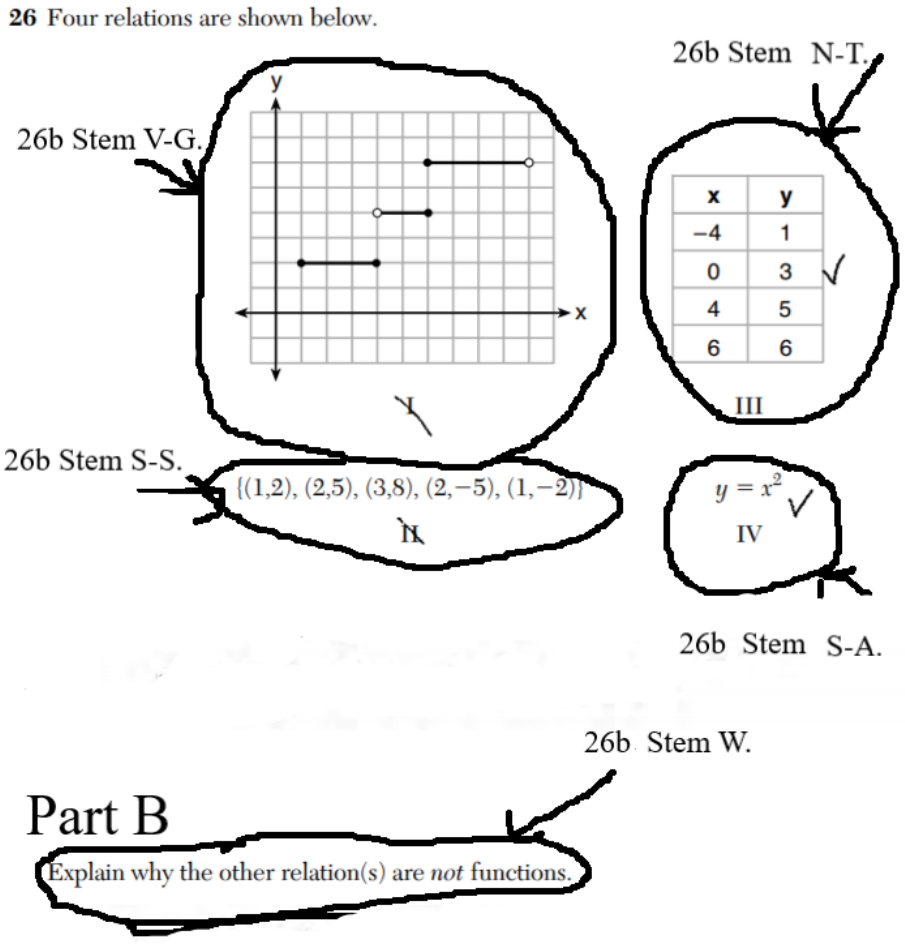
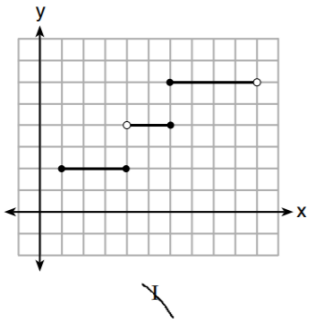
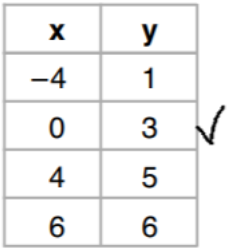


Figure 23. Example of the labeling of representations on the *current* August 2018 exam, item part 26b.

Table 17

Current *August 2018, item part 26b responses*

Coding	Response
V-G.	 I There is an indication that this is not a function.
S-S.	$\{(1,2), (2,5), (3,8), (2,-5), (1,-2)\}$ II There is an indication that this is not a function.
N-T.	 ✓ III There is an indication here that it is a function.
S-A.	$y = x^2$ ✓ IV There is an indication here that it is a function.
W.	I does not pass the vertical line test and II has two outputs for the input 2. This is a written description of why I and II are not functions.

After counting the number of representational subcategories identified for each item part, *old* exams had fewer representational subcategories overall per item part than *current* exams. Table 18 described the number of representations in subcategories based on each exam (*old*, *recent*, and *current*).

Table 18

Number of Representational Subcategories Present in Item Parts Type for Old, Recent, *and* Current Exams

Number of Representational Subcategories in Item Parts	Multiple Choice Items						Constructed Response Items					
	Old		*Recent*		*Current*		*Old*		*Recent*		*Current*	
	#	%	#	%	#	%	#	%	#	%	#	%
2	226	41	227	35	73	22	27	12	41	15	29	10
3	265	48	335	51	163	49	33	15	121	44	115	40
4	57	10	78	12	77	23	75	33	75	27	93	32
5	6	1	13	2	17	5	68	30	28	10	39	14
6	-	-	-	-	5	1	14	6	7	3	9	3
7	-	-	-	-	-	-	6	3	4	1	1	-
8	-	-	-	-	1	-	4	2	-	-	-	-
9	-	-	-	-	-	-	-	-	-	-	1	-
10	-	-	-	-	-	-	-	-	-	-	1	-

When looking more specifically at item stems and item responses, there has been a decrease in the number of representational subcategories in both item stems and responses for constructed response items. However, there has been an increase in the number of representational subcategories for both item stems and responses for multiple choice items. Table 19 illustrates the average number of representational subcategories and how they changed in both constructed response and multiple choice items in *old*, *recent*, and *current* exams.

Table 19

Average Number of Representational Subcategories Present in Item Stems and Responses by Item Parts Type for Old, Recent, *and* Current *Exams*

	Constructed Response Items			Multiple Choice Items		
	Old	*Recent*	*Current*	*Old*	*Recent*	*Current*
Item Stems	2.1	1.9	2.1	1.7	1.8	2.0
Item Responses	2.1	1.5	1.5	1.0	1.1	1.2

Types of representations by exam version.

After determining the representations for each item part, the number of representational subcategories for each stem and response were determined. For example, on the August 2018 *current* exam, there were seven stems of the item parts that had visual-graphs (V-G). There were 45 item parts in total. Therefore, 16% of the item part stems on the August 2018 exam had visual-graphs and were coded V-G.

After assigning applicable representations to each stem and response, there were four representational subcategories that were significant through all versions of the NYS algebra exams (*old*, *recent*, and *current*): visual-graph (V-G), symbolic-algebraic (S-A), numeric-numeral (N-N), and written (W). Since the unit of analysis was an exam, the relative representation percentage was determined for all 20 *old*, 22 *recent*, and 14 *current* exams. For each exam administration, the number of the type of representational subcategories was counted for both the stem and response of each item part and a relative representational percentage was calculated for each representation. Table 20 illustrates the four prevalent subcategories along with the rest of the subcategories.

Table 20

Prominent Representations in Stems and Responses on Old, Recent, *and* Current *Exams*

	Medians of Stems			Medians of Responses		
	Old	*Recent*	*Current*	*Old*	*Recent*	*Current*
Visual						
Graph	4.0	7.0	14.0	3.0	7.0	9.5
Chart	3.0	2.0	-	-	-	-
Geometric Figure	10.0	7.0	-	3.0	-	-
Diagram	-	2.5	1.0	-	-	-
Picture	-	-	-	-	-	-
Symbolic						
Algebraic	36.5	38.0	62.0	32.0	29.0	47.5
Sets	-	2.5	-	1.0	5.0	-
Intervals	-	-	2.0	-	2.0	4.5
Probability Statistics	-	-	-	3.0	-	-
Geometric Notation	9.0	2.5	-	3.0	2.0	-
Sequences and Series	-	-	2.0	-	-	2.0
Numeric						
Numeral	35.0	32.5	30.0	53.5	49.0	33.5
Point	5.0	2.0	1.0	5.0	5.0	4.5
Table and Matrix	-	3.5	9.0	-	-	2.0
Written	75.5	77.0	80.0	22.5	14.5	33.0

Visual-graph (V-G) representations by exam and item type.

Since the unit of analysis was an exam, the relative visual-graph percentage was determined for stems on all 20 *old*, 22 *recent*, and 14 *current* exams. Overall, for each exam, the medians of relative visual-graph percentage for stems increased from *old* (4%) to *recent* (7%) and again to *current* (14%). Through further testing, statistically significant differences were found between *old* and *recent*, *old* and *current*, as well as *recent* and *current* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in visual-graph stems among the different exam versions, $\chi^2(2) = 27.730$, $p < 0.001$ with a mean rank score of 15.65 for *old* exams, 29.48 for *recent* exams, and 45.32 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *recent* exams

(z = -13.827, p = 0.017) (adjusted Bonferroni). There was a statistically significant difference between the *old* and *current* exams (z = -29.671, p < 0.001) (adjusted Bonferroni). There was a statistically significant difference between the *recent* and *current* exams (z = -15.844, p = 0.013) (adjusted Bonferroni). Figure 24 displays the increase in medians from *old* to *current* exams for visual-graph stems.

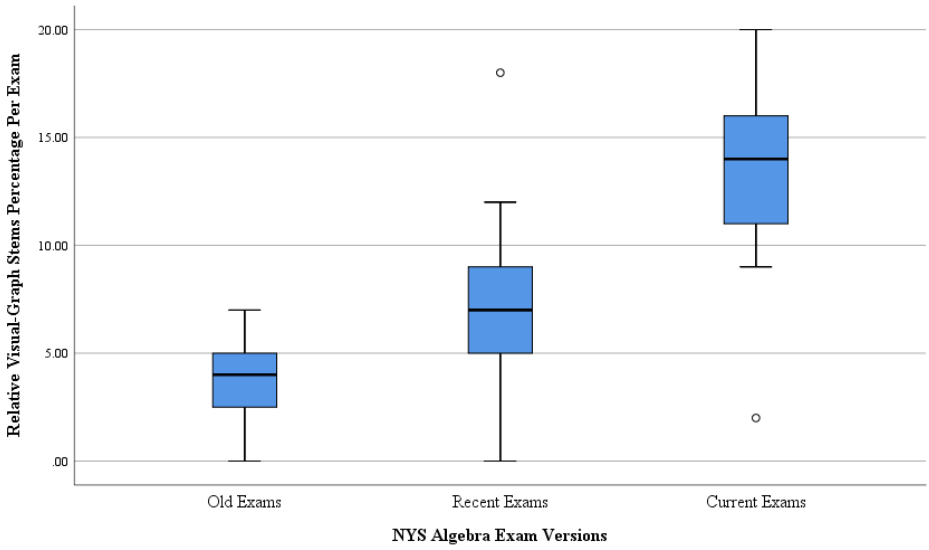


Figure 24. Box and whisker plots of relative visual-graph percentage for stems on *old*, *recent*, and *current* exams.

Since the unit of analysis was an exam, the relative representation percentage was determined for all responses on 20 *old*, 22 *recent*, and 14 *current* exams. Overall, for each exam, the medians of relative visual-graph percentage for responses increased from *old* (3%) to *recent* (7%) and again to *current* (9.5%) exams. Through further testing, there were statistically significant differences between *old* and *recent* as well as *old* and *current* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in visual-graph responses among the different exam versions, $\chi^2(2) = 25.601$, $p < 0.001$ with a mean rank score

of 15.53 for *old* exams, 30.57 for *recent* exams, and 43.79 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *recent* exams (z = -15.043, p = 0.008) (adjusted Bonferroni). There was a statistically significant difference between the *old* and *current* exams (z = -28.261, p < 0.001) (adjusted Bonferroni). Figure 25 illustrates the increase in medians for visual-graph responses.

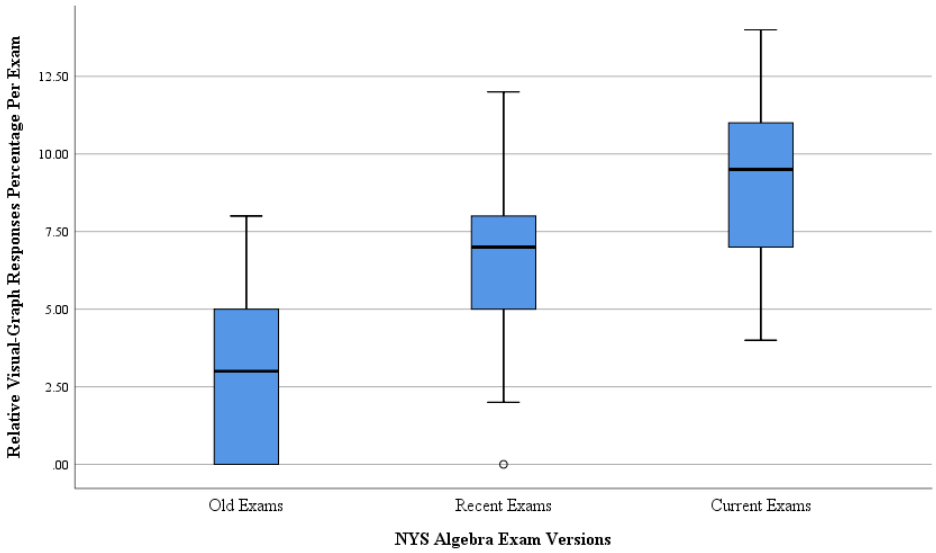


Figure 25. Box and whisker plots of relative visual-graph percentage for responses on *old*, *recent*, and *current* exams.

These trends were also apparent when item parts were compared within the categories of constructed response and multiple choice item parts: *old* constructed response (n = 221), *recent* constructed response (n = 276), and *current* constructed response (n = 288) as well as *old* multiple choice (n = 560), *recent* multiple choice (n = 653), and *current* multiple choice (n = 336). After calculating the relative visual-graph percentage for constructed response and multiple choice, there was an increase from *old* to *recent* to *current* exams of visual-graph (V-G) representations when considering multiple choice item parts and constructed response item parts for both stems and responses.

For constructed response and multiple choice stems, there was a statistically significant difference among *old* and *current* as well as *old* and *recent* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in visual-graph constructed response stems among the different exam versions, $\chi^2(2) = 20.994$, $p < 0.001$ with a mean rank score of 18.70 for *old* exams, 27.70 for *recent* exams, and 43.75 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = -25.050$, $p < 0.001$) (adjusted Bonferroni). There was a statistically significant difference between the *recent* and *current* exams ($z = -16.045$, $p = 0.009$) (adjusted Bonferroni).

For constructed response and multiple choice responses, there was a statistically significant difference among *old* and *current* exams. For multiple choice responses only, there was a statistically significant difference among *old* and *recent* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in visual-graph constructed response responses among the different exam versions, $\chi^2(2) = 6.595$, $p = 0.037$ with a mean rank score of 23.33 for *old* exams, 27.36 for *recent* exams, and 37.68 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = -14.354$, $p = 0.034$) (adjusted Bonferroni).

For visual-graph multiple choice responses, a Kruskal-Wallis H test showed that there was a statistically significant difference among the different exam versions, $\chi^2(2) = 20.539$, $p < 0.001$ with a mean rank score of 16.28 for *old* exams, 33.39 for *recent* exams, and 38.29 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *recent* exams ($z = -17.111$, $p = 0.001$) (adjusted Bonferroni). There was a statistically significant difference between the *old* and *current* exams ($z = -22.011$, $p < 0.001$) (adjusted Bonferroni). Table 21 overall illustrates the median relationships.

Table 21

Relative Visual-Graph Percentage for Stems and Responses by Item Type (Constructed Response, Multiple Choice) for Old, Recent, *and* Current *Exams*

Item Stems and Responses with Visual-Graphs by Item Type	Percentage Median		
	Old	*Recent*	*Current*
Stem			
Constructed Response	-	7.7	18.2
Multiple Choice	3.3	6.7	12.5
Total Stem (Weighted)	4.0	7.0	14.0
Response			
Constructed Response	10.0	10.0	14.3
Multiple Choice	-	3.3	8.3
Total Response (Weighted)	3.0	7.0	9.5

Symbolic-algebraic representations (S-A) by exam and item type.

Since the unit of analysis was an exam, the relative symbolic-algebraic percentage was determined for all stems on 20 *old*, 22 *recent*, and 14 *current* exams. Overall, for each exam, the medians of relative percentage of stems with symbolic-algebra (S-A) representations increased from *old* (36.5%) to *recent* (38%) and again to *current* (62%) exams. Further statistical testing indicated that there were significant differences between *old* and *current* as well as *recent* and *current* exams for symbolic-algebraic stems. A Kruskal-Wallis H test showed that there was a statistically significant difference in symbolic-algebraic stems among the different exam versions, $\chi^2(2) = 33.146$, $p < 0.001$ with a mean rank score of 17.68 for *old* exams, 24.98 for *recent* exams, and 49.50 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = -31.825$, $p < 0.001$)

(adjusted Bonferroni). There was a statistically significant difference between the *recent* and *current* exams (z = -24.523, p < 0.001) (adjusted Bonferroni).This increase in symbolic-algebra (S-A) representations is further illustrated in Figure 26.

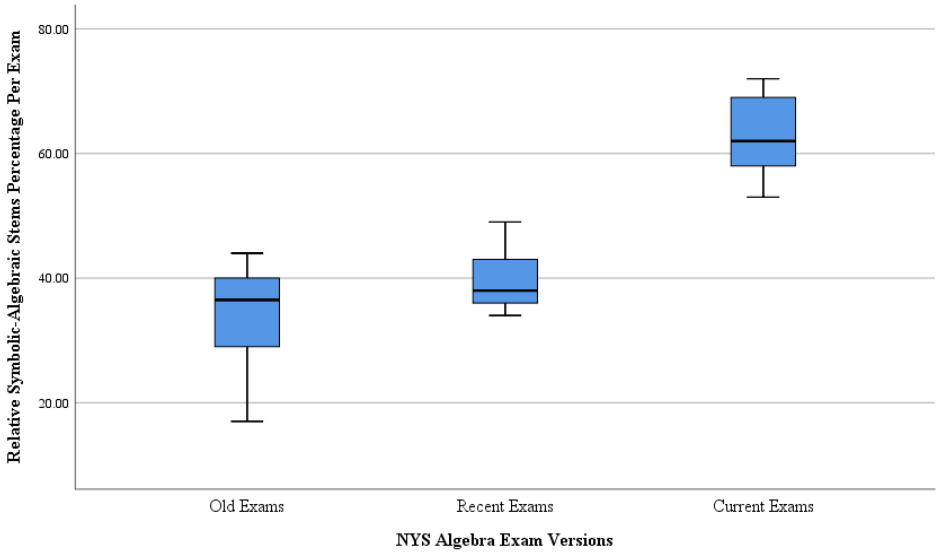


Figure 26. Box and whisker plots of relative symbolic-algebraic percentages for stems on *old*, *recent*, and *current* exams.

Since the unit of analysis was an exam, the relative symbolic-algebraic percentage was determined for all responses on 20 *old*, 22 *recent*, and 14 *current* exams. Overall, for each exam, the medians of relative percentage of responses with symbolic-algebraic representations decreased, and then increased from *old* (32%) to *recent* (29%) and again to *current* (37.5%) exams. Further statistical testing indicated that there were significant differences between *old* and *recent* as well as *recent* and *current* exams for symbolic-algebraic stems. A Kruskal-Wallis H test showed that there was a statistically significant difference in symbolic-algebraic responses among the different exam versions, $\chi^2(2) = 28.645$, $p < 0.001$ with a mean rank score of 24.38 for *old* exams, 19.61 for *recent* exams, and 48.36 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *recent* and *current* exams (z = -28.744, p < 0.001) (adjusted Bonferroni). There was a statistically significant

difference between the *old* and *recent* exams (z = -23.982, p < 0.001) (adjusted Bonferroni). This decrease and then increase in symbolic-algebraic responses is illustrated in figure 27.

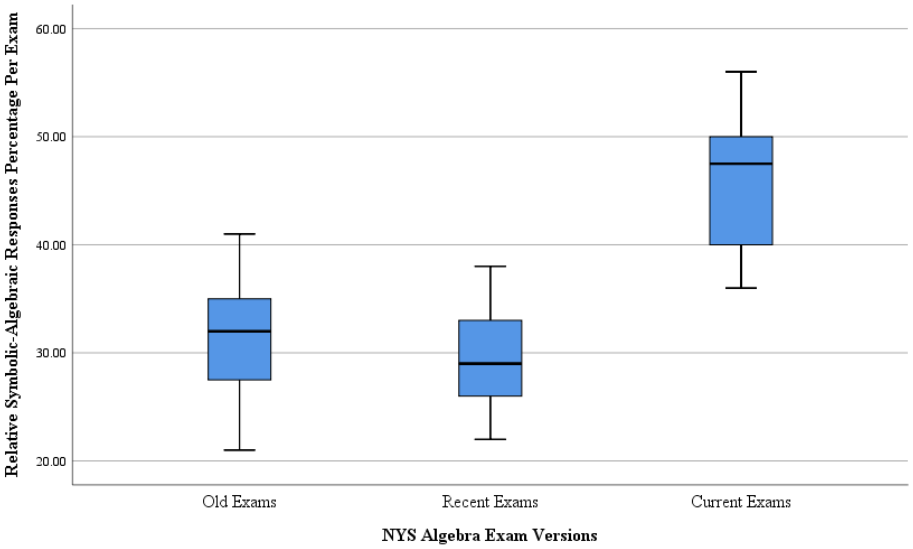


Figure 27. Box and whisker plots of relative symbolic-algebraic percentages for responses on *old*, *recent*, and *current* exams.

These trends were also apparent when item parts were compared within the categories of constructed response and multiple choice item parts: *old* constructed response (n = 221), *recent* constructed response (n = 276), and *current* constructed response (n = 288) as well as *old* multiple choice (n = 560), *recent* multiple choice (n = 653), and *current* multiple choice (n = 336). After calculating the relative symbolic-algebraic percentage for constructed response and multiple choice, this is indicative of an increase from *old* to *current* exams of symbolic-algebraic representations when considering constructed response and multiple choice item parts.

For constructed response stems, there was a statistically significant difference among *old* and *current* exams as well as *old* and *recent* exams. For multiple choice stems only, there was a statistically significant difference among *recent* and *current* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in symbolic-algebraic constructed response stems among the different exam versions, $\chi^2(2) = 29.814$, $p < 0.001$ with a mean rank

score of 19.98 for *old* exams, 23.25 for *recent* exams, and 48.93 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = -28.954$, $p < 0.001$) (adjusted Bonferroni). There was a statistically significant difference between the *old* and *recent* exams ($z = -25.679$, $p < 0.001$) (adjusted Bonferroni).

For symbolic-algebraic multiple choice stems, a Kruskal-Wallis H test showed that there was a statistically significant difference in these stems among the different exam versions, $\chi^2(2) = 27.762$, $p < 0.001$ with a mean rank score of 19.23 for *old* exams, 24.61 for *recent* exams, and 47.86 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the old and *current* exams ($z = -28.632$, $p < 0.001$) (adjusted Bonferroni). There was a statistically significant difference between the *recent* and *current* exams ($z = -23.244$, $p < 0.001$) (adjusted Bonferroni).

For constructed response and multiple choice responses, there was a statistically significant difference among *recent* and *current* exams. For constructed response responses only, there was a statistically significant difference among *old* and *recent* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in symbolic-algebraic constructed response responses among the different exam versions, $\chi^2(2) = 32.510$, $p < 0.001$ with a mean rank score of 40.00 for *old* exams, 13.18 for *recent* exams, and 36.14 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *recent* and *current* exams ($z = -22.961$, $p < 0.001$) (adjusted Bonferroni). There was a statistically significant difference between the *old* and *recent* exams ($z = 26.818$, $p < 0.001$) (adjusted Bonferroni).

For symbolic-algebraic multiple choice responses, a Kruskal-Wallis H test showed that there was a statistically significant difference among the different exam versions, $\chi^2(2) = 31.219$, $p < 0.001$ with a mean rank score of 14.58 for *old* exams, 30.02 for *recent* exams, and 46.00 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the old and recent exams ($z = -15.448$, $p = 0.006$) (adjusted Bonferroni). There was a statistically significant difference between the *old* and *current* exams ($z = -31.425$, $p < 0.001$) (adjusted Bonferroni). There was a statistically significant difference between the *recent* and *current* exams ($z = -15.977$, $p = 0.012$) (adjusted Bonferroni). These differences in medians are illustrated in Table 22.

Table 22

Relative Symbolic-Algebraic Percentage for Stems and Responses by Item Type (Constructed Response, Multiple Choice) for Old, Recent, *and* Current *Exams*

Item Stems and Responses with Symbolic-Algebraic Representations by Item Type	Percentage Median		
	Old	*Recent*	*Current*
Stem			
Constructed Response	21.1	28.7	64.1
Multiple Choice	38.4	43.4	64.6
Total Stem (Weighted)	36.5	38.0	62.0
Response			
Constructed Response	60.0	23.3	57.1
Multiple Choice	20.0	30.0	37.5
Total Response (Weighted)	32.0	29.0	47.5

Numeric-Numeral representations (N-N) by exam and item type.

Overall, for each exam, the medians of the relative percentage of stems with numeric-numerals representations slightly decreased from *old* (35%) to *recent* (32.5%) and again to *current* (30%) exams. Through further statistical testing, there was a statistically significant difference between *old* and *current* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in numeric-numeral stems among the different exam versions, $\chi^2(2) = 6.068$, $p = 0.048$ with a mean rank score of 34.88 for *old* exams, 27.41 for *recent* exams, and 21.11 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = 13.768$, $p = 0.045$) (adjusted Bonferroni). Figure 28 illustrates the decrease in medians for numeric-numeral representations in stems.

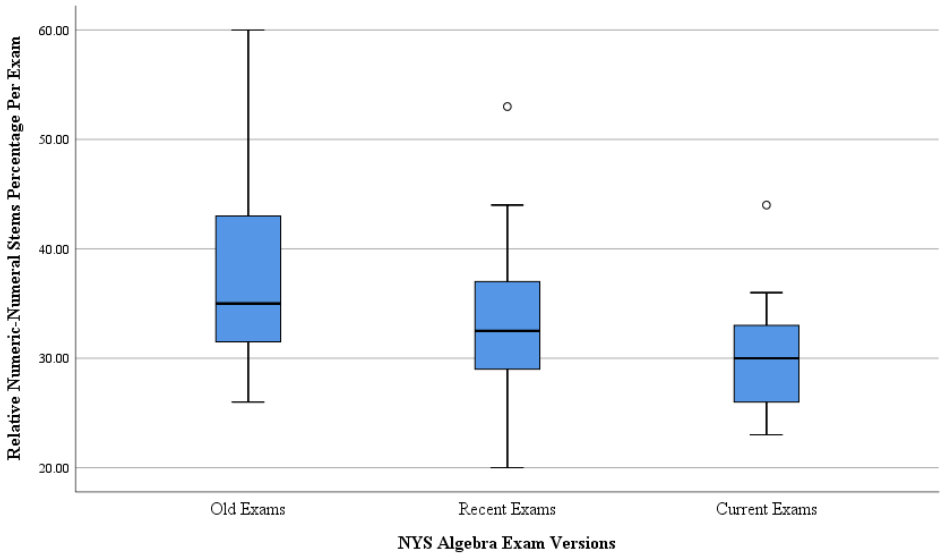


Figure 28. Box and whisker plots of relative numeric-numeral percentages for stems on *old*, *recent*, and *current* exams.

Overall, for each exam, the medians of relative percentage of item responses with numeric-numeral representations decreased from *old* (53.5%) to *recent* (49%) and again to *current* (33.5%) exams. Through further statistical testing, there was a statistically significant

difference between *old* and *current* exams as well as *recent* and *current* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in numeric-numeral responses among the different exam versions, $\chi^2(2) = 28.961$, $p < 0.001$ with a mean rank score of 40.40 for *old* exams, 29.50 for *recent* exams, and 9.93 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *recent* and *current* exams ($z = 19.571$, $p = 0.001$) (adjusted Bonferroni). There was a statistically significant difference between the *old* and *current* exams ($z = 30.471$, $p < 0.001$) (adjusted Bonferroni). Figure 29 illustrates the decrease in medians for numeric-numeral representations in responses.

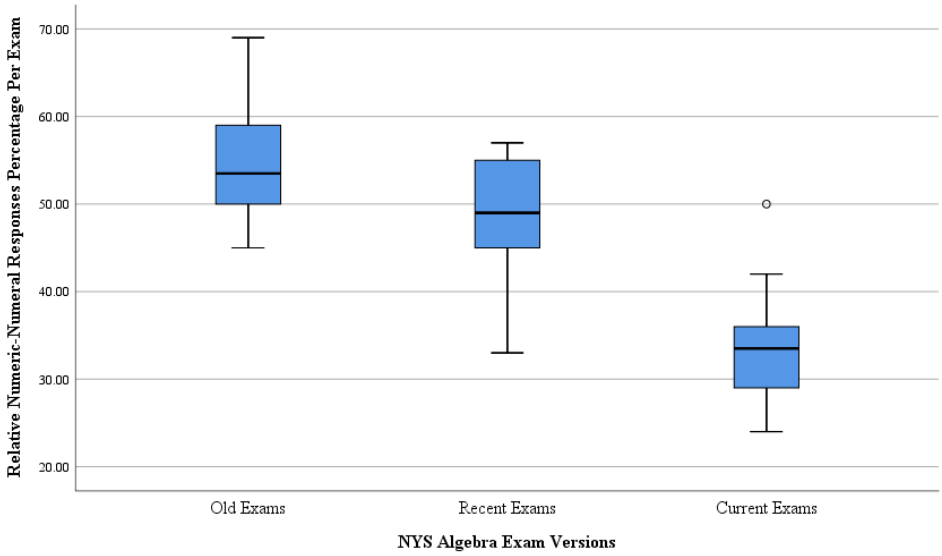


Figure 29. Box and whisker plots of relative numeric-numeral percentages for responses on *old*, *recent*, and *current* exams.

These trends were also apparent when item parts were compared for both item types (i.e., constructed response and multiple choice items). This is indicative of a decrease from *old* to *recent* to *current* exams of numeric-numeral representations when considering item parts. For constructed response stems, there was a statistically significant difference among *old* and *current* exams. For multiple choice items, there were no statistically significant differences among exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in

numeric-numeral constructed response stems among the different exam versions, $\chi^2(2) = 14.077$, $p = 0.001$ with a mean rank score of 37.55 for *old* exams, 28.05 for *recent* exams, and 16.29 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = 21.264$, $p = 0.001$) (adjusted Bonferroni). However, a Kruskal-Wallis H test showed that there was no statistically significant difference in numeric-numeral multiple choice stems among the different exam versions, $\chi^2(2) = 0.040$, $p = 0.980$ with a mean rank score of 28.58 for *old* exams, 28.89 for *recent* exams, and 27.79 for *current* exams.

For constructed response and multiple choice responses, there was a statistically significant difference between *old* and *current* exams. For multiple choice responses only, there was a statistically significant difference between *recent* and *current* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in numeric-numeral constructed response responses among the different exam versions, $\chi^2(2) = 29.170$, $p < 0.001$ with a mean rank score of 38.48 for *old* exams, 32.02 for *recent* exams, and 8.71 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = 21.264$, $p = 0.001$) (adjusted Bonferroni).

For numeric-numeral multiple choice responses, a Kruskal-Wallis H test showed that there was a statistically significant difference among the different exam versions, $\chi^2(2) = 20.312$, $p < 0.001$ with a mean rank score of 38.60 for *old* exams, 29.11 for *recent* exams, and 13.11 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *current* exams ($z = 25.493$, $p < 0.001$) (adjusted Bonferroni). There was a statistically significant difference between the *recent* and *current* exams ($z = 16.006$, $p = 0.012$) (adjusted Bonferroni). Table 23 indicates these relationships.

Table 23

Relative Numeric-Numeral Percentage for Stems and Responses by Item Type (Constructed Response, Multiple Choice) for Old, Recent, *and* Current *Exams*

Item Stems and Responses with Numeric-Numeral Representations by Item Type	Percentage Median		
	Old	*Recent*	*Current*
Stem			
Constructed Response	60.0	44.0	32.6
Multiple Choice	26.7	28.4	29.2
Total Stem (Weighted)	35.0	32.5	30.0
Response			
Constructed Response	70.0	63.6	34.2
Multiple Choice	48.4	41.7	33.3
Total Response (Weighted)	53.5	49.0	33.5

Written representations by exam and item type.

Overall, for each exam, the medians of relative percentage of written representation stems slightly increased from *old* (75.5%) to *recent* (77%), and again to *current* (80%) exams. However, further statistical testing did not find this increase significant. A Kruskal-Wallis H test showed that there was no statistically significant difference in written stems among the different exam versions, $\chi^2(2) = 1.277$, $p = 0.528$ with a mean rank score of 26.88 for *old* exams, 27.27 for *recent* exams, and 32.75 for *current* exams. This slight increase in written representations is indicated on figure 30.

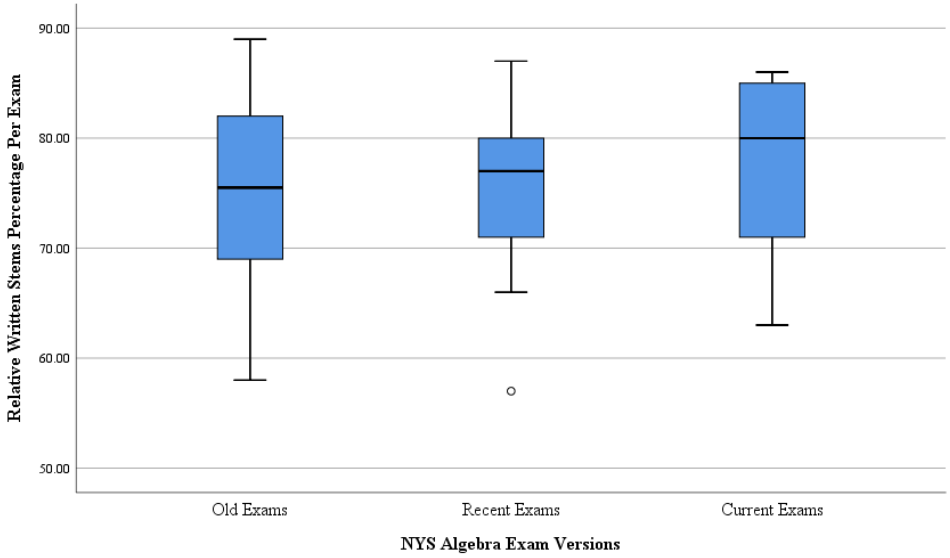


Figure 30. Box and whisker plots of relative percentages of written stems for *old*, *recent*, and *current* exams.

Overall, for each exam, the medians of relative percentage of responses with written representations dipped and rebounded from *old* (22.5%) to *recent* (14.5%) and then to *current* (33%) exams. Further statistical testing indicated that there was a significant difference between *old* and *recent* as well as *recent* and *current* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in written responses among the different exam versions, $\chi^2(2) = 32.012$, $p < 0.001$ with a mean rank score of 32.90 for *old* exams, 14.25 for *recent* exams, and 44.61 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *recent* exams ($z = 18.650$, $p = 0.001$) (adjusted Bonferroni). There was a statistically significant difference between the *recent* and *current* exams ($z = -30.357$, $p < 0.001$) (adjusted Bonferroni). This decrease and increase of written responses is illustrated in figure 31.

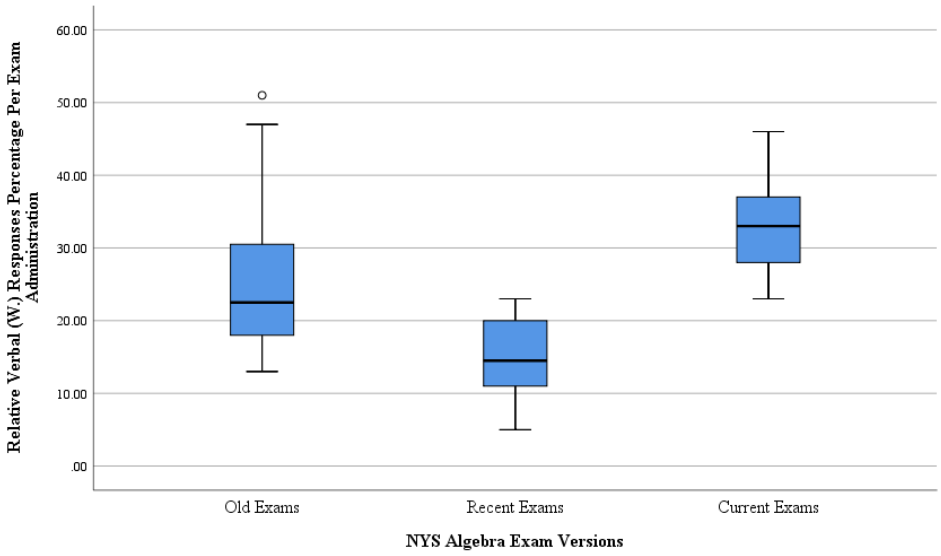


Figure 31. Box and whisker plots of relative percentages of written responses for *old*, *recent*, and *current* exams.

These trends were also apparent when item parts were compared within item types. This is indicative of a slight increase from *old* to *current* exams of written representations when considering both multiple choice and constructed response item parts. For constructed response and multiple choice item stems, by using further testing, there were no statistically significant differences. A Kruskal-Wallis H test showed that there was no statistically significant difference in written constructed response stems among the different exam versions, $\chi^2(2) = 0.420$, $p = 0.810$ with a mean rank score of 26.88 for *old* exams, 28.68 for *recent* exams, and 30.54 for *current* exams. Additionally, a Kruskal-Wallis H test showed that there was no statistically significant differences in written multiple choice stems among the different exam versions, $\chi^2(2) = 1.083$, $p = 0.582$ with a mean rank score of 28.28 for *old* exams, 26.39 for *recent* exams, and 32.14 for *current* exams.

For constructed response and multiple choice responses, there was a statistically significant difference between *recent* and *current* exams. For constructed response item responses only, there was a statistically significant difference between *old* and *recent* exams. A

Kruskal-Wallis H test showed that there was a statistically significant difference in written constructed response stems among the different exam versions, $\chi^2(2) = 18.142$, $p < 0.001$ with a mean rank score of 34.48 for *old* exams, 17.09 for *recent* exams, and 37.89 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *old* and *recent* exams ($z = 17.384$, $p = 0.002$) (adjusted Bonferroni). There was a statistically significant difference between the *recent* and *current* exams ($z = -20.802$, $p = 0.001$) (adjusted Bonferroni). A Kruskal-Wallis H test showed that there was a statistically significant difference in written multiple choice responses among the different exam versions, $\chi^2(2) = 20.293$, $p < 0.001$ with a mean rank score of 29.63 for *old* exams, 18.20 for *recent* exams, and 43.07 for *current* exams. When running a Dunn-Bonferroni test, there was a statistically significant difference between the *recent* and *current* exams ($z = -24.867$, $p < 0.001$) (adjusted Bonferroni). The differences in medians are noted in Table 24.

Table 24

Relative Written Percentage for Stems and Responses by Item Type (Constructed Response, Multiple Choice) for Old, Recent, *and* Current *Exams*

Item Stems and Responses with Written Representation by Item Type	Percentage Median		
	Old	*Recent*	*Current*
Stem			
Constructed Response	73.9	80.0	81.0
Multiple Choice	75.9	73.3	79.2
Total Stem (Weighted)	75.5	77.0	80.0
Response			
Constructed Response	33.3	12.6	39.9
Multiple Choice	18.4	13.3	25.0
Total Response (Weighted)	22.5	14.5	33.0

Research Question 7: Stem-Response Representation Relationships for Exam Versions and Standards

After determining the most prominent representational subcategories (visual-graph, symbolic-algebraic, numeric-numeral, and written), I decided to see if certain stem-response relationship combinations were more prevalent for the different exam versions. Since the unit of analysis was an exam, a relative representation relationship percentage was determined for all 20 *old*, 22 *recent*, and 14 *current* exams. A median was determined for each representation relationship.

For each item, there was a labeling of the stem and response of the item part. The stem and response could be each labeled with multiple representations based on the item part. On the August 2017 exam, the stem was labeled with V-D, W, S-Q, and S-A. The response was labeled

with N-N. For the written (W) stem, the words “recursive function” were circled as this is target vocabulary for algebra 1. Of course, the other words in the stem also indicate the written representation. An example of this is on the *current* August 2017 exam item 15 shown in Figure 32 below.

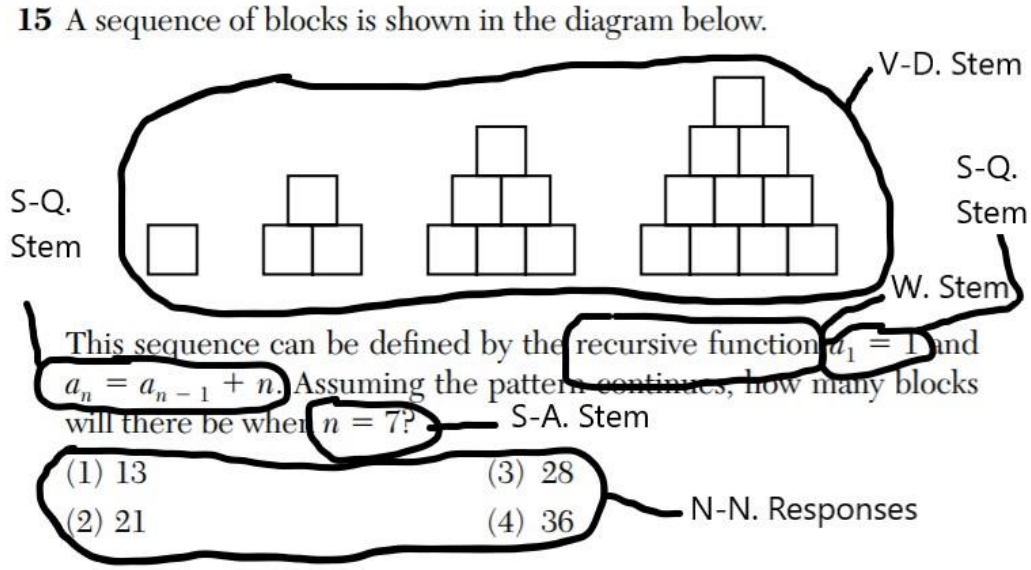


Figure 32. Example of the labeling of representations on the *current* August 2017 exam, item 15.

After labeling all of the item parts on all exams (*old*, *recent*, and *current*), the most prominent representational subcategories were determined for both stem and responses: visual-graph, symbolic-algebraic, numeric-numeral, and written. In order to find out patterns for these prominent subcategories, if the item had a visual-graph stem and a visual-graph response, it was counted as a 1 for that column. For the previous example (August 2017, #15), there would be a point assigned for written to numeral and a point assigned for symbolic-algebraic to numeral. Then, a relative percentage of these representation relationships was calculated per exam, and medians were found based on all 20 *old*, 22 *recent*, and 14 *current* exams as indicated in Figure 33.

The most prominent representation relationship medians were very similar when comparing the *old* and *recent* exams. For example, in the top five representation relationships for *old* and *recent* exams, the first, second, and fifth relationship median were the same in both in the *old* and *recent* exams with the third and fourth relationships being the same but switched in order. Different relationships were more prevalent on the *current* exams than on either the *old* or *recent* exams. This may indicate that in terms of representation relationships, the *old* and *recent* exam item parts had more similarities than compared to the *current* exam item parts.

Old Exam Relationships and Medians	*Recent Exams Relationships and Medians*	*Current Exams Relationships and Medians*
1. Written to Numeric-Numeral (41%) 2. Numeric-Numeral to Numeric-Numeral (28.2%) 3. Written to Symbolic-Algebraic (20.5%) 4. Symbolic-Algebraic to Symbolic-Algebraic (20%) 5. Written to Visual-Graph (17.9%)	1. Written to Numeric-Numeral (38.9%) 2. Numeric-Numeral to Numeric-Numeral (24.7%) 3. Symbolic-Algebraic to Symbolic-Algebraic (20%) 4. Written to Symbolic-Algebraic (19.3%) 5. Written to Visual-Graph (12.8%)	1. Written to Symbolic-Algebraic (38%) 2. Symbolic-Algebraic to Symbolic-Algebraic (35.3%) 3. Written to Visual-Graph (31.1%) 4. Written to Numeric-Numeral (26.7%) 5. Symbolic-Algebraic to Written (21.6%)

Figure 33. Prevalence of representation relationships per exam administration for *old*, *recent*, and *current* exam versions.

***Old* exams.**

On the old exams, there were no visual-graph stems. Visual-graphs were only found in responses after a written stem. The prominent stem-response relationship response by far was written stem and a numeric-numeral response. The wider bands on the Sankey diagram indicate more prevalent stem-response relationships among the four most common representation subcategories. Figure 34 and Table 25 illustrates these relationships for the *old* exam.

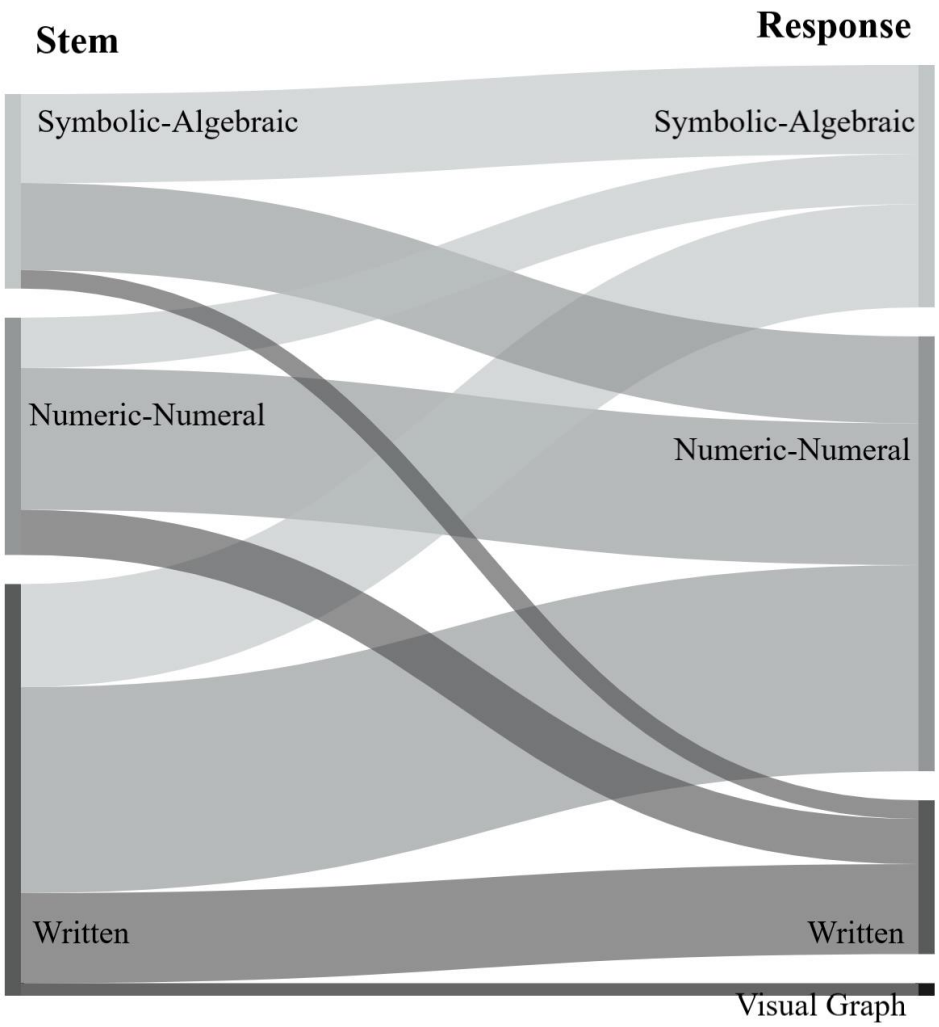


Figure 34. Sankey diagram of the medians of stem-to-response relationships for *old* exams.

Table 25

Stem-to-Response Relationship Medians for Old *Exams.*

Stems	Responses			
	V-G	S-A	N-N	W
V-G	-	-	-	-
S-A	-	17.7	17.3	3.7
N-N	-	10.0	28.2	9.0
W	2.5	20.5	41.0	17.9

***Recent* exams.**

On the *recent* exams, there is a noticeable difference in the interactions of different stems and responses. Written stem to numeral response could still be considered a strong relationship on this exam. There are thinner bands connecting representations which indicate a greater diversity in representation relationships as compared to *old* exams. Figure 35 and Table 26 illustrates these relationships for the *recent* exam.

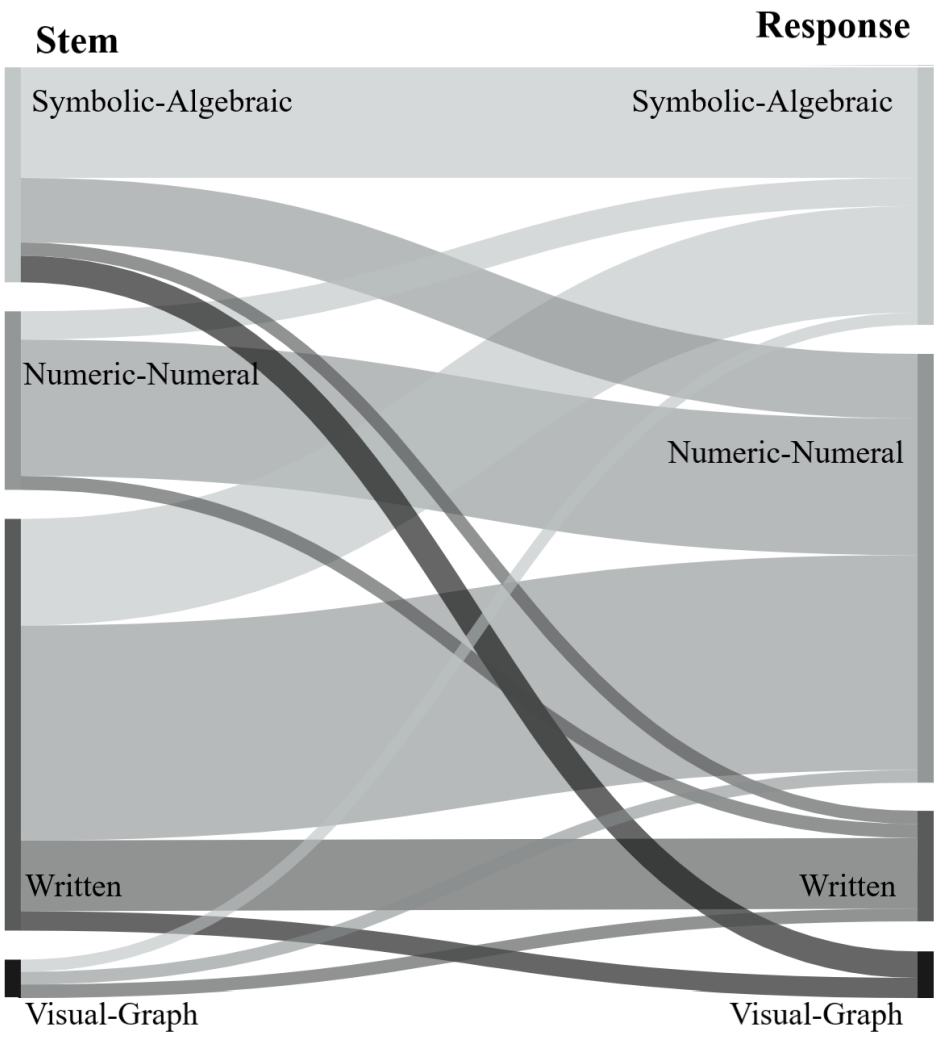


Figure 35. Sankey diagram of the medians of stem-to-response relationships for *recent* exams.

Table 26

Stem-to-Response Relationship Medians for Recent *Exams.*

Stems	Responses			
	V-G	S-A	N-N	W
V-G	-	2.2	2.3	2.3
S-A	4.8	20.0	11.8	2.4
N-N	-	5.1	24.7	2.5
W	3.5	19.3	38.9	12.8

***Current* exams.**

The *current* exam item parts have interaction among all four stem and response subcategories. Now, the most prevalent relationships are written stem to symbolic-algebraic response, symbolic-algebraic stem to symbolic-algebraic response, and written stem to visual-graph response. By far, the *current* exams have the most stem-to-response relationships and indicate the highest amount of diversity of stems to response relationships. Figure 36 and Table 27 illustrates these relationships for the *current* exam.

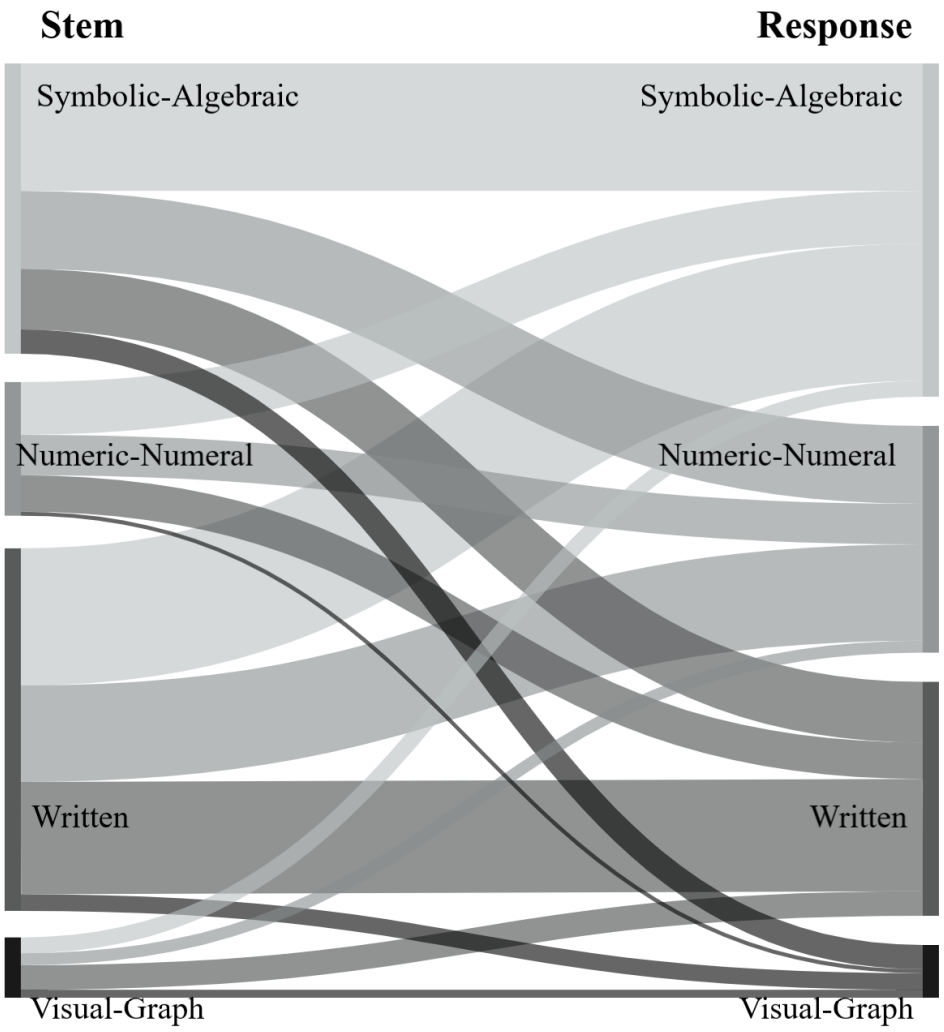


Figure 36. Sankey diagram of the medians of stem-to-response relationships for *current* exams.

Table 27

Stem-to-Response Relationship Medians for Current *Exams.*

Stems	Responses			
	V-G	S-A	N-N	W
V-G	2.2	4.4	3.3	6.8
S-A	6.7	35.3	21.6	16.8
N-N	1.1	14.6	11.2	10.1
W	4.6	38.0	26.7	31.1

Additional tests for Visual-Graph stems.

Tests were run on the connections between visual-graph stems and the responses of visual-graph, symbolic-algebraic, numeric-numeral, and written. Further tests indicated that there was a statistically significant difference between *old* and *current* exams for visual-graph, symbolic-algebraic, numeric-numeral, and written responses. This indicates that there was a significant difference between *old* and *current* exams for visual-graph stems across all types of major responses (visual-graph, symbolic-algebraic, numeric-numeral, and written). The tests also indicated that there was a statistically significant difference between *recent* and *current* exams for visual-graph, symbolic-algebraic, and written responses.

For visual-graph stem to visual-graph responses, a Kruskal-Wallis H test showed that there was a statistically significant difference in this relationship among the different exam versions, $\chi^2(2) = 11.365$, $p = 0.003$ with a mean rank score of 23.15 for *old* exams, 27.05 for *recent* exams, and 38.43 for *current* exams. There was a statistically significant difference between the *old* and *current* exams ($z = -15.279$, $p = 0.003$) (adjusted Bonferroni). There was a

statistically significant difference between the *recent* and *current* exams (z = -11.383, p = 0.036) (adjusted Bonferroni).

For visual-graph stems to symbolic-algebraic responses, a Kruskal-Wallis H test showed that there was a statistically significant difference in this relationship among the different exam versions, $\chi^2(2) = 15.892$, $p < 0.001$ with a mean rank score of 19.75 for *old* exams, 28.36 for *recent* exams, and 41.21 for *current* exams. There was a statistically significant difference between the *old* and *current* exams (z = -21.464, p < 0.001) (adjusted Bonferroni). There was a statistically significant difference between the *recent* and *current* exams (z = -12.851, p = 0.045) (adjusted Bonferroni).

For visual-graph stems to numeric-numeral responses, a Kruskal-Wallis H test showed that there was a statistically significant difference in this relationship among the different exam versions, $\chi^2(2) = 11.543$, $p = 0.003$ with a mean rank score of 19.98 for *old* exams, 30.27 for *recent* exams, and 37.89 for *current* exams. More specifically, there was a statistically significant difference between the *old* and *current* exams (z = -17.918, p = 0.003) (adjusted Bonferroni).

For visual-graph stems to written responses, a Kruskal-Wallis H test showed that there was a statistically significant difference in this relationship among the different exam versions, $\chi^2(2) = 19.042$, $p < 0.001$ with a mean rank score of 20.78 for *old* exams, 25.89 for *recent* exams, and 43.64 for *current* exams. There was a statistically significant difference between the *old* and *current* exams (z = -22.868, p < 0.001) (adjusted Bonferroni). There was a statistically significant difference between the *recent* and *current* exams (z = -17.756, p = 0.002) (adjusted Bonferroni).

Additional tests for Symbolic-Algebraic stems.

Tests were run on the connections between symbolic-algebraic stems and the responses of visual-graph, symbolic-algebraic, numeric-numeral, and written. Further tests indicated that there was a statistically significant difference between *old* and *current* exams for visual-graphs, symbolic-algebraic, and written responses. The tests also indicated that there was a statistically significant difference between *recent* and *current* exams for symbolic-algebraic, numeric-numeral, and written responses. Finally, the tests also indicated that there was a statistically significant difference between *old* and *recent* exams for visual-graph responses.

For symbolic-algebraic stems to visual-graph responses, a Kruskal-Wallis H test showed that there was a statistically significant difference in this relationship among the different exam versions, $\chi^2(2) = 24.948$, $p < 0.000$ with a mean rank score of 14.73 for *old* exams, 32.11 for *recent* exams, and 42.50 for *current* exams. There was a statistically significant difference between the *old* and *recent* exams ($z = -17.389$, $p = 0.002$) (adjusted Bonferroni). There was a statistically significant difference between the *old* and *current* exams ($z = -27.775$, $p < 0.001$) (adjusted Bonferroni).

For symbolic-algebraic stems to symbolic-algebraic responses, a Kruskal-Wallis H test showed that there was a statistically significant difference in this relationship among the different exam versions, $\chi^2(2) = 25.901$, $p < 0.001$ with a mean rank score of 18.58 for *old* exams, 25.80 for *recent* exams, and 46.93 for *current* exams. There was a statistically significant difference between the *old* and *current* exams ($z = -28.354$, $p < 0.001$) (adjusted Bonferroni). There was a statistically significant difference between the *recent* and *current* exams ($z = -21.133$, $p < 0.001$) (adjusted Bonferroni).

For symbolic-algebraic stems to numeric-numeral responses, a Kruskal-Wallis H test showed that there was a statistically significant difference in this relationship among the different exam versions, $\chi^2(2) = 19.466$, $p < 0.001$ with a mean rank score of 29.88 for *old* exams, 18.23 for *recent* exams, and 42.68 for *current* exams. There was a statistically significant difference between the *recent* and *current* exams ($z = -24.451$, $p < 0.001$) (adjusted Bonferroni).

Finally, for symbolic-algebraic stems to written responses, a Kruskal-Wallis H test showed that there was a statistically significant difference in this relationship among the different exam versions, $\chi^2(2) = 32.066$, $p < 0.001$ with a mean rank score of 23.68 for *old* exams, 19.52 for *recent* exams, and 49.50 for *current* exams. There was a statistically significant difference between the *recent* and *current* exams ($z = -29.977$, $p < 0.001$) (adjusted Bonferroni). There was a statistically significant difference between the *old* and *current* exams ($z = -25.825$, $p < 0.001$) (adjusted Bonferroni).

Additional tests for Numeric-Numeral stems.

Tests were run on the connections between numeric-numeral stems and the responses of visual-graph, symbolic-algebraic, numeric-numeral, and written. Further tests indicated that there was no statistically significant difference among exams for visual-graph responses. There was a statistically significant difference between *old* and *current* exams for symbolic-algebraic responses. The tests also indicated that there was a statistically significant difference between *recent* and *current* exams for numeric-numeral and written responses. Finally, the tests also indicated that there was a statistically significant difference between *old* and *recent* exams for visual-graph responses.

For numeric-numeral stems to visual-graph responses, a Kruskal-Wallis H test showed that there was no statistically significant difference in numeric-numeral to visual-graph

relationships among the different exam versions, $\chi^2(2) = 5.585$, $p = 0.061$ with a mean rank score of 32.20 for *old* exams, 23.20 for *recent* exams, and 31.54 for *current* exams.

For numeric-numeral stems to symbolic-algebraic responses, a Kruskal-Wallis H test showed that there was a statistically significant difference in numeric-numeral to symbolic-algebraic relationships among the different exam versions, $\chi^2(2) = 28.955$, $p < 0.001$ with a mean rank score of 31.03 for *old* exams, 15.59 for *recent* exams, and 45.18 for *current* exams. There was a statistically significant difference between the *old* and *recent* exams (z = 15.434, p = 0.007) (adjusted Bonferroni). There was a statistically significant difference between the *recent* and *current* exams (z = -29.588, p < 0.001) (adjusted Bonferroni). There was a statistically significant difference between the *old* and *current* exams (z = -14.154, p = 0.038) (adjusted Bonferroni).

For numeric-numeral stems to numeric-numeral responses, a Kruskal-Wallis H test showed that there was a statistically significant difference in numeric-numeral to numeric-numeral relationships among the different exam versions, $\chi^2(2) = 30.267$, $p < 0.001$ with a mean rank score of 39.75 for *old* exams, 30.77 for *recent* exams, and 8.86 for *current* exams. There was a statistically significant difference between the *recent* and *current* exams (z = 21.916, p < 0.001) (adjusted Bonferroni). There was a statistically significant difference between the *recent* and *current* exams (z = 30.893, p < 0.001) (adjusted Bonferroni).

For numeric-numeral stems to written responses, a Kruskal-Wallis H test showed that there was a statistically significant difference in numeric-numeral to written relationships among the different exam versions, $\chi^2(2) = 9.589$, $p = 0.008$ with a mean rank score of 33.05 for *old* exams, 20.18 for *recent* exams, and 35.07 for *current* exams. There was a statistically significant difference between the *old* and *recent* exams (z = 12.868, p = 0.032) (adjusted Bonferroni).

There was a statistically significant difference between the *recent* and *current* exams (z = -14.890, p = 0.022) (adjusted Bonferroni).

Additional tests for Written stems.

Tests were run on the connections between written stems and the responses of visual-graph, symbolic-algebraic, numeric-numeral, and written. Further tests indicated that there was no statistically significant difference among exams for written responses. There was a statistically significant difference between *old* and *current* exams for symbolic-algebraic and numeric-numeral responses. The tests also indicated that there was a statistically significant difference between *recent* and *current* exams for visual-graph, symbolic-algebraic and numeric-numeral responses. Finally, the tests also indicated that there was a statistically significant difference between *old* and *recent* exams for visual-graph responses.

For written stem to visual-graph response, a Kruskal-Wallis H test showed that there was a statistically significant difference in these relationships among the different exam versions, $\chi^2(2) = 26.683$, $p < 0.001$ with a mean rank score of 31.13 for *old* exams, 16.02 for *recent* exams, and 44.36 for *current* exams. There was a statistically significant difference between the *old* and *recent* exams (z = 15.102, p = 0.008) (adjusted Bonferroni). There was a statistically significant difference between the *recent* and *current* exams (z = -28.334, p < 0.001) (adjusted Bonferroni).

For written stem to symbolic-algebraic response, a Kruskal-Wallis H test showed that there was a statistically significant difference in these relationships among the different exam versions, $\chi^2(2) = 25.821$, $p < 0.000$ with a mean rank score of 24.93 for *old* exams, 19.80 for *recent* exams, and 47.29 for *current* exams. There was a statistically significant difference between the *recent* and *current* exams (z = -27.490, p < 0.001) (adjusted Bonferroni). There was

a statistically significant difference between the *old* and *current* exams (z = -22.361, p < 0.001) (adjusted Bonferroni).

For written stem to numeric-numeral response, a Kruskal-Wallis H test showed that there was a statistically significant difference in these relationships among the different exam versions, $\chi^2(2) = 16.860$, $p < 0.001$ with a mean rank score of 35.05 for *old* exams, 32.32 for *recent* exams, and 13.14 for *current* exams. There was a statistically significant difference between the *recent* and *current* exams (z = 19.175, p = 0.002) (adjusted Bonferroni). There was a statistically significant difference between the *old* and *current* exams (z = 21.907, p < 0.001) (adjusted Bonferroni).

For written stem to written response, a Kruskal-Wallis H test showed that there was no statistically significant difference in these relationships among the different exam versions, $\chi^2(2) = 4.911$, $p = 0.086$ with a mean rank score of 23.83 for *old* exams, 27.84 for *recent* exams, and 36.21 for *current* exams.

Research Question 8: Representation Relationships

The relationships among representations were determined for each item part. Then, a relative representation relationship percentage was determined for all exam administrations from June 2002 to August 2018 on all 20 *old* exams, 22 *recent* exams, and 14 *current* exams.

By exam version and item type.

Median scores were determined for each of the three types of exams (i.e., *old*, *recent*, *current*) for all the different representation relationships of representation treatment (RT), representation conversion (RC), and representation synergy (RS).

For each exam, the medians of the percentage of item parts with representation treatments slightly decreased: *old* (14.8%), *recent* (14.8%), and *current* (10.5%) as displayed in Figure 37.

Further testing also indicated that there was a significant difference between the *old* and *current* exams as well as between the *recent* and *current* exams. A Kruskal-Wallis H test showed that there was a statistically significant difference in representational treatments relationships among the different exam versions, $\chi^2(2) = 10.085$, $p = 0.006$ with a mean rank score of 31.25 for *old* exams, 33.55 for *recent* exams, and 16.64 for *current* exams. There was a statistically significant difference between the *old* and *current* exams (z = 14.607, p = 0.030) (adjusted Bonferroni). There was a statistically significant difference between the *recent* and *current* exams (z = 16.903, p = 0.007) (adjusted Bonferroni).

The decrease occurred more with multiple choice items than with constructed response item parts as shown in Table 28. However, after further testing, neither constructed response nor multiple choice items had a significant difference between different exam versions. A Kruskal-Wallis H test showed that there was no statistically significant difference in representational treatments relationships among the different exam versions, $\chi^2(2) = 10.085$, $p = 0.218$ with a mean rank score of 24.20 for *old* exams, 32.73 for *recent* exams, and 28.00 for *current* exams. A Kruskal-Wallis H test showed that there was no statistically significant difference in representational treatments relationships among the different exam versions, $\chi^2(2) = 5.438$, $p = 0.066$ with a mean rank score of 31.03 for *old* exams, 31.73 for *recent* exams, and 19.82 for *current* exams.

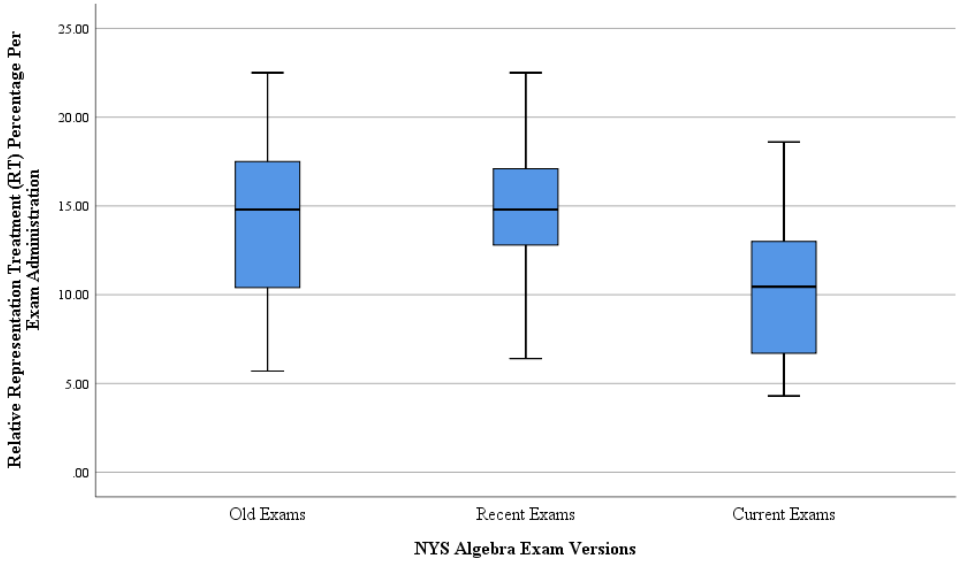


Figure 37. Box and whisker plots of relative representation treatment percentage for *old*, *recent*, and *current* exams.

Table 28

Relative Representation Treatments Percentage Medians by Item Type for Old, Recent, and Current *Exams*

Representation Treatments by Item Type	Percentage Median		
	Old	*Recent*	*Current*
Constructed Response	-	9.1	5.5
Multiple Choice	16.7	16.7	14.6
Total	14.8	14.8	10.5

For each exam, the medians of the percentage of item parts with representation conversions decreased: *old* (25.9%), *recent* (21.5%), and *current* (19.9%), with much more variability in *current* (IQR = 8.7-37.8) as compared to *old* (IQR = 11.4-37.5) and *recent* (IQR =

12.2-31.0) as noted in current exams than recent exams as shown in Figure 38. With further testing, there were no statistically significant differences among the exam versions. A Kruskal-Wallis H test showed that there was no statistically significant difference in representational conversions relationships among the different exam versions, $\chi^2(2) = 2.054$, $p = 0.358$ with a mean rank score of 32.45 for *old* exams, 25.27 for *recent* exams, and 27.73 for *current* exams.

While the multiple choice items also followed this trend, there was an increase in the medians of the percentage of item parts for constructed response items as reported in Table 29. However, after further statistically testing, there were no statistically significant differences in the representation conversions for either constructed response or multiple choice. For constructed response, Kruskal-Wallis H test showed that there was no statistically significant difference in representational conversions relationships among the different exam versions, $\chi^2(2) = 4.798$, $p = 0.091$ with a mean rank score of 34.75 for *old* exams, 26.00 for *recent* exams, and 23.50 for *current* exams.

For multiple choice, a Kruskal-Wallis H test showed that there was no statistically significant difference in representational treatments relationships among the different exam versions, $\chi^2(2) = 4.320$, $p = 0.115$ with a mean rank score of 24.78 for *old* exams, 27.02 for *recent* exams, and 36.14 for *current* exams.

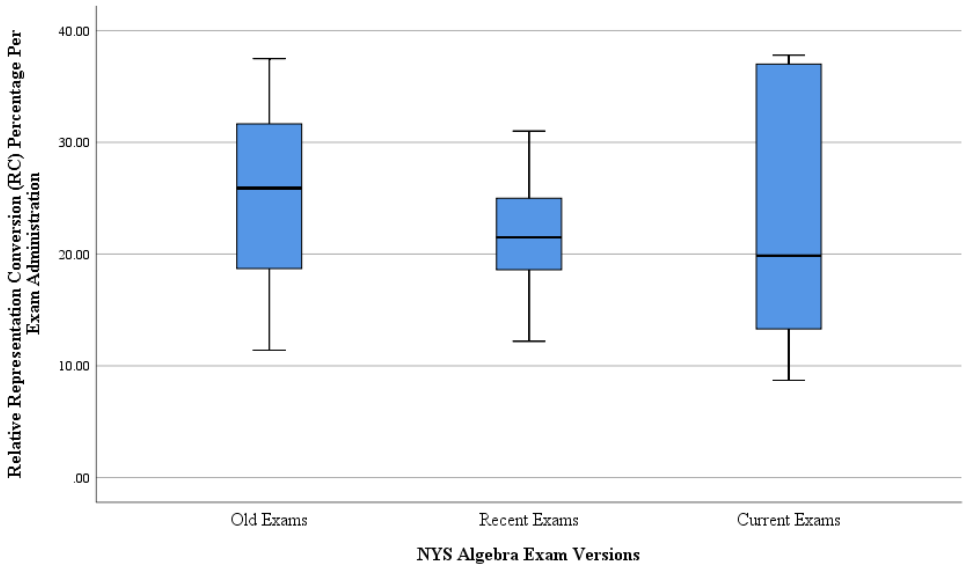


Figure 38. Box and whisker plots of relative representation conversion percentage for *old*, *recent*, and *current* exams.

Table 29

Relative Representation Conversions Percentage Medians by Item Type for Old, Recent, and Current *Exams*

Representation Conversion by Item Type	Percentage Median		
	Old	*Recent*	*Current*
Constructed Response	13.9	15.4	19.0
Multiple Choice	30.0	24.2	18.5
Total	25.9	21.5	19.9

For each exam, the medians of the percentage of item parts with representation synergy increased: *old* (58.9%), *recent* (63.4%), and *current* (68.0%), with the *current* exams having similar variability (IQR = 44.2-80.0) to the *old* exams (IQR = 45.0-82.9) and more variability than *recent* (IQR = 52.4-76.6) exams as displayed in Figure 39. However, after further testing, a

Kruskal-Wallis H test showed that there was no statistically significant difference in representational synergy relationships among the different exam versions, $\chi^2(2) = 3.388$, $p = 0.184$ with a mean rank score of 23.35 for *old* exams, 30.20 for *recent* exams, and 33.18 for *current* exams.

This trend was present when considering multiple choice items; there was a slight decrease in the percentage of item parts with representation synergy when considering constructed response item parts as noted in Table 30. After further testing, there was a statistically significant difference between *old* and *current* exams for multiple choice items only, but no significant difference for constructed response items. For constructed response, a Kruskal-Wallis H test showed that there was no statistically significant difference in representational synergy relationships among the different exam versions, $\chi^2(2) = 4.347$, $p = 0.114$ with a mean rank score of 33.48 for *old* exams, 28.34 for *recent* exams, and 21.64 for *current* exams. For multiple choice, a Kruskal-Wallis H test showed that there was a statistically significant difference in representational treatments relationships among the different exam versions, $\chi^2(2) = 9.205$, $p = 0.010$ with a mean rank score of 20.93 for *old* exams, 29.32 for *recent* exams, and 38.04 for *current* exams. There was a statistically significant difference between the *old* and *current* exams ($z = -17.111$, $p = 0.008$) (adjusted Bonferroni).

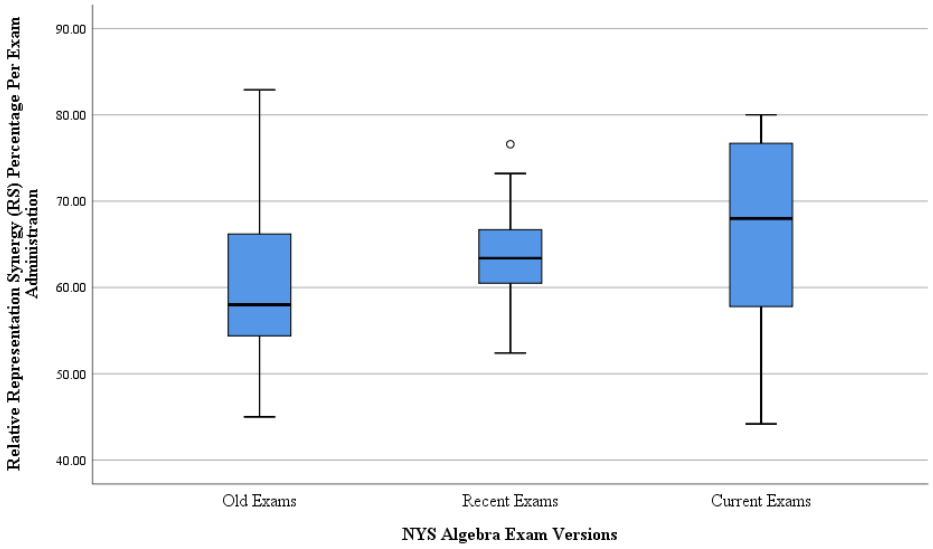


Figure 39. Box and whisker plots of relative representation synergy percentage for *old*, *recent*, and *current* exams.

Table 30

Representation Synergy by Item Type for Old, Recent, *and* Current *Exams*

Representation Synergy by Item Type	Percentage Median		
	Old	*Recent*	*Current*
Constructed Response	82.3	77.8	77.5
Multiple Choice	51.7	57.5	66.7
Total	58.9	63.4	68.0

Summary

Overall, there are significant differences among the *old*, *recent*, and *current* New York State algebra exams. When looking at multiple choice and constructed response items, these patterns were found significantly when considering multiple choice items. In general, there was an increase in the number of item parts from the *old* exams to the *current* exams. Also, by using multiple tests of readability, there was a general increase in the level of reading difficulty,

including a significant increase in the number of words from *old* to *current* exams especially within the multiple choice items. Additionally, there was an increase in the relative percentage of context among exam administrations from *old* to *current* exams, with a definite increase when considering the multiple choice items. However, when considering readability and context together, there was no significant relationship between these two factors.

Although there was no significant increase or decrease of complexity, there was more variability of levels of relative complexity on *current* exam administrations as compared to *old* exam administrations. A significant positive correlation was found between relative complexity and relative context.

Finally, several significant patterns emerged in terms of representations and their corresponding relationships among other representations. There were more representations in *current* multiple choice items than in *old* or *recent* multiple choice items. However, there were more representations in *old* constructed response item parts than in *recent* or *current* constructed response items. More representations were included in the stem than in the response for multiple choice items. When looking at all of the possible representations, four major representation subcategories emerged: visual-graphs (V-G), symbolic-algebraic (S-A), numeric-numeral (N-N), and written (W) representations. From *old* exams to *current* exams, there was an increase in visual-graphs (V-G) and symbolic-algebraic (S-A) representations, a slight increase in written (W) representations, and a decrease in number-numeral (N-N) representations.

Chapter 5 – Summary, Discussions, Implications

This study aimed to provide teachers with a more holistic view of state algebra exams by looking at the item parts' representations and relationships between or among the readability, context, cognitive complexity, representations, and the relationships between or among representations. In terms of readability, context, complexity, representations, and the relationships among representation, there was a noticeable increase and/or increase in the diversity of item parts by comparing *old*, *recent*, and *current* exams. This chapter has five sections. The first section summarized the major findings of the study based on the research questions. The second section summarized the study by exam (*old*, *recent*, and *current*). The third section described the limitations of this study of New York state algebra exam item parts. Finally, the fourth section described recommendations for future research.

Discussion

In this dissertation, the evolution of three different New York state algebra assessments (*old*, *recent*, and *current*) was analyzed in terms of the exams' readability, context, complexity, representations, and relationships among representations. Readability measures were calculated for each item and each exam administration for all *old*, *recent*, and *current* items and exam administrations. Context, complexity, representations in both stems and responses, and overall relationships were determined for each item part for all *old*, *recent*, and *current* item parts. Overall, there are significant differences among the *old*, *recent*, and *current* exams not only in terms of the standards but also in terms of readability, context, complexity, representations, and the relationships among representations.

Research Question 1: What is the average grade level and average level of reading ease for each exam (*old, recent,* or *current)* and item part type (constructed response, multiple choice)?

In the first research question, several readability tests were used to analyze *old*, *recent*, and *current* exams. All readability tests used (e.g., Flesch Reading Ease (1975), Flesch-Kincaid Reading Level (1979), ATOS score, average word length, average sentence length, average vocabulary level, and word count) indicated similar patterns: there was a general increase in the median level of readability from *old* to *current* exam administrations. There was a very notable increase in the number of words from *old* to *current* exam administrations. When considering the items, there was a more marked increase in readability difficulty on multiple choice items as compared to constructed response items when comparing *old* and *recent* items to *current* items. Overall, c*urrent* exams require more reading and more difficult reading than *old* or *recent* exams.

Previous research also indicated an increase in readability with algebra problems in a middle and high school online tutoring program (Walkington, Clinton, & Howell, 2015), on *current* New York State Algebra 1 Regents exams (Spoth, 2016), and ninth grade mathematics (Leiss, Schukajlow, Blum, Messner, & Pekrun, 2010). Additionally, similar vocabulary trends among New York State middle school mathematics assessments (Czekanski, 2018) and New York State algebra 1 examinations (Spoth, 2016) have been determined by rating vocabulary on the assessment into tiers. Numerically, the readability grade level is close to the grade level of ninth grade students which is comparable to King and Burge's (2015) analysis of PISA assessments. In the current study, there were more words on average per sentence from *old* to *current* state algebra exams while there were fewer words per sentences from 1987 to 2011 in Morgan and Sfard's (2016) research. This may have occurred because Morgan and Sfard (2016)

measured high stakes examinations in England while this study is concentrated on high stakes algebra exams in New York State. Overall, this study aligns well with previous research.

Research Question 2: Do certain exams (*old, recent,* or *current*) rely more on a "real word" context? Which item part types (constructed response, multiple choice) rely the most on using "real world" contexts?

The second research question was an analysis of the presence or absence of context was analyzed by exam version and items. There was a definite increase in the relative context per exam administration. When comparing multiple choice and constructed response items, the median of the multiple choice items increased more from the *old* to the *current* exams. This supports the NCTM's (2014) recommendations in the *Principles to Action* where real world contexts are promoted as a way to encourage mathematical thinking and the ability to apply knowledge of mathematics to every day concepts. Also, the idea of real world context has been emphasized more in the *current* exam standards as compared to previous standards. This increased emphasis in the standards has affected the writing of the item parts.

Previous studies did not measure the percentage of contextual problems. However, previous studies did indicate that understanding the problem solving context is an important step in solving algebraic item parts. However, this finding that *current* exams have more contextual problems in the assessments than *old* or *recent* exams supports the underlying ideas of reform curriculum and the mission of the Common Core Algebra test as "the middle school and high school standards call on students to practice applying mathematical ways of thinking to real-world issues and challenges" (Common Core State Standards Initative, 2020).

Research Question 3: Is there a correlation between item readability and presence of context?

The third research question examined a possibility of a relationship between readability and the relative context. There was no significant relationship found between readability and context. This could be due to the fact that readability tests often become averages of the text level. A higher readability score may occur when there is a higher concentration of multi-syllable words when thinking about the Flesch reading ease and Flesch Kincaid, or when there is a higher concentration of difficult words when calculating with the ATOS level calculator.

Also, word complexity does not necessary indicate mathematical complexity. When analyzing the ATOS level of the phrase "system of inequalities," its determined ATOS score is a 6.5 average vocabulary level. Of course, the concept of translating a word problem into a systems of equations, graphing this system, and then interpreting it based on the previous context of the problem is not at a sixth grade level of comprehension.

Research Question 4: What is the average complexity of the items labeled by exam (*old*, *recent*, and *current*)? What is the average complexity of items labeled by item part types (constructed response, multiple choice)?

Cognitive complexity was analyzed by exam version and items. Although there was not a significant increase in complexity from *old* to *current* exams, there was more of an interquartile range with *current* exams as compared to *old* exams, indicating that it may be less predictable on the *current* exams to know the percentage of complex items per administration of the exam. When looking at seventh grade Korean algebra standards-based curriculum, Cho and Kwon (2012) found that Korean algebra textbooks had a higher level of cognitive demand through the years, and Park (2011) also found that more recent algebra 1 textbooks had fewer low level tasks.

There are several reasons why there might not a general increase in cognitive demand when comparing *old*, *recent*, and *current* items. All item parts were considered singly when considering their level of cognitive demand. The increased number of items parts in the *current* exams may indicate that there is more guidance in the problem solving process.

When considering the greater interquartile range in the cognitive demand levels on *current* exams versus previous versions, this means that the *current* exams are less predictable than the *old* or *recent*. If all item parts for the standard are typically at a level two, then there needs to be more analysis that may possibly indicate that the items are constructed similarly. When considering the standards, if the item parts are at various levels, this may indicate that the student would need to be more careful in reading the problem as the item part may not always ask something similar. For example, with an inequality item, students may need to solve the inequality for the variable (i.e., a solution of $x > 2$), or they may need to label the smallest integer value in the inequality set or the value three. The students may also need to reason within the problem situation and explain why three would be an appropriate answer for the context of the problem as well.

Research Question 5: Are there correlations among readability, context, and complexity?

The fifth research question determined correlations among cognitive complexity, context, and readability. There was a significant positive correlation between complexity and context. Previously, Berqvist, Dryvold, and Osterholm (2012) found that there were connections between vocabulary and task difficulty on the PISA assessments in mathematics in all grade levels. Complexity does not equate to task difficulty in terms of the mathematics; rather, it refers to the overall problem solving process and how representations may need to be used to find a response that used an algorithm or even inventive non-algorithmic thinking. Although one of the

readability measures was average vocabulary level, it is not the same as only selecting words that are difficult from the problem and seeing how many of those level words are present on the exams.

Problems that include context may tend to be more complex because of the introduction of the problem solving situation. By introducing a context, it adds to the semiotic bundle of the problem which can be comprised of other representations that the learner must need to understand in order to answer the item. This context can make it easier, more distracting, or more complicated to answer the item. Drake and colleagues (2012) previously found that items on the GSCE mathematics examinations from England that contained context may still follow a routine that does not require the context in order to solve. This may also be similar to how the New York State *old*, *recent*, and *current* exams were written. For instance, on the *recent* exams June 2008 #12, the problem read the following way:

> "Carrie bought new carpet for her living room. She calculated the area of the living room to be 174.2 square feet. The actual area was 149.6 square feet. What is the relative error of the area to the nearest ten-thousandth?"

For the *recent* exams August 2011 #23, the problem was written very similarly:

> "Jack wants to replace the flooring in his rectangular kitchen. He calculates the area of the floor to be 12.8 square meters. The actual area of the floor is 13.5 square meters. What is the relative error of calculating the area of the floor, to the nearest thousandth?"

However, in the *current* standards, the problems are written so that students need to use the given contexts to solve the problems, which increases the overall complexity of the problem. The *current* exam January 2017 #26 illustrates this complexity:

"A typical marathon is 26.2 miles. Allan averages 12 kilometers per hour when running in marathons. Determine how long it would take Allan to complete a marathon, to the nearest tenth of an hour. Justify your answer."

This type of item is not present on all *current* exams and it requires students to understand how to exchange units by using the reference sheet. On the *current* June 2018 #15, the stem of the multiple choice item is asked in this way:

"The Utica Boilermaker is a 15-kilometer road race. Sara is signed up to run this race and has done the following training runs:

I. 10 miles

II. 44,880 feet

III. 15,560 yards

Which run(s) are at least 15 kilometers?"

This item requires multiple conversions and it is distinctly different than the *current* January 2017, #26 item. It requires more of a knowledge of the problem situation and the ability to do conversions versus the *recent* items where the student would remember the "recipe" of relative error calculation.

Research Question 6: How many representations appear on average for each standard and overall for each exam (*old, recent,* or *current*)? What types of mathematical representations appear most frequently by item part types (constructed response, multiple choice)?

In the sixth research question, the number and type of representations were determined by exam version. There were more representations in the *current* multiple choice than in the *old* or *recent* multiple choice. However, there were more representations in the *old* constructed response than in the *recent* or *current* constructed response. *Current* exams included more

representations in the stem than in the response for multiple choice item parts. By analyzing state algebra exam item parts, four major representations emerged as being prominent across *old*, *recent*, and *current* exam administrations: visual-graph, symbolic-algebra, numeric-numeral, and written representation. Visual-graphs and symbolic-algebra were more present in *current* exam item parts than in *recent* or *old* exam item parts in both the multiple choice and constructed response. There was a decrease in the number of representations from *old* to *current exams*, and a slight increase in written representations from *old* to *current exams*.

Symbolic notation has also been mentioned as an important representation in algebraic research. Students may not naturally use a symbolic method to solve items on assessments (Swafford & Lagrall, 2000). Essentially, since the *current* exams are richer in symbols, when compared to previous exams, students need to be able to be fluent with these symbols and able to compare symbolic representations, such as equations,to the other representations present in the problem solving process. These multiple representations can increase the level of understanding of symbolic notation as found on assessments for pre-algebra students in high school (Moseley & Brenner, 1997).

Although students can have difficulty with number types (Fischbein, Jehiam, & Cohen) including decimals and fractions (Calhoon et al., 2007; Hiebert & Wearne, 1983; Tien & Siegler, 2017), there were fewer numbers in both the stem and responses of items on the *current* exams versus previous versions of exams. This may indicate that numbers are instead being used within other representations (such as the *y*-intercept of the symbolic algebraic equation $E(x) = 30 + 2.40x$ on the *current* August 2018, #17).rather than by itself.

Research Question 7: How do representations interact with one another in the *stem* and *response* of the item? Do certain types of relationships occur with representations occur on certain exams (*old, recent,* or *current*) or by item part types (constructed response, multiple choice)?

The seventh research question addressed the stem-to-response relationships among representations by exams. Visual-graph, symbolic-algebra, numeric-numeral, and written representations were found to be major representations through all exams (*old*, *recent*, and *current*). This is similar to the research of Panasuk (2010) who used a three-phase ranking framework of verbal, visual, and symbolic representations. In Panasuk's (2010) research, students who used multiple representations were able to answer items more successfully than students who did not use multiple representations. On *current* exams, I found that there was more of diversity in the connections among representations as compared to the previous two versions of the state algebra exams.

Research Question 8: What types of representation relationships occur most frequently according to exam (*old, recent,* or *current*) and item part types (constructed response, multiple choice)?

Finally, the eighth research question was an analysis of the patterns of representation relationships. Overall, there was a decrease in the relative percentage of representation treatments from *old* to *current* that occurred more with multiple choice items than with constructed response items. There was also a decrease in the relative percentage of representation conversions which was more present in the multiple choice items. There was definitely more variability in the relative percentages of item parts with representation conversions on *current* exams as compared to previous versions, indicating that the number of items with representation

conversions did not stay consistent across *current* exams. Finally, there was an increase in the relative percentage of representation synergy which was more present in the multiple choice and an increased variability in the relative percentages of item parts with representation synergy on *current* exams as compared to previous versions.

Representation relationships have not been studied in comparison to their inherent levels of difficulty. As a theory, it may be assumed that the more representations displayed in a semiotic bundle, the more difficult an item is on an examination. However, it is more about looking at how the representations would work together in order to help students understand the overall stem as representations may either increase or decrease cognitive load that exists within the item.

Summary by Exam Type

***Old* exams.**

The *old* exams had a median of 39 item parts, which was the smallest number of item parts compared to *recent* and *current* exams. *Old* exams only included **important** (10% to 25% of the exam items) and **minor** (less than 10% of the exam) with no **major** topic areas (more than 25% of the exam). This is unlike the *recent* exam which had one **major** topic area and the *current* exam which had two **major** topic areas. *Old* exams had the lowest reading level overall across all readability tests. *Old* exams had a similar amount of words (940.5) to *recent* exams (932.5) but this was smaller than the number of words on *current* exams (1281.5).

The *old* exam relative context percentage was in the middle between the lower *recent* exam percentage and the higher *current* exam percentage. However, the relative cognitive complexity percentages on the *old* exams was the highest for high cognitive complexity and the lowest for low cognitive complexity items among the three exams. The interquartile range for

both high and low cognitive complexity was smaller than *recent* and *current* exams, indicating that *old* items were predictably more complex than *recent* or *current* exam item parts.

The *old* exams had fewer representations overall in each item part as compared to *recent* and *current*. While the relative representation percentages of visual-geometric figure (10.0%) and symbolic-geometric notation (9.0%) were prominent on the *old* exams, the visual-graph, symbolic algebraic, numeric-numeral, and written representations were chosen to be analyzed so that general patterns could be seen through all three exams. The relative visual-graph for stems and responses and the symbolic-algebraic percentage for stems was the smallest for the *old* exam item parts as compared to *recent* and *current* exam item parts. In the relative symbolic-algebraic percentage for responses, the *old* exams were in the middle: slightly more *old* item parts than *recent,* but less *old* than *current* item parts. The largest relative numeric-numeral percentage occurred on both stems and responses for the *old* exam item parts as compared to *recent* and *current*. There was the smallest relative written percentage for stems for the *old* as compared to the *recent* and *current*. The relative written percentage for responses for the *old* item parts was in the middle: larger than the *recent* and a smaller than the *current*.

When analyzing the stem-response relationships between *old, recent,* and *current* exams, *old* exams had similar stem-response relationships to *recent,* but different than *current* with the least amount of connections among the prominent representational relationships (visual-graph, symbolic-algebraic, numeric-numeral, and written representations). When considering the relative representation relationship percentages, there was a higher percentage of representation treatments and representation conversions on the *old* item parts as compared to the *recent* and *current*. There was the lowest percentage of representation synergy on the *old* as compared to *recent* and *current*.

***Recent* exams.**

The *recent* exams had a median of 42 item parts, which was larger than the number of item parts on *recent*, but smaller than the number of item parts on the *current* exams. *Recent* exams included one **major** (more than 25% of the exam), two **important** (10% to 25% of the exam items) and two **minor** (less than 10% of the exam) topic areas which was similar to the *current* exam and different than the *old* exam. *Recent* exams had a higher reading level as compared to *old* exams but a lower level of readability as compared to *current* exams. *Recent* exams had a similar amount of words (932.5) to *recent* exams (940.5) but this was smaller than the number of words on *current* exams (1281.5). The *recent* exam relative context percentage was smaller than the *old* or *current* exam. The relative cognitive complexity percentage for high cognitive complexity items for *recent* items was in the middle: smaller than *old* exams and larger than *current* exams, but it was higher for low cognitive complexity items than the *recent* and *current* exams. The interquartile range for high and low cognitive complexity percentage was larger than *old* and smaller than *current* exams. This indicates that *recent* items were less complex than either *old* or *current* exams, but less predictably so than on *old* exam items.

The *recent* exams had more representations overall in each item part as compared to *old* exams, but had fewer representations than *current* exams. While the relative representation percentage of visual-geometric figure was 7.0%, the visual-graph, symbolic algebraic, numeric-numeral, and written representations were chosen to be analyzed so that general patterns could be seen through all three exams. A larger relative visual-graph percentage occurred in the *recent* exam item parts as compared to *old* exam item parts. There was a smaller relative visual-graph percentage that occurred on *recent* exam item parts as compared to *current* exam item parts. Similarly, a larger relative symbolic-algebraic percentage occurred in the *recent* exam item parts

as compared to *old* exam item parts. There was a smaller relative visual-graph percentage that occurred on *recent* exam item parts as compared to *current* exam item parts. Similarly, the relative symbolic-algebraic percentage on the *recent* exams was larger than the *old* exams, but smaller than the *current* exams.. In the relative symbolic-algebraic percentage for responses, there was a smaller percentage for *recent* item parts as compared to *old* or *current* item parts. The relative numeric-numeral percentage for *recent* exam item parts was larger than *old* and smaller than *current* exam item parts. There was a larger relative written percentage for stems for the *recent* item parts as compared to the *old* and a smaller relative written percentage for stems for *recent* item parts as compared to the *current* item parts. There was the smallest relative written percentage for stems and responses for the *recent* item parts as compared to the *old* and *current* item parts.

When analyzing the stem-response relationships between *old, recent,* and *current* exams, *recent* exams had similar stem-response relationships to *old* and different than *current* with more connections among the prominent representational relationships (visual-graph, symbolic-algebraic, numeric-numeral, and written representations) than *old* but less than *current*. When considering the relative representation relationship percentages, there were a smaller representation treatment percentage in *recent* item parts as compared to the *old*, but the *recent* had a larger percentage than the *current*. Also, there was a smaller relative representation conversion percentage in the *recent* as compared to the *old* but a larger percentage in the *recent* as compared to the *current*. There were more representation synergy item parts on the *recent* as compared to *old* and fewer parts as compared as to the *current*.

***Current* exams.**

The *current* exams had a median of 45 item parts, which was the largest number of item parts compared to *recent* and *current* exams. *Current* exams included two **major** (more than 25% of the exam) and two **minor** (less than 10% of the exam) with no **important** (10% to 25% of the exam items) topic areas which was similar to the *current* exam and different than the *old* exam. *Current* exams had the highest reading level as compared to *old* and *recent*. *Current* exams had the largest number of words per exam (1281.5). The *current* exam relative context percentage was the highest out as compared to the *old* and *recent* exam. The relative cognitive complexity percentage for high cognitive complexity items on the *current* exams was lower than the *recent* and *current* exams, and it was higher for low cognitive complexity items than the *old exams* and lower for *recent* exams. The interquartile range for both high and low cognitive complexity was larger than *recent* and *current* exams, indicating that *current* items, although possibly less cognitively complex, varied more in their levels of cognitive complexity.

The *current* exams had more representations overall in each item part as compared to *old* and *recent* exams. The visual-graph, symbolic algebraic, numeric-numeral, and written representations were chosen to be analyzed so that general patterns could be seen through all three exams without any other prominent representations on the *current* exams. The largest relative visual-graph percentage occurred on the *current* exam item parts as compared to *old* and *recent* exam item parts. Similarly, the largest relative symbolic algebraic percentage occurred on the *current* exam item parts as compared to *old* and *recent* exam item parts. The smallest relative numeric-numeral percentage occurred on both stems and responses for the *current* exam item parts as compared to *old* and *recent* exam item parts. There was the largest relative written

percentage for stem and responses for the *current* item parts as compared to the *old* and *recent* item parts.

When analyzing the stem-response relationships between *old*, *recent*, and *current* exams, *current* exams had different stem-response relationships than the *old* and *recent* with the most connections among the prominent representational relationships (visual-graph, symbolic-algebraic, numeric-numeral, and written representations). When considering the relative representation relationship percentages, there were fewer representation treatments on the *current* item parts as compared to the *old* and *recent*. Also, there was a smaller relative representation conversion percentage but with more variability in the percentage in the *current* versus *old* and *recent*. There was a higher percentage of representation synergy on the *current* as compared to *old* and *recent* with more variability in the percentage than *old* or *recent*. Table 31 further summarized the changes in the schema from *old*, *recent* and *current* exams.

Table 31

Major Findings for Old, Recent, *and* Current *Exams Organized by Schema*

Schema	Findings		
	Old	*Recent*	*Current*
Schema 1: Readability	• Lowest level	• Middle level	• Highest level
Schema 2: "Real World" Context	• Middle percentage	• Lowest percentage	• Highest percentage
Schema 3: Cognitive Complexity	• Highest percentage of high complexity items • Lowest percentage of low complexity items • Complexity levels varied the least	• Middle percentage of high complexity items • Highest percentage of low complexity items • Complexity levels varied more than *old*, but less than *current* exams	• Lowest percentage of high complexity items • Middle percentage of complexity items • Complexity levels varied the most
Schema 4: Representations	• Lowest percentage of Visual-Graph • Middle percentage of Symbolic-Algebraic • Highest percentage of Numeric-Numeral • Lowest percentage of Written	• Middle percentage of Visual-Graph • Lowest percentage of Symbolic-Algebraic • Middle percentage of Numeric-Numeral • Middle percentage of Written	• Highest percentage of Visual-Graph • Highest percentage of Symbolic-Algebraic • Lowest percentage of Numeric-Numeral • Highest percentage of Written
Schema 5: Stem-to-Response and Representation Relationships	• Similar Stem-to-response as *recent* • Highest percentage of Representation Treatment • Highest percentage of Representation Conversion • Lowest percentage of Representation Synergy	• Similar Stem-to-response as *old* • Middle percentage of Representation Treatment • Middle percentage of Representation Conversion • Middle percentage of Representation Synergy	• Different stem-to-response • Lowest percentage of Representation Treatment • Lowest percentage of Representation Conversion • Highest percentage of Representation Synergy

Limitations of Findings

There are several limitations to these findings. First of all, these are based on items on the state examinations in New York. Although different states may also use Common Core state standards, it does not mean that the New York State interpretation of these standards is consistent across states. Although Common Core standards (*current* standards) were written with the idea of competing with other countries around the world, there is no specific correlation that has been published by state, national, or international organizations that shows the alignment among the state, national, and international standards.

Another important limitation is that these results are by no means to encourage "gaming the system" in order to help students pass. These results are not intended for general "teaching to the test practices" where students take previous exams over and over to hopefully learn enough content for a passing grade. In the *current* exams especially, it simply cannot work as every exam can be quite different. My findings can actually allude to the fact that the items are different enough from *current* exam to exam that students should be learning the different topics and subtopic areas in a way in which topics naturally build on each other. Students should also be exploring their interrelatedness in class to reduce cognitive load and to organize their thoughts so that excellent problem solving methods are defined and carried out.

By studying *old*, *recent*, and *current* exams, it shows that educators should be interested in how students are answering items. Students should also be analyzing how others answer items to see where common mistakes are made, and therefore, correcting these mistakes within themselves. Instead of spending time going over and over a concept that the student may already know, educators need to find value in the statement "I don't know where to start". This is not always indicative of not knowing the math. Instead, it may be indicative of not uncovering the

main idea in the problem, difficulty with how different representations relate to the problem, not understanding the relation of the context with the mathematics, and/or retrieving a similar procedure to the correct procedure that is not the complete or accurate procedure.

It is certainly difficult to align previous state standards as they have been written with various different topic areas and were developed in distinctly different time periods of mathematics education in New York State. Students now have the ability to use graphing calculators on these exams which changes the way students can view algebra in real time. The difficulty in aligning previous standards together (*old*, *recent*, and *current*) does not necessarily indicate that there is not overlap. Teachers may want to utilize *old* and *recent* items in their instructional practices by using them without any changes if they meet *current* standards, modifying them to meet *current* standards by changing the wording and/or the scope of the item, or using them as an item to build understanding so that students will be able to achieve *current* standards in the future. Previous items that were a part of *old* and *recent* exams can be useful, but they should always be used with the *current* standards in mind.

Readability was measured through various readability tests and similar trends in readability were found among the *old*, *recent*, and *current* item parts. However, the actual grade levels found with the readability tests should be used with caution. Reading levels do not indicate complexity levels in terms of context. For example, finding the answer to a systems of equations algebraically may not require many words while understanding what a function means may require the reading of a paragraph. The reading level may be more of an average of the text difficulty rather than an indicator of the number of difficult words present in the text.

When the presence or absence of context was indicated, there was not an indication of how much knowing about the context would affect the problem solving process. An item that has

context may be able to be solved without directly understanding the problem solving context depending on how it was written.

Complexity scores were assigned with the item in mind rather than the body of items in mind. Through more standards based research, there may be more trends in the similarity of the writing of the item. For instance, the *old* and *recent* exams may have more patterning in the amount of words and key phrases in items while the *current* exams may have more differences in how items with certain standards are represented on different exams.

There were more representations on *current* exams than on *old* or *recent* exams. The number and type of representations on exams may not necessarily correlate with the number and type of representations that should be used in the curriculum. For instance, when a function is graphed on an exam, the student may only need to graph the points properly from a graphing calculator for full credit. However, having students write out the representations of the table and set of points in the lessons during the year may allow students to see the necessary connections among these representations, the symbolic algebraic equation, and the graph.

Implications

Standards based education, cognitive load, and semiotics.

This research brings together the ideas of standards based education, cognitive load, and semiotics. NCTM's *Principles and Standards for School Mathematics* (2000) evolved to the current NCTM (2014) *Principles to Action* by developing specific mathematical teaching practices such as establishing connections among representations and using conceptual reasoning as a part of the educational process. States like New York adapted their assessments to these changes by NCTM. Unfortunately, when the standards in New York state changed from *old* to *current* exams and instead of having an evolution of the standards, there was more of an "adopt

and toss" approach where the teachers and students would be discouraged from using previous curriculum and/or assessments that were from previous versions of exams. The hope is that by focusing only on the standards present in the state that the teachers would learn this curriculum. In fact, with the beginning of Common Core, instruction only using the modules occurred in schools without the flexibility to instead "adapt and reuse" applicable materials. This dissertation research along with previous research component research indicates that while there have been changes in topic areas there was also significant changes in wording and connections among representation.

When students and teachers are focused on standards as the guide to organize meaning, there are areas of germane load that can occur. This can be overwhelming for students as they may be attempting to memorize routines of mathematical practices based on the topics in their math class versus understanding how each representation may be interwoven through multiple topic areas. A strategy to interpret the context of the written stem in a unit on linear functions may also apply to interpreting the context of the written stem in a unit about parabolic functions. Also, students and teachers may need to associate appropriate representations together in order to show how they are related with one another. This transforms the representations from single units of information that have their own systems of rules to a member of a bundle of representations that can help students understand the concepts more effectively.

The idea of "graphing a line" can be easily seen through the *old*, *recent*, and *current* exam items. This object has not changed. Although the objects (or formalized algebraic concepts embedded in the standards) may be different from exam to exam, there will always be interpretation of these objects by the learner, educator, and overall mathematical community. Educators will always be interpreting these signs that are within the curriculum in order to meet

the needs of the diverse group of students present in their classrooms. Students will always be editing their own definitions as interpretants of the formalized object in mathematics by using the signs or examples of the mathematical topic. In order to help students remember a concept, teachers may sometimes not fully encapsulate the idea of the object. In fact, formalized definitions of an object may not make sense to a student whose perspective of the mathematical world is only in basic numeracy and algebra. By changing the standards (and possibly the amount of information necessary to understand an object in algebra), it leads to more disjointed explanations of algebra and therefore more confusion.

As standards continue to evolve through New York State and national mathematics organizations should be considering the concepts of cognitive load in terms of the standards and representations on the exams. The unnecessary rewording of standards that occurred in the *old*, *recent*, and *current* standards affects the interpretation of algebraic concepts. An uncertainty with how the standard would be represented on an exam or how much of the standard is developmentally appropriate to be taught in a *current* algebra classroom may also greatly affect teaching practices. Rewording a standard can lead to an unnecessary reconceptualization of a topic which could then unfortunately lead to cognitive overload for both the learner and the instructor.

Fortunately, the Next Generation Standards in Algebra 1 (2017) are an edit of the *current* standards. This has been a positive step forward in New York State as the *current* standards system was not completely abolished. Instead, this is an evolution of the *current* standards which makes it more manageable for students, educators, families, curriculum experts, faculty members of universities, and other important members of the mathematical community. When mathematical standards evolve, mathematical practices become more sustainable and teachers

have the ability to craft curriculum that can be adapted over the years for different groups of students.

Algebra is more than a collection of standards.

This research indicates that algebra is more than a collection of standards. Although a standard may have a related semiotic bundle of representations, students and teachers should be aware that these representations overlap and that the "rules" of the representation may or may not be dictated by the standard area. What may not be considered is the rest of the item: readability, context, complexity, representations, and representational relationships affect the understanding of the standard. If each standard is considered singly without connection to other standards and the outside mathematical context, there could be a redundancy and/or an uncertainty of how these algebraic topics are a part of the study of mathematics.

For instance, reading and understanding the stem of an item part and explaining how that fits in the world (both real world and abstract mathematical world) may be the way for students to better interpret objects. When students understand that their own definitions are flexible, they may be better able to develop better definitions of concepts over time. Each student will understand the standard slightly differently based on their previous exposure to mathematics and mathematical concepts. Integrating the standard into familiar contexts may build students' mathematical meaning. Educators and students need to realize that today's algebra is a complex system of thinking which not only includes an strong interpretation of an algebraic object, but the ability to decode language and manipulate other representations in order to find a response that hopefully makes sense within the context of the problem.

However, general knowledge and reading can also really hinder the students' response to a stem. Students actively find meaning of an object by creating conversions of the sign and then

interpreting them (Sáenz-Ludlow & Kadunz, 2016). When decoding language on an exam, students may be trying to use their previous general knowledge to first understand what they are reading. Students who already struggle in reading now are just trying to perceive the problem. Then, the student has to figure out what wording to react to so that they can bring in appropriate background knowledge in order to develop a correct response. Again, cognitive overload can exist in this type of situation. In this instance, students may try to remember different ways that they have decoded word problems in the past and struggle to make the connections with the appropriate algebra required to solve.

Cognition and brain-based research.

These findings may align with previous cognitive studies and brain based research. More words on the *current* exam may increase or decrease cognitive load depending on how wording is used in the stem of the item. This increased wording may also be as a result of the general increase of context on items. Students would need to understand how to interpret the general knowledge along with the algebra specific vocabulary to properly respond to the stem.

In terms of complexity, *old* exam items were often more complex on average than *current* exam items. However, *current* exam items varied more in levels of complexity. This may indicate that students would need to have a variety of approaches for problem solving in *current* exam items than in *old* or *recent* exam items. This increased variability may have significant impacts on working memory. For example, if on *old* exam items, if right triangle trigonometry problems were often written in a context such as an angle of elevation for a lighthouse, students would be trained to focus more on the diagram and applicable numbers with having the context to help the student visualize the "why" of the problem solving process. However, this is not the case with *current* exam items. One exam may have a system of equations with a context in the

constructed response while the next exam may have a system of equations where the student only needs to determine the next step. The type of task for systems of equations is different per exam. When comparing the trigonometry problem with context to the system of equations in context, it would be clear that the context in the system of equation is more involved in the problem solving process. Students need to translate the written word into two symbolic equations and use the context in the system of equations as a constraint to their response. In this case the context may help the student if the student understands the context itself. For example, it is not possible to have a negative number of dogs and cats which would indicate that there was a mistake in solving the system. If the student only provides one solution, he or she may realize that the response is incorrect as there were not two separate responses for both dogs and cats in this system of equations.

With an increase number of representations and connections among representations in items on *current* exam items versus *old* and *recent* exam items, there may be a variety of ways that representations are affecting the learner. Knowledge of these representations becomes more flexible with more experience and practice with the representations. There is more of an ability to compare and contrast representations when students know beyond the initial definition of a single representation. When there are more representations in the stem of an item, a student has to store these in working memory in order to process the problem and eventually come to the correct response. If the student understands how the representations are related, this would decrease cognitive load and help the student derive the correct response. However, if the student does not understand how representations are related, there could be cognitive overload where the student is overwhelmed with the germane load of the stem.

As mathematics becomes harder and different wording, contexts, and representations are used with students, the students who were persistent in practicing mathematics have more flexibility in problem solving and can be more successful than others who did not practice which leads to the Matthew Effect. Successful students have persisted through difficult problem solving contexts before and have found favorable results. This is why they are able to focus and give the incredible amount of working memory necessary to solve complex problems. Students who obtain early success in their mathematical procedures and make connections quicker than other students continue to practice and continue to acquire skills much faster than students who do not make these connections as readily. Struggling students unfortunately endure a longer period of time of frustration, and have a higher likelihood that they will give up before understanding a concept.

As years progress and mathematics becomes more difficult, there can be a divide among the students are able to provide a response in multiple ways and guarantee that they are correct versus students who still cannot solve basic mathematics problems. Students learn what they practice, and practice is what leads to habits. This is why all students (with struggling students in particular) need to develop mathematical endurance so that they guarantee that their rationale is correct by using integrated methods that combine representations. Teachers and students should understand that *current* algebra items require advanced reading comprehension as well many intricate connections among representations. This requires both teachers and students to have flexibility and patience in problem solving to build mathematical endurance.

Are the *current* standards more complex?

Since there is an increase in readability, context, representations, and use of complex representational processes, this could be indicative of an increased complexity of the items from

old to *current* state algebra exams. It could also be indicative of a transformation of how algebra could be better approached with our diverse group of students. The mathematical community may be under the impression that the *current* standards demonstrate the rigor that is needed in order to compete with the international mathematics community.

Yet, when I analyzed the items for complexity, I was not able to obtain a clear indication that *current* item parts are any more difficult than *old* item parts. Instead, I saw an inconsistency with the number of high cognitive demand items per *current* exam as compared to previous exams. This adds on a different level of complexity: the level of the standard is not easily predicable from exam to exam. Students could only have to find out the next step of solving a written systems of equations on one exam administration, but on the next exam administration, there could be a translation of a systems of equations based on a contextual problem. The next exam administration may test this standard by making the students graph a system of equation by having the student convert the equations from a standard form into slope intercept form.

Recommendations

After completing this research, there is a clear need for multifaceted algebra instruction that takes into account the ideas of literacy, contextual factors, and representational fluency and the relationships among representations. A heavy focus on procedural thinking without the necessary conceptual components will not prepare students for success in the *current* algebra exam. Instead, there needs to be more explicit cross curricular and cross unit connections along with an emphasis on how representational signs can aid in a correct interpretation of an object.

More teaching has to be done where teachers make cross curricular connections as representations can contribute to cognitive overload or it can lead to cognitive offloading. For instance, English teachers should be working with Algebra teachers in order to help students

with specific reading comprehension strategies. Reading strategies that work for passages on English exams need to be adapted for word problems in the algebra classroom.

Teaching practices.

There are a variety of teaching practices that would be supported by this research based on the readability, context, complexity, representations, and relationships among representations. This includes improving reading fluency, relating problem solving processes to contexts, changing previous items to meet the complexity of the *current* items, presenting representations in certain orders, and connecting multiple representations together.

Improving reading fluency.

Written representations can either increase or decrease the amount of cognitive load for the reader. Most items were found to have some sort of written description with them, and it was clear that from *old* to *current* exams, written descriptions have been a part of all exams. Students are being asked to communicate more about their mathematical understanding using writing which was found by the increase of written representations in the responses of *current* exam items, as opposed to drawing a graph of filling in a table. This may indicate that using worked examples (Mwangi & Sweller, 1998) as well as creating mathematical stories (Acari, 2003) and combining visuals with the written information (Rau, 2016) could be effective ways to encourage literacy in the mathematics classroom.

"Real world" context.

Since there has been an increase in the relative amount of "real world" context per exam from *old* exams to *current* exams, there needs to be explicit links between the problem solving processes to the real world context represented in the problem (Ruthven, 1994; Boston, 2012). A mention of a familiar context may make the algebra more meaningful for the student if the

connections among the algebraic representations and context are made explicit (Mousavi, Low, & Sweller, 1995).

Cognitive demand.

The findings that cognitive demand may not have increased over the years could be helpful to teachers. These teachers may be able to use items from previous exam versions (*old*, *recent*) or even modify these items so that they have a similar style to the *current* items. Students may be able to practice with items from other standards as well for additional practice. There may be ways in which different algebra topics were solved in the past that may help students with the *current* standards. For instance, a simpler system of equations in the *old* or *recent* exam may help students to understand how systems of equations work without the added complexity of decimals or fractions within the systems of equations in the *current* exam.

Connections among representations.

Although Nathan et al. (2002) found that visual representations should be presented before symbolic and written (verbal) representations, it should be noted that there is definitely a higher concentration of these written representations on the *current* exam as compared to previous exams. Previously, teachers have had difficulty with transferring graph to numerical representations (Cunningham, 2005; Kili*ç*, 2017) and graph to algebraic representations to oral representations (Bal, 2015). Similarly, visual graphs were not always correctly connected with analytic and graphic representations (Hitt, 1998; van den Bosch et al., 2017). This indicates that students need to be able to critically read the stems of items and able to understand what this stem is asking for. They also need to have background knowledge of both the mathematics and context in order to correctly answer (Ernest, 2008b; Radford, Bardini, & Sabena, 2005). With more variability in the complexity of items, an increased relative percentage of context, and

increased readability in the *current* exams as compared to previous exam versions, written representations are an integral part of mathematics instruction. Rather than presenting the problem solving situation and allowing students to possibly misapply routine procedures based on a faulty reading of the item, teachers and students need to develop better strategies in terms of reading the item, using contextual clues to uncover the connections in the item, and then summarizing the problem situation so that the problem situation is more easily understood by the students.

Since the meaning of the representation lies within the learner (Otte, 2011), both teachers and students have to be aware of the body of knowledge that is present in the stem in order to provide the correct representation(s) in the response. In previous studies, it was found that being able to manipulate and see connections among representations led to higher levels of achievement on assessments (Lepak, Wernet, & Ayieko, 2018; Daryaee et al., 2018). Students need to actually understand how these representations connect with one another rather than using trial and error as seen with sixth grade algebra students who experienced lower assessment scores (Rossi Becker & Rivera, 2005) or transferring the rules of one representation to another as seen with middle schools students who were interviewed about their algebra solving abilities (Bieda & Nathan, 2009; Stacey & MacGregor, 1995). With an increased number of problems requiring representational conversions and synergy, it is even more important that students are not only learning about each representation and how it is used, but they are also learning about how the representation is connected with other representations.

Conclusion

Students who are successful in algebra often make the connections among representations because of their strong organization of basic mathematics in the lower grades. They are able to

conceptualize an item in its various representations whether these representations are part of the problem or ones that they create in order to check over their work. Students who struggle may have more difficulty organizing this information together, and therefore, would need more explicit instruction using cross unit connections. Teachers should be explaining how the different units relate with one another. For example, the process of graphing a line is similar to graphing a parabola. The difference is the type of function and how the variable is manipulated. Students who struggle may organize the concept of lines as separate from parabolas in all instances and therefore incur cognitive overload. Students need to be able to comfortably utilize multiple representations in their thinking. Elementary and middle school mathematics have now included multiple representations in both the curriculum and assessments for these grades. The reason why these representations were included in these years is to better equip students for the rigors of algebraic thinking in the future.

In the future, my hope is that the representational framework is used to identify patterns and trends in representations. This would allow the mathematical community to define mathematical objects so that students of mathematics are able to use these representations in the future. By analyzing the readability, context, complexity, representations, and representational relationships of other branches of mathematics, consistent definitions of signs and objects can be understood in order to decrease the cognitive load for all mathematicians involved.

References

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

Aiken, L. R. (1972). Language factors in learning mathematics. *Review of Educational Research, 42*(3), 359-385. Retrieved from https://files.eric.ed.gov/fulltext/ED068340.pdf

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16*(3), 183-198. https://doi.org/10.1016/j.learninstruc.2006.03.001

Ainsworth, S. (1999). The functions of multiple representations. *Computers & Education, 33*(2-3), 131-152. https://doi.org/10.1016/S0360-1315(99)00029-9

Algebra as a strand of school mathematics for all students. (2014, April). Retrieved from https://www.nctm.org/uploadedFiles/Standards_and_Positions/Position_Statements/Algebra_2014-04.pdf

Amit, M., & Fried, M. N. (2005). Multiple representations in 8th grade algebra lessons: Are learners really getting it. *Proceedings of the 29th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 57-63). Melbourne, Australia: Psychology of Mathematics Education.

Anderson, L. W., Krathwohl, D. R. & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (Complete ed.). New York: Longman.

Andrade, H. L. (2010). Students as the definitive source of formative assessment: Academic self assessment and the self-regulation of learning. *AERA Conference Proceedings.* Retrieved from http://opencommons.uconn.edu/nera_2010125

Arcari, A. (2003). The role of visual representations in the learning of mathematics. *Educational Studies in Mathematics, 52*, 215-241. Retrieved from https://link.springer.com/content/pdf/10.1023/A:1024312321077.pdf

Arzarello, F. (2006). Semiosis as a multimodal process. *Relime, Número Especial,* 267-299.

Arzarello, F., Kenett, R. S., Robutti, O., Shafrir, U., Prodromou, T., & Carante, P. (2015). Teaching and Assessing with New Methodological Tools (MERLO): a New Paradigm? *International Conference on Barriers and Enablers to Learning Maths: Enhancing Learning and Teaching for All Learners*. Institute of Mathematics & Its Applications.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research, 70*(2), 181-214.

ATOS text analyzer. (2020). Retrieved from http://www1.renaissance.com/Products/Accelerated-Reader/ATOS/ATOS-Analyzer-for-Text?_ga=2.216123541.2001539103.1562206383-116860732.1562206383

Bal, A. P. (2015). Skills of using and transform multiple representations of the prospective teachers. In 7th World Conference on Educational Sciences (WCES-2015), 05-07 February 2015, Novotel Athens Convention Center, Athens, Greece. doi:10.106/j.sbspro/2015.07.197

Bal, A. P. (2014). The examination of representations used by classroom teacher candidates in solving mathematical problems. *Educational Sciences: Theory & Practice, 14*(6), 2349-2365. doi: 10.12738/estp.2014.6.2189

Bandeira de Mello, V. (2011). *Mapping state proficiency standards onto the NAEP Scales: Variation and change in state standards for reading and mathematics, 2005-2009*. Retrieved from Washington, D.C.: United States Department of Education.

Bandeira de Mello, V., Bohrnstedt, G., Blankenship, C., & Sherman, D. (2015). *Mapping state proficiency standards onto NAEP Scales: Results from the 2013 NAEP reading and mathematics assessments*. Retrieved from Washington, DC: United States Department of Education.

Bandeira de Mello, V., Rahman, T., & Park, B. J. (2018). *Mapping state proficiency standards onto NAEP Scales: Results from the 2015 NAEP Reading and Mathematics Assessments* Retrieved from Washington, DC: United States Department of Education.

Banerjee, R. & Subramaniam, K. (2012, July). Evolution of a teaching approach for beginning algebra. *Educational Studies in Mathematics, 80*(3), 351-367. Retrieved from https://www.jstor.org/stable/41485986

Barbosa, A. & Vale, I. (2015). Visualization in pattern generalization: Potential and challenges. *Journal of the European Teacher Education Network, 10*, 57-70.

Barbosa, A. Vale, I., & Palhares, P. (2009, December). Exploring generalization with visual patterns: Tasks developed with pre-algebra students. In I. Vale & A. Barbosa, (Eds.), *Patterns: Multiple perspectives and contexts in mathematics education.* (pp. 137-150). Escola Superior de Educação do Instituto Politécnico de Viana do Castelo - Projecto Padrões.

Bardini, C. & Pierce, R. (2015). Assumed mathematics knowledge: The challenge of symbols. *International Journal of Innovation in Science and Mathematics Education, 23*(1), 1-9.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan., A. … Tsai, Y. (2010, March). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133-180. Retrieved from https://www.jstor.org/stable/40645421

Becker, J. B. (2007). *Effect of readability on Living Environment Regents exam question success.* (Master's Thesis). Retrieved from https://fisherpub.sjfc.edu/mathcs_etd_masters/5/

Berg, C. V. (2013, February). Enhancing mathematics student teachers' content knowledge: Conversion between semiotic representations. *Proceedings of CERME8 conference*. Retrieved from http://cerme8.metu.edu.tr/wgpapers/WG17/WG17_Berg.pdf

Bergqvist, E., Dyrvold, A., & Österholm, M. (2012). Relating vocabulary in mathematical tasks to aspects of reading and solving. In C. Bergsten, E. Jablonka & M. Raman (Eds.), *Evaluation and comparison of mathematical achievement: Dimensions and perspectives. Proceedings of MADIF 8, the Eighth Mathematics Education Research Seminar, Umeå, January 24-25, 2012* (pp. 61-70). Linköping: SMDF

Bicer, A., Perihan, C., & Lee, Y. (2018). The impact of writing practices on students' mathematical attainment. *International Electronic Journal of Mathematics Education, 13*(3), 305-313. https://doi.org/10.12973/iejme/3922

Bieda, K. N. & Nathan, M. J. (2009). Representational disfluency in algebra: Evidence from student gestures and speech. *ZDM Mathematics Education, 41*(5), 637-650.

Blanton, M. L. & Kaput, J. J. (2005a, February). Helping elementary teachers build mathematical generality into curriculum and instruction. *ZDM Mathematics Education, 37*(1), 34-42. Retrieved from https://doi.org/10.1007/BF02655895

Blanton, M. L. & Kaput, J. J. (2005b, November). Characterizing a classroom practice that promotes algebraic reasoning. *Journal for Research in Mathematics Education, 36*(5), 412-446. Retrieved from https://www.jstor.org/stable/30034944

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals*. New York: Longmans, Green.

Boardman, A. G. & Woodruff, A. L. (2004). Teacher change and "high stakes" assessment: What happens to professional development? *Teaching and Teacher Education, 20,* 545-557. doi:10.1016/j.tate.2004.06.001

Bofah, E. A. T., & Hannula, M. (2011). The case of calculus: comparative look at task representation in textbooks. In *7th Congress of European Research in Mathematics Education* (pp. 9-13). Rzeszów (Poland): University of Rzeszów.

Bonner, P. J. (2006). Transformation of teacher attitude and approach to math instruction through collaborative action research. *Teacher Education Quarterly, 33*(3), 21-44. Retrieved from http://www.caddogap.com

Bossé, M. J., Adu-Gyamfi, K., & Cheetham, M. (2011). Translations among mathematical representations: Teacher beliefs and practices. *International Journal of Mathematics Teaching and Learning, 15*(6), 1-23.

Bossé, M. J. & Faulconer, J. (2008). Learning and assessing mathematics through reading and writing. *School Science and Mathematics, 108*(1), 8-19.

Boston, M. (2012, September). Assessing instructional quality in mathematics. *The Elementary School Journal, 113*(1), 76-104. Retrieved from https://www.jstor.org/stable/10.1086/666387

Boston, M. D. & Smith, M. S. (2009). Transforming secondary mathematics teaching: Increasing the cognitive demands of instructional tasks used in teachers' classrooms. *Journal for Research in Mathematics Education, 40*(2), 119-156. Retrieved from https://www.jstor.org/stable/40539329

Bradstreet, T. E. (1996). Teaching introductory statistics courses so that nonstatisticians experience statistical research. *The American Statistician, 50*(1), 69-78.

Brenner, M. E., Mayer, R.E., Moseley, B., Brar, J., Durán, R., Reed, B. S., & Webb, D. (1997). Learning by understanding: The role of multiple representations in learning algebra. *American Educational Research Journal, 34*(4), 663-689. Retrieved from https://www.jstor.org/stable/pdf/1163353.pdf

Brizuela, B. M. & Lara-Roth, S. (2002). Additive relations and function tables. *Journal of Mathematical Behavior, 20,* 309-319.

Bull, R. & Lee, K. (2014). Executive functioning and mathematics achievement. *Child Development Perspectives, 8*(1), 36-41. http://dx.doi.org/10.1111/cdep.12059

Burton, L. & Morgan, C. (2000, July). Mathematicians writing. *Journal for Research in Mathematics Education, 31*(4), 429-453. Retrieved from https://www.jstor.org/stable/749652

Bussi, M. B., & Mariotti, M. A. (2008). Semiotic mediation in the mathematics classroom: Artifacts and signs after a Vygotskian perspective. *Handbook of international research in mathematics education, New York,* 746-783.

Cai, J., Nie, B., & Moyer, J. C. (2010). The teaching of equation solving: Approaches in Standards-based and traditional curricula in the United States. *Pedagogies: An International Journal*, *5*(3), 170-186.

Cai, J., Moyer, J.C., Nie, B., & Wang, N. (2009). Learning mathematics from classroom instruction using standards-based and traditional curricula: An analysis of instructional tasks. In S. L. Swars, D. W. Stinson, & S. Lemons-Smith, (Eds.) *Proceedings of the 31st Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education.* Atlanta, GA: Georgia State University.

Calhoon, M. B., Emerson, R. W., Flores, M., & Houchins, D. E. (2007). Computational fluency performance profile of high school students with mathematics disabilities. *Remedial and Special Education, 28*(5), 292-303. doi: 10.1177/07419325070280050401

Caponera, E., Sestito, P., & Russo, P. M. (2016). The influence of reading literacy on mathematics and science achievement. *The Journal of Educational Research, 109*(2), 197-204. http://dx.doi.org/10.1080/00220671.2014.936998

Care, E. & Griffin, P. (2014). An approach to assessment of collaborative problem solving. *Research and Practice in Technology Enhanced Learning, 9*(3), 367-388.

Carraher, D. W., Martinez, M. V., & Schliemann, A. D. (2008). Early algebra and mathematical generalization. *ZDM Mathematics Education, 40,* 3-22. doi: 10.1007/s11858-007-0067-7

Carraher, D., Schliemann, A. D., & Brizuela, B. M. (2000). *Early algebra, early arithmetic: Treating operations as functions.* Presented at PME-NA XXII, Tucson, AZ.

Carraher, D. W., Schliemann, A. D., Brizuela, B. M., & Earnest, D. (2006). Arithmetic and algebra in early mathematics education. *Journal for Research in Mathematics Education,* 37(2), 87-115. Retrieved from https://www.jstor.org/stable/30034843

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*(8), 723-733.

Carroll, W. M. (1998). Geometric knowledge of middle school students in a reform-based mathematics curriculum. *School Science and Mathematics, 98*(4), 188-197.

Carter, J. (2018). Graph-algebras - Faithful representations and mediating objects in mathematics. *Endeavour.* https://doi.org/10.1016/j.endeavour.2018.07.006

Ceuppens, S., Deprez, J., Dehaene, W., & De Cock, M. (2018). Design and validation of a test for representational fluency of 9th grade students in physics and mathematics: The case of linear functions. *Physical Review Physics Education Research, 14*(020105), 1-19. Retrieved from https://journals.aps.org/prper/pdf/10.1103/PhysRevPhysEducRes.14.020105

Charalambous, C. Y. (2008). Mathematical knowledge for teaching and the unfolding of tasks in mathematics lessons: Integrating two lines of research. In *Proceedings of the 32nd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 281-288).

Charalambous, C. Y. & Philippou, G. N. (2010, September). Teachers' concerns and efficacy beliefs about implementing a mathematics curriculum reform: Integrating two lines of inquiry. *Educational Studies in Mathematics, 75*(1), 1-21. https://www.jstor.org/stable/40928544

Chen, J. & Lai, Y. (2015). A brief review of researches on the use of graphing calculator in mathematics classrooms. *International Journal of Learning, Teaching, and Educational Research, 14*(2), 163-172.

Cho, H., & Kwon, O. N. (2012). An Analysis of the Mathematical Tasks in the Korean Seventh Grade Mathematics Textbooks and Workbook. *Research in Mathematical Education, 16*(4), 245-263.

Clarke, D. J., Waywood, A., & Stephens, M. (1993). Probing the structure of mathematical writing. *Educational Studies in Mathematics, 25*(3), 235-250. Retrieved from https://www.jstor.org/stable/3482735

Coburn, C. E., Hill, H. C., & Spillane, J. P. (2016). Alignment and accountability in policy design and implementation: The Common Core State Standards and implementation research. *Educational Researcher, 45*(4), 243-251.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Common Core State Standards Initative. (2020). "Do the Common Core State Standards incorporate both content and skills?" Retrieved from http://www.corestandards.org/faq/do-the-common-core-state-standards-incorporate-both-content-and-skills/

Corder, G. W., & Foreman, D. I. (2009). *Nonparametric statistics for non-statisticians: A step-by-step approach.* Hoboken, N.J: Wiley.

Crowley, M. L. (1987). "The van Hiele model of the development of geometric thought." In *Learning and Teaching Geometry, K-12,* 1987 Yearbook of the National Council of Teachers of Mathematics, Mary Montgomery Lindquist, 1-16. Reston, VA: National Council of Teachers of Mathematics.

Cunningham, R. F. (2005, February). Algebra teachers' utilization of problems requiring transfer between algebraic, numeric, and graphic representations. *School Science and Mathematics, 105*(2), 73-81.

Czekanski, S. (2018). A vocabulary analysis of the New York State 2017 mathematics assessment constructed response questions (Grades 6-8). (Master's thesis). Retrieved from https://dspace.sunyconnect.suny.edu/bitstream/handle/1951/70445/Czekanski_Shana_MastersProject_Spring2018.pdf?sequence=1

Daryaee, A., Shahvarani, A., Tehranian, A., Lotfi, F. H., & Rostamy-Malkhalifeh, M. (2018). Executable functions of the representations in learning the algebraic concepts. *PNA, 13*(1), 1-18.

De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science, 38,* 105-134. doi: 10.1007/s11251-009-9110-0

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44,* 1-42.

Dehaene, S., & Akhavein, R. (1995). Attention, automaticity, and levels of representation in number processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(2), 314–326. https://doi.org/10.1037/0278-7393.21.2.314

Dehaene, S., & Cohen, L. (1995). Towards an anatomical and functional model of number processing. *Mathematical Cognition, 1,* 83–120.

Delaney, S., Charalambous, C. Y., Hsu, H., & Mesa, V. (2007). The treatment of addition and subtraction of fractions in Cypriot, Irish, and Taiwanese textbooks. *Proceedings of the 31st Conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 193-200). Retrieved from https://files.eric.ed.gov/fulltext/ED499417.pdf#page=199

Dewolf, T., Van Dooren, W., Hermens, F., & Verschaffel, L. (2015). Do students attend to the representational illustrations of non-standard mathematical word problems, and, if so, how helpful are they? *Instructional Science, 43*(1), 147-171. doi: 10.1007/s11251-014-9332-7

Doig, B. & Groves, S. (1999). Putting meaning behind bars: Children's interpretations of bar graphs. Paper presented at the combined *Annual Meeting of the Australian Association for Research in Education and the New Zealand Association for Research in Education*: Melbourne, Australia.

Drake, C. & Sherin, M. G. (2006). Practicing change: Curriculum adaptation and teacher narrative in the context of mathematics education reform. *Curriculum Inquiry, 36*(2), 153-187. Retrieved from https://www.jstor.org/stable/3698503

Drake, P., Wake, G., & Noyes, A. (2012). Assessing 'functionality' in school mathematics examinations: What does being human have to do with it? *Research in Mathematics Education, 14,* 237–252.

Draper, R. J. & Siebert, D. (2004). Different goals, similar practices: Making sense of the mathematics and literacy instruction in a standards-based mathematics classroom. *American Educational Research Journal, 41*(4), 927-962.

Dubinsky, E. & Wilson, R. T. (2013). High school students' understanding of the function concept. *The Journal of Mathematical Behavior, 32*(1), 83-101.

Duff, D., Tomblin, J. B. & Catts, H. (2015). The influence of reading on vocabulary growth: A case for a Matthew Effect. *Journal of Speech, Language, and Hearing Research, 58*(3), 853-864. doi: 10.1044/2015 JSLHR-L-13-0310

Dündar, S. (2016, January). Does writing have any effect on mathematical success? *Journal of Education and Training Studies, 4*(1), 1-10. Retrieved from http://www.redfame.com/journal/index.php/jets/article/viewFile/989/1056

Duval, R. (2006). A cognitive analysis of problems of comprehension in a learning of mathematics. *Educational Studies in Mathematics, 61*(1-2), 103-131. https://www.jstor.org/stable/2542062

Dyrvold, A. (2016). *Difficult to read or difficult to solve?: The role of natural language and other semiotic resources in mathematical tasks* (Doctoral dissertation). Umeå Universitet.

Elia, I. & Philippou, G. (2004). The functions of pictures in problem solving. In Høines, M. J. & Fuglestad, A. B. (Ed.), *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education,* Bergen, Norway: Bergen University College.

Ernest, P. (2006). A semiotic perspective of mathematical activity: The case of number. *Educational Studies in Mathematics, 61*, 67-101. doi: 10.1007/s10649-006-6423-7

Ernest, P. (2008a). Towards a semiotics of mathematical text (Part 2). *For the Learning of Mathematics, 28*(2), 39-47. Retrieved from https://www.jstor.org/stable/40248607

Ernest, P. (2008b). Towards a semiotics of mathematical text (Part 3). *For the Learning of Mathematics, 28*(3), 42-49. Retrieved from https://www.jstor.org/stable/40248628

Etkind, M., Kenett, R. S., & Shafrir, U. (2010). The evidence-based management of learning: Diagnosis and development of conceptual thinking with meaning equivalence reusable learning objects (MERLO). In *Invited paper. Proceedings of the 8th International Conference on Teaching Statistics (ICOTS8).* Ljubljana, Slovenia.

Fan, L. & Zhu, L. (2008, December). Using performance assessment in secondary school mathematics: An empirical study in a Singapore classroom. *Journal of Mathematics Education, 1*(1), 132-152.

Fasano, D. D. (2014). *Redefining algebra: The recent history of policies and politics of the algebra curriculum in New York State and its instructional implications.* (Doctoral dissertation). St. John's University, New York, NY.

Ferrini-Mundy, J., Floden, R. E., McCrory, R., Burrill, G., & Sandow, D. (2005). *Knowledge for teaching school algebra: Challenges in developing an analytic framework.* East Lansing, MI: Michigan State University, Knowledge of Algebra for Teaching Project.

Fischbein, E., Jehiam, R., & Cohen, D. (1995). The concept of irrational numbers in high-school students and prospective teachers. *Educational Studies in Mathematics, 29*(1), 29-44. Retrieved from https://www.jstor.org/stable/3482830

Fisher-Hoch, H. & Hughes, S. (1996). *What makes mathematics exam questions difficult?* Paper presented at the 1996 British Educational Research Association (BERA) Conference, Lancaster, United Kingdom.

Flesch, R. F. (1979). *How to write plain English: A book for lawyers and consumers.* Michigan: Harper & Row.

Font, V., Bolite, J., & Acevedo, J. (2010, November). Metaphors in mathematics classrooms: Analyzing the dynamic process of teaching and learning of graph functions. *Educational Studies in Mathematics, 75*(2), 131-152. https://www.jstor.org/stable/40928552

Font, V., Godino, J. D., & Gallardo, J. (2013, January). The emergence of objects from mathematical practices. *Educational Studies in Mathematics, 82*(1), 97-124. https://www.jstor.org/stable/23434841

Freitag, M. (1997). Reading and writing in the mathematics classroom. *The Mathematics Educator, 18*(1), 16-21.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education, 32*(2), 124-158.

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Hamlett, C. L., Seethaler, P. M., … Schatschneider, C. (2010). Do different types of school mathematics development depend on different constellations of numerical versus general cognitive abilities? *Developmental Psychology, 46*(6): 1731-1746.

Gabriele, A. J. & Joram, E. (2007). Teachers' reflections on their reform-based teaching in mathematics: Implications for the development of teacher self-efficacy. *Action in Teacher Education, 29*(3), 60-74. doi: 10.1080/01626620.2007.10463461

Gamson, D. A., Lu, X., & Eckert, S. A. (2013). Challenging the research base of the common core state standards: A historical reanalysis of text complexity. *Educational Researcher, 42*(7), 381-391. Retrieved from https://www.jstor.org/stable/24571234

Garfield, J. & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice.* New York: Springer.

Garfield, J. & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistics Review, 75*(3), 372-396. doi: 10.1111/j.1751-5823.2007.00029.x

Germain-Williams, T. (2014). Mathematical modeling in algebra textbooks at the onset of the common core state standards (Doctoral dissertation, Columbia University). Available from the ProQuest Dissertations and Theses database. Retrieved from http: //eduproxy.tc-library.org/?url=/docview/1537949683?accountid=14258

Giardino, V. (2018). Manipulative imagination: How to move things around in mathematics. *Theoria, 33*(2), 345-360. doi: 10.1387/theoria.17871

Godino, J. D., Batanero, C., & Font, V. (2007). The onto-semiotic approach to research in mathematics education. *ZDM. The International Journal on Mathematics Education, 39*(1-2), 127-135.

Godino, J. D., Font, V., Wilhelmi, M. R., & Lurduy, O. (2011, July). Why is the learning of elementary arithmetic concepts difficult? Semiotic tools for understanding the nature of mathematical objects. *Educational Studies in Mathematics, 77*(2/3), 247-265. https://www.jstor.org/stable/41485928

Gjoci, B. (2017). *Analyzing Mathematics High School State Examinations in Albania in the 1970s and 2006-2015: Two Decades, Two Historical Periods* (Doctoral dissertation, Columbia University).

Greenes, C., Chang, K. Y., & Ben-Chaim, D. (2007). International survey of high school students' understanding of key concepts of linearity. In J. H. Woo, H. C. Lew, K. S. Park, & D. Y. Seo, (Eds.). *Proceedings of the 31st Conference of the International Group for the Psychology of Mathematics Education,* (Vol. 2, pp. 273-280). Seoul, Korea: PME.Greiffenhagen, C. (2014). The materiality of mathematics: Presenting mathematics at the blackboard. *The British Journal of Sociology, 65,* 502–528. doi:10.1111/1468-4446.12037

Grouws, D. A., Tarr, J. E., Chávez, Ó., Sears, R., Soria, V. M., & Taylan, R. D. (2013). Curriculum and implementation effects on high school students' mathematics learning from curricula representing subject-specific and integrated content organizations. *Journal for Research in Mathematics Education, 44*(2), 416-463. Retrieved from https://www.jstor.org/stable/10.5951/jresematheduc.44.2.0416

Guven, B. (2012). Using dynamic geometry software to improve eighth grade students' understanding of transformation geometry. *Australasian Journal of Educational Technology, 28*(2), 364-382. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.888.3025&rep=rep1&type=pdf

Halliday, M. A. K. (1978). *Language as a social semiotic: The social interpretation of language and meaning.* London: Edward Arnold.

Handal, B. & Herrington, A. (2003). Mathematics teachers' beliefs and curriculum reform. *Mathematics Education Research Journal, 15*(1), 59-69.

Harris, L. R. & Brown, G. T. L. (2010). Mixing interview and questionnaire methods: Practical problems in aligning data. *Practical Assessment, Research, & Evaluation, 15*(1), 1-19. Retrieved from repositorio.minedu.gob.pe

Hazzan, O. & Zazkis, R. (2005, January). Reducing abstraction: The case of school mathematics. *Educational Studies in Mathematics, 58,* 101-119.

Heddins, J. W. & Smith, K. J. (1964). The readability of elementary mathematics books. *The Arithmetic Teacher, 11*(7), 466-468. Retrieved from https://www.jstor.org/stable/41186851

Helwig, R., Rozek-Tedesco, M. A., Tindal, G., Heath, B., & Almond, P. J. (1999). Reading as an access to mathematics problem solving on multiple-choice test for sixth-grade students. *The Journal of Educational Research, 93*(2), 113-125. Retrieved from https://www.jstor.org/stable/27542254

Henningsen, M. & Stein, M. K. (1997, November). Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. *Journal for Research in Mathematics Education, 28*(5), 524-549. https://www.jstor.org/stable/749690

Heritage, M. & Niemi, D. (2006). Toward a framework for using student mathematical representations as formative assessments. *Educational Assessment, 11*(3 & 4), 265-282.

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments. *Applied Measurement in Education, 20*(1), 101-126. Retrieved from http://cresst.org/wp-content/uploads/R653.pdf

Hess, K. K., Jones, B. S., Carlock, D., & Walkup, J. R. (2009). Cognitive rigor: Blending the strengths of Bloom's taxonomy and Webb's depth of knowledge to enhance classroom-level processes. Retrieved from ERIC database (ED517804).

Hewitt, M. & Homan, S. P. (2004). Readability of standardized test items and student performance: The forgotten validity variable. *Reading Research and Instruction, 43*(2), 1-16.

Hickendorff, M. (2013). The language factor in elementary mathematics assessments: Computational skills and applied problem solving in a multidimensional IRT framework. *Applied Measurement in Education, 26*, 253-278. doi: 10.1080/08957347.2013.824451

Hiebert, J., & Carpenter, T. P. (1992). Learning and teaching with understanding. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 65-97). New York, NY, England: Macmillan Publishing Co, Inc.

Hiebert, J. & Wearne, D. (1983). *Students' conceptions of decimal numbers.* (Grant SPE-8218387). Washington: National Science Foundation.

Hitt, F. (1998). Difficulties in the articulation of different representations linked to the concept of function. *Journal of Mathematical Behavior, 17*(1), 123-134. ISSN: 0364-0213

Holmes, V. (2012). Depth of teachers' knowledge: Frameworks for teachers' knowledge of mathematics. *Journal of STEM Education, 13*(1), 55-71.

Homan, S., Hewitt, M., & Linder, J. (1994). The development and validation of a formula for measuring single-sentence test item readability. *Journal of Educational Measurement, 31*(4), 349-358.

Horz, H., & Schnotz, W. (2009). Cognitive load in learning with multiple representations. In J. Plass, R. Moreno, & R. Brunken (Eds.), Cognitive load theory (pp. xx–xx). New York: Cambridge University Press.

Hudson, P., Miller, S. P., & Butler, F. (2006). Adapting and merging explicit instruction within reform based mathematics classroom. *American Secondary Education, 35*(1), 19-32. Retrieved from https://search.proquest.com/docview/195185376?accountid=60358

Iori, M. (2017). Objects, signs, and representations in the semio-cognitive analysis of the processes involved in teaching and learning mathematics: A Duvalian perspective. *Educational Studies in Mathematics, 94*, 275-291. doi: 10.1007/s10649-016-9726-3

Iori, M. (2018). Teachers' awareness of the semio-cognitive dimension of learning mathematics. *Educational Studies in Mathematics, 98,* 95-113. https://doi.org/10.1007/s10649-018-9808-5

Jacobs, J. K., Hiebert, J., Givvin, K. B., Hollingsworth, H., Garnier, H., & Wearne, D. (2006, January). Does eighth-grade mathematics teaching in the United States align with the NCTM standards? Results from the TIMSS 1995 and 1999 video studies. *Journal for Research in Mathematics Education, 37*(1), 5-32. Retrieved from https://www.jstor.org/stable/30035050

Johnson, D. A. (1957). The readability of mathematics books. *The Mathematics Teacher, 50*(2), 105-110. Retrieved from https://www.jstor.org/stable/2795536

Jones, D. L. & Tarr, J. E. (2007). An examination of the levels of cognitive demand required by probability tasks in middle grades mathematics textbooks. *Statistics Education Research Journal, 6*(2), 4-27. Retrieved from http://www.stat.auckland.ac.nz/serj

Jones, M. (2006). Demystifying functions: The historical and pedagogical difficulties of the concept of the function. *Rose-Hulman Undergraduate Mathematics Journal, 7*(2), 1-20. Retrieved from https://scholar.rose-hulman.edu/rhumj/vol7/iss2/5

Kane, R. B. (1970). The readability of mathematics textbooks revisited. *The Mathematics Teacher, 63*(7), 579-581. Retrieved from https://www.jstor.org/stable/27958458

Katz, V. J. & Barton, B. (2007). Stages in the history of algebra with implications for teaching. *Educational Studies in Mathematics, 66*(2), 185-201. Retrieved from https://www.jstor.org/stable/27822699

Kerckhoff, A. C. & Glennie, E. (1999). The Matthew effect in American education. *Research in Sociology of Education and Socialization, 12*, 35-66.

Kiliç, Ç. (2017). The ability of pre-service primary teachers to produce figural patterns based on algebraic formulas. *Turkish Journal of Computer and Mathematics Education, 8*(2), 261-283.

Kim, H., Fukawa-Connelly, T., & Cook, S. A. (2012). Student understanding of symbols in introductory statistics. In Brown, S., Larsen, S., Marrongelle, K., & Oehrtman, M. (EDs.)*15th annual conference on research in undergraduate mathematics education* (pp. 285-294). Portland, Oregon.

Kincaid, J. P., Fisburne, R. P. Jr., Rogers, R. L., & Chissom, B. S. (1975, February). Derivation of readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. *Research Branch Report 8-75*. Millington, TN: Naval Training, U.S. Naval Air Station, Memphis, TN.

Koedinger, K. R. & Terao, A. (2002). A cognitive task analysis of using pictures to support pre-algebraic reasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society, 24*(24). Retrieved from https://escholarship.org/uc/item/2649r1xt

Kirshner, D. & Awtry, T. (2004, July). Visual salience of algebraic transformations. *Journal for Research in Mathematics Education, 35*(4), 224-257. Retrieved from https://www.jstor.org/stable/30034809

Kralemann, B. & Lattmann, C. (2013, November). Models as icons: Modeling models in the semiotic framework of Peirce's theory of signs. *Synthese, 190*(16), 3397-3420. Retrieved from https://www.jstor.org/stable/24019866

Kurz, A., Elliott, S. N., Wehby, J. H., & Smithson, J. L. (2010). Alignment of the intended, planned, and enacted curriculum in general and special education and its relation to student achievement. *Journal of Special Education, 44*(3), 131-145. doi: 10.1177/0022466909341196

Kuzle, A. (2013). Promoting writing in mathematics: Prospective teachers' experiences and perspectives on the process of writing when doing mathematics as problem solving. *CEPS Journal, 3*(4), 41-59. Retrieved from https://www.pedocs.de/volltexte/2014/8500/pdf/cepsj_2013_4_Kuzle_Promoting_writing_in_mathematics.pdf

Lamb, J. H. (2010). Reading grade levels and mathematics assessment: An analysis of Texas mathematics assessment items and their reading difficulty. *The Mathematics Educator, 20*(1), 22-34.

Landy, D., Brookes, D., & Smout, R. (2014). Abstract numeric relations and the visual structure of algebra. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(5), 1404-1418. http://dx.doi.org/ 10.1037/a0036823

Landy, D., & Goldstone, R. L. (2007). How abstract is symbolic thought? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 720–733. http://dx.doi.org/10.1037/0278-7393.33.4.720

Lee, K., Ng, S. F., & Bull, R. (2017). Learning and solving more complex problems: The roles of working memory, updating, and prior skills for general mathematics achievement and algebra. In D. C. Geary, D. B. Berch, R. Ochsendorf, & K. M. Koepke (Eds.), *Acquisition of complex arithmetic skills and higher-order mathematics concepts* (pp. 197-220). New York: Academic Press.

Lee, J., & Reeves, T. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity, and resources: State NEAP 1990-2009 Reading and Math achievement gaps and trends. *Educational Evaluation and Policy Analysis, 34*(209), 209-231. doi: 10.3102/0162373711431604

Lehrer, R., Kim, M., & Schauble, L. (2007). Supporting the development of conceptions of statistics by engaging students in measuring and modeling variability. *International Journal of Computer Mathematics Learning, 12*, 195-216. doi: 10.1007/s10758-007-9122-2

Leinhardt, G., Zaslavsky, O., & Stein, M. K. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of Educational Research, 60*(1), 1-64. Retrieved from https://www.jstor.org/stable/1170224

Leiss, D., Schukajlow, S., Blum, W., Messner, R., & Pekrun, R. (2010). The role of the situation model in mathematical modelling - Task analyses, student competencies, and teacher interventions. *Journal für Mathematik-Didaktik, 31,* 119-141. doi: 10.1007/s13138-010-0006-y

Lemke, J. L. (2003). Mathematics in the middle: Measure, picture, gesture, sign, and word. *Educational perspectives on mathematics as semiosis: From thinking to interpreting to knowing, 1,* 215-234.

Lennon, C. & Burdick, H. (2004). The lexile framework as an approach for reading measurement and success. (The Lexile Framework for Reading white paper). Durham, NC: MetaMetrics. Retrieved from https://web.archive.org/web/20100103044012/http://lexile.com/m/resources/materials/Lennon__Burdick_2004.pdf

Lepak, J. R., Wernet, J. L. W., & Ayieko, R. A. (2018, June). Capturing and characterizing students' strategic algebraic reasoning through cognitively demanding tasks with focus on representations. *The Journal of Mathematical Behavior, 50,* 57-73. https://doi.org/10.1016/j.jmathb.2018.01.003

Lepik, M. (1990). Algebraic word problems: Role of linguistic and structural variables. *Educational Studies in Mathematics, 21,* 83-90. Retrieved from https://www.jstor.org/stable/3482220

Lian, L. H. & Yew, W. T. (2012). Assessing algebraic solving ability: A theoretical framework. *International Education Studies, 5*(6), 177-188. https://dx.doi.org/10.5539/ies.v5n6p177

Lianghuo, F. (2004, October 16). Improving mathematics teaching and learning through effective classroom assessment: Experiences and perspectives from Singapore schools. In *The Association of Mathematics Education, Singapore & The Department of Mathematical Science, Singapore Polytechnic*. Singapore: National Institute of Education, Nanyany Technological University.

Lloyd, G. M. & Behm, S. L. (2005). Preservice elementary teachers analysis of mathematics instructional materials. *Action in Teacher Education, 26*(4), 48-62.

MacGregor, M. & Stacey, K. (1997, June). Students' understanding of algebraic notation: 11-15. *Educational Studies in Mathematics, 33*(1), 1-19. Retrieved from https://www.jstor.org/stable/3483002

MacKenzie, D. & Barany, M. (2014). Chalk: Materials and concepts in mathematics research. In C Coopmans, M Lynch, J Vertesi & S Woolgar (eds), *Representation in Scientific Practice Revisited.* MIT Press, Cambridge, MA. doi: 10.7551/mitpress/9780262525381.001.0001

Maffia, A., & Mariotti, M. A. (2015, February). Introduction to arithmetical expressions: a semiotic perspective. In *CERME 9-Ninth Congress of the European Society for Research in Mathematics Education* (pp. 1947-1953).

Maffia, A. & Sabena, C. (2015). Networking of theories as resource for classroom activities analysis: The emergence of multimodal semiotic chains. In C. Sabena, B. Di Paola (eds.), *Teaching and learning mathematics: Resources and obstacles, Proceedings of CIEAEM 67, Quaderni di ricerca didattica, 25-2* (pp. 405-417). Aosta, July 20-24, 2015. ISSN: 1592-4424

Major, T. E. & Mangope, B. (2012). The constructivist theory in mathematics: The case of Botswana primary schools. *International Review of Social Studies and Humanities, 3*(2), 139-147. ISSN 2250-0715

Martin, D. B., Gholson, M. L., & Leonard, J. (2010). Mathematics as gatekeeper: Power and privilege in the production of knowledge. *Journal of Urban Mathematics Education, 3*(2), 12-24. Retrieved from http://education.gsu.edu/JUME

Martin, K. R. (2018). Teacher and student discourse and vocabulary in secondary mathematics classrooms during whole class discussions. (Doctoral dissertation). University at Buffalo, Buffalo, New York.

McCaffrey, D. F., Hamilton, L. S., Stecher, B. M., Klein, S. P, Bugliari, D., & Robyn, A. (2001, November). Interactions among instructional practices, curriculum, and student achievement: The case of standards-based high school mathematics. *Journal for Research in Mathematics Education, 32*(5), 493-517. Retrieved from https://www.jstor.org/stable/749803

McGee, D. L. & Moore-Russo, M. (2015). Impact of explicit presentation of slopes in three dimensions on students' understanding of derivatives in multivariable calculus. *International Journal of Science and Mathematics Education, 13*(2), 357-384.

McLaughlin, H. (1969). SMOG grading - a new readability formula. *Journal of Reading, 22*, 639-646.

Meltzer, D. E. (2005). Relation between students' problem-solving performance and representational format. *American Journal of Physics, 73*(5), 463-478. https://doi.org/10.1119/1.1862636

Merton, R. K. (1968, January 5). The Matthew effect in science. *Science, New Series, 159*(3810), 56-63. Retrieved from http://links.jstor.org/sici?sici=0036-8075%2819680105%293%3A159%3A3810%3C56%3ATMEIS%3E2.0.CO%3B2-K

Mesa, V. (2004). Characterizing practices associated with functions in middle school textbooks: An empirical approach. *Educational Studies in Mathematics, 56,* 255-286.

Mevarech, Z. R. (1983). A deep structure model of students' statistical misconceptions. *Educational Studies in Mathematics, 14*(4), 415-429. Retrieved from https://www.jstor.org/stable/3482437

Mistretta, R. M. (2000). Enhancing geometric reasoning. *Adolescence, 35*(138), 365-379.

Moll, K., Göbel, S. M., & Snowling, M. J. (2015). Basic number processing in children with specific learning disorders: Comorbidity of reading and mathematics disorders. *Child Neuropsychology, 21,* 399-417. doi: 10.1080/09297049.2014.899570

Monroe, E. E. & Pendergrass, M. R. (1997). Effects of mathematical vocabulary instruction on fourth grade students. In *BYU/Public School Partnership Symposium on Education, Provo, UT.*

Moore, K. C. & Thompson, P. W. (2015). Shape thinking and students' graphing activity. In T. Fukawa-Connelly, N. Infante, K. Keene, & M. Zandieh (Eds.), *Proceeedings of the Eighteenth Annual Conference on RUME* (p. 782-789). Pittsburgh, PA: WVU.

Moore, K. C. (2016). Graphing as figurative and operative thought. In Csíkos, C., Rausch, A., & Szitányi, J. (Eds.). *Proceedings of the 40th Conference of the International Groups for the Psychology of Mathematics Education,* Vol. 3, pp. 323-330. Szeged, Hungary: PME.

Moore-Russo, D. M. & Viglietti, J. M. (2012). Using the K_5 *connected cognition diagram* to analyze teachers' communication and understanding of regions in three-dimensional space. *The Journal of Mathematical Behavior, 31*, 235-251. Retrieved from www.elsevier.com/locate/jmathb

Moreno, R. & Durán, R. (2004). Do multiple representations need explanations? The role of verbal guidance and individual differences in multimedia mathematics learning. *Journal of Educational Psychology, 96*(3), 492-503. doi: 10.1037/0022-0663.96.3.492

Morgan, C. (1996). "The language of mathematics": Towards a critical analysis of mathematics texts. *For the Learning of Mathematics, 16*(3), 2-10.

Morgan, C. (2006). What does social semiotics have to offer mathematics education research? *Educational Studies in Mathematics, 61*(½), 219-245. Retrieved from https://www.jstor.org/stable/25472066

Morgan, C., & Alshwaikh, J. (2009). Mathematical activity in a multi-semiotic environment. In *Sixth Congress of the European Society for Research in Mathematics Education, Working Group* (Vol. 6, pp. 993-1002).

Morgan, C. & Sfard, A. (2016). Investigating changes in high-stakes mathematics examinations: A discursive approach. *Research in Mathematics Education, 18*(2), 93-119. http://dx.doi.org/10.1080/14794802.2016.1176596

Morgan, C. & Tang, S. (2016, July). To what extent are students expected to participate in specialised mathematical discourse? Change over time in school mathematics in England. *Research in Mathematics Education, 18*(2), 142-164. https://doi.org/10.1080/14794802.2016.1174145

Moritz, J., Meyerhoff, H. S., Meyer-Derbecher, C., & Schwan, S. (2018). Representation control increases task efficiency in complex graphical representations. *PLoS One, 13*(4), 1-14. https://doi.org/10.1371/journal.pone.0196420

Morton, D. E. (2015). *Readability of the New York State Regents exam in United States history and government.* (Master's thesis). Retrieved from https://dspace.sunyconnect.suny.edu/bitstream/handle/1951/65753/Morton_Daniel_Masters_Spring2015.pdf?sequence=1&isAllowed=y

Moschkovich, J. (2012, January). Mathematcs, the common core, and language: Recommendations for mathematics instruction for ELs aligned with the common core. Paper presented at the Understanding Language Conference, Stanford, CA. Retrieved from http://ell.stanford.edu/papers/language

Moseley, B. & Brenner, M. E. (1997). Using multiple representations for conceptual change in pre-algebra: A comparison of variable usage with graphic and text based problems. Office of Educational Research and Improvement (ED), Washington, D.C.

Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology, 87*(2), 319-334.

Mwangi, W. & Sweller, J. (1998). Learning to solve compare word problems: The effect of example format and generating self-explanations. *Cognition and Instruction, 16*(2), 173-199. Retrieved from https://www.jstor.org/stable/3233720

Nathan, M. J., Stephens, A. C., Masarik, D. K., Alibali, M. W., & Koedinger, K. R. (2002). Representational fluency in middle school: A classroom study. In *Proceedings of the twenty-fourth annual meeting of the North American chapter of the International Group for the Psychology of Mathematics Education* (Vol. 1, pp. 462-472). Columbus, OH: ERIC Clearinghouse for Science, Mathematics and Environmental Education.

National Commission on Excellence in Education (NCEE). (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Department of Education.

National Council for Teachers of Mathematics (NCTM). (1989). *Commission on Standards for School Mathematics. Curriculum and evaluation standards for school mathematics*. Reston, VA: NCTM.

National Council of Teachers of Mathematics (NCTM). (2000). *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA: NCTM.

National Council of Teachers of Mathematics. (2014). *Principles to actions: Ensuring mathematical success for all*. Reston, VA: NCTM, National Council of Teachers of Mathematics.

Nelson, E. A. (2017). Cognitive science and the Common Core mathematics standards. *Nonpartisan Education Review/Articles*, *13*(3), 1-19. Retrieved from http://nonpartisaneducation.org/Review/Articles/v13n3.pdf

Neria, D. & Amit, M. (2004). Students preference of non-algebraic representations in mathematical communication. In Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education, (vol. 3), p. 409-416.

Neuhoff, E., Feeser, K. M., Sutherland, K., & Hovatter, T. (2016). Flesh-Kincaid reading grade level re-examined: Creating a uniform method for calculating readabilty on a certification exam. *Journal for Workforce Education and Development*, *9*(1), 2-13. Retrieved from https://opensiuc.lib.siu.edu/cgi/viewcontent.cgi?article=1189&context=ojwed

"New research on text complexity". (2017). Retrieved from http://www.corestandards.org/wp-content/uploads/Appendix-A-New-Research-on-Text-Complexity.pdf

Nistal, A. A., Van Dooren, W., Clarebout, G., Elen, J., & Verschaffel, L. (2009). Conceptualising, investigating, and simulating representational flexibility in mathematical problem solving and learning: A critical review. *ZDM Mathematics Education, 41*, 627-636. doi: 10.1007/s11858-009-0189-1

Nizaruddin, M. & Murtianto, Y. H. (2017). Exploring of multimathematical representation capability in problem solving on senior high school students. *Problems of Education in the 21st Century*, *75*(6), 591-598. Retrieved from oaji.net/articles/2017/457-1513710677.pdf

Norton, A. (2014). In a search of a common core for algebra. In L. P. Steffe, K. C. Moore, & L. L. Hatfield (Eds.), *Epistemic algebra students: Emerging models of students' algebraic knowing (Vol. 4)* (pp. 127-133). WISDOM^{e} Monographs, Laramie, WY: University of Wyoming.

Notari-Syverson, A. & Sadler, F. H. (2008). Math is for everyone: Strategies for supporting early mathematical competencies in young children. *Young Exceptional Children, 11*(3), 2-16. doi: 10.1177/1096250608314589

No Child Left Behind Act [NCLB]. (2002). P.L. 107-110, 20 U.S.C. § 6319. Washington, D.C.: United States Department of Education.

O'Halloran, K. L. (2010). The semantic hyperspace: Accumulating mathematical knowledge across semiotic resources and modalities. In Christie, F. & Maton, K. (eds). Disciplinary: Functional Linguistic and Sociological Perspectives. London: Continuum, 217-236.

Olande, O. (2014). Graphical artefacts: Taxonomy of students' response to test items. *Educational Studies in Mathematics, 85*(1), 53-74. Retrieved from https://www.jstor.org/stable/43589806a

Österholm, M. & Bergqvist, E. (2012). Methodological issues when studying the relationship between reading and solving mathematical tasks. *Nordisk matematikkdidaktikk, 17*(1), 5-30. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-42287

Otte, M. F. (2011). Evolution, learning, and semiotics from a Peircean point of view. *Educational Studies in Mathematics, 77*(⅔), 313-329. Retrieved from https://www.jstor.org/stable/41485931

Ottmar, E., Landy, D., & Goldstone, R. (2012). Teaching the perceptual structure of algebraic expressions: Preliminary findings from the pushing symbols intervention. In *Proceedings of the Annual Meeting of the Cognitive Science Society, 34(34)* (pp. 2156-2161). Sapporo, Japan: Cognitive Science Society.

Panaoura, G., & Gagatsis, A. (2009). The geometrical reasoning of primary and secondary school students. In *The 6th Conference of the European Society for Research in Mathematics Education: Working Group 5, Geometrical Thinking* (pp. 746-755).

Panasuk, R. M. (2010). Three Phase Ranking Framework for Assessing Conceptual Understanding in Algebra Using Multiple Representations. *Education, 131*(2). Retrieved from https://pdfs.semanticscholar.org/ff5f/1a2be947be6957d85e1369a4de313c4965fe.pdf

Panasuk, R. M. & Beyranevand, M. L. (n.d.) Algebra students' ability to recognize multiple representations and achievement. *International Journal for Mathematics Teaching & Learning*. Retrieved from http://www.cimt.org.uk/journal/panasuk.pdf

Pape, S. J. & Tchoshanov, M. A. (2001). The role of representation(s) in developing mathematical understanding. *Theory Into Practice, 40*(2), 118-127. Retrieved from https://www.jstor.org/stable/1477273

Park, A. M. (2011). Comparing the cognitive demand of traditional and reform algebra 1 textbooks. Claremont, CA: Harvey Mudd College.

Paul, D. J., Nibbelink, W. H., & Hoover, H. D. (1986). The effects of adjusting readability on the difficulty of mathematics story problems. *Journal for Research in Mathematics Education, 17*(3), 163-171. Retrieved from https://www.jstor.org/stable/749299

Pepin, B., & Haggarty, L. (2001). Mathematics textbooks and their use in English, French and German classrooms: a way to understand teaching and learning cultures. *ZDM, The International Journal on Mathematics Education, 33*(5), 158-75.

Pfost, M., Hattie, J., Dörfler, T., & Artelt, C. (2014, June). Individual differences in reading development: A review of 25 years of empirical research on Matthew Effects in reading. *Review of Educational Research, 84*(2), 203-244. doi: 10.3102/0034654313509492

Phonapichat, P. & Wongwanich, S. (2014). *Effects of using K-N-W-S on mathematical problem-reading ability development among elementary school students*. Paper presented at the 5th World Conference on Educational Sciences - WCES 2013, Rome, Italy. Procedia - Social and Behavioral Sciences, 116, 5081-5084. doi: 10.1016/j.sbspro.2014.01.1077

Plunk, A. D., Tate, W. F., Bierut, L. J., & Grucza, R. A. (2014). Intended and unintended effects of state-mandated high school science and mathematics course graduation requirements on educational attainment. *Educational Researcher, 43*(5), 230-241. doi: 10.3102/0013189X14540207

Pointon, A. & Sangwin, C. J. (2003). An analysis of undergraduate core material in the light of hand-held computer algebra systems. *International Journal of Mathematical Education in Science and Technology, 34*(5), 671-686.

Polly, D. (2012). Supporting mathematics instruction with an expert coaching model. *Mathematics Teacher Education and Development, 14*(1), 78-93.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practices. *Educational Researcher, 31*(7), 3-14.

Porter, A., McMaken, J., Hwang, J., & Yang, R. Common core standards: The new U.S. intended curriculum. *Educational Researcher, 40*(3), 103-116. Retrieved from https://www,jstor.org/stable/41238917

Powell, S. R., Hebert, M. A., Cohen, J. A., Casa, T. M., & Firmender, J. M. (2017). A synthesis of mathematics writing: Assessments, interventions, and surveys. *Journal of Writing Research, 8*(3), 493-526. doi: 10.17239/jowr-2017.08.03.04

Presmeg, N. (2006). Semiotics and the "connections" standard: Significance of semiotics for teachers of mathematics. *Educational Studies in Mathematics, 61*(½), 163-182. Retrieved from https://www.jstor.org/stable/25472064

Presmeg, N., Radford, L., Roth, W., & Kadunz, G. (2016). *Semiotics in mathematics education*. Rotterdam, The Netherlands: Springer Nature.

Prins, E. D. & Ulijn, J. M. (1998). Linguistic and cultural factors in the readability of mathematics texts: The Whorfian hypothesis revisited with evidence from the South American context. *Journal of Research in Reading, 21*(2), 139-159.

Pugalee, D. K. (2001). Writing, mathematics, and metacognition: Looking for connections through students' work in mathematical problem solving. *School Science and Mathematics, 101*(5), 236-245.

Puteh, M. & Ibrahim, M. (2010). The usage of self-regulated learning strategies among form four students in the mathematical problem-solving contexts: A case study. *Procedia Social and Behavioral Sciences, 8*, 446-452.

Quinnell, L. (2014). Enhancing the teaching and learning of mathematical visual images. *Australian Mathematics Teacher, 70*(1), 18-25. Retrieved from https://files.eric.ed.gov/fulltext/EJ1093264.pdf

Radford, L. (2010, March). Algebraic thinking from a cultural semiotic perspective. *Research in Mathematics Education, 12*(1), 1-19 . doi: 10.1080/14797800903569741

Radford, L. (2014). On the role of representations and artefacts in knowing and learning. *Educational Studies in Mathematics, 85*, 405-422. doi: 10.1007/s106749-013-9527-x

Radford, L. (2000). Signs and meanings in students' emergent algebraic thinking: A semiotic analysis. *Educational Studies in Mathematics, 42*, 237-268. Retrieved from https://link.springer.com/content/pdf/10.1023/A:1017530828058.pdf

Radford, L. (2002). The seen, the spoken and the written: A semiotic approach to the problem objectification of mathematical knowledge. *For the Learning of Mathematics, 22*(2), 14-23. Retrieved from https://www.jstor.org/stable/40248387

Radford, L., Bardini, C., & Sabena, C. (2005). Perceptual semiosis and the microgenesis of algebraic generalizations. Fourth Congress of the European Society for Research in Mathematics Education (CERME 4), 17 - 21 February 2005, Sant Feliu de Guíxols, Spain. [http://laurentian.ca/educ/lradford/cerme4.pdf]

Rau, M. (2017). Conditions for the effectiveness of multiple visual representations in enhancing STEM learning. *Educational Psychology Review, 29*, 717-761. doi: 10.1007/s10648-016-9365-3

Raub, L. A., Shukor, N. A., Arshad, M. Y. & Rosli, M. S. (2015). An integrated model to implement contextual learning with virtual learning environment for promoting higher order thinking skills in Malaysian secondary schools. *International Education Studies, 8*(13), 41-46. doi: 10.5539/ies.v8n13p41

Regents exams: Mathematics - Multiple representations (2013). Retrieved from https://www.engageny.org/resource/regents-exams-mathematics-multiple-representations

Remillard, J. T., Baker, J. Y., Steele, M. D., Hoe, N. D., & Traynor, A. (2017). Universal algebra I policy, access, and inequality: Findings from a national survey. *Educational Policy Analysis Archives*, *25*(101). https://dx.doi.org/10.14507/epaa.25.2970

Remillard, J. T. & Bryans, M. B. (2004). Teachers' orientations toward mathematics curriculum materials: Implications for teacher learning. *Journal for Research in Mathematics Education*, *35*(5), 352-388. doi: 10.2307/30034820

Remillard, J. T. & Heck, D. (2014, June). Conceptualizing the curriculum enactment process in mathematics education. *ZDM Mathematics Education, 46,* 705-718. doi: 10.1007/s11858-014-0600-4

Renkl, A., Hilbert, T., & Schworm, S. (2009). Example-based learning in heuristic domains: A cognitive load theory account. *Educational Psychology Review*, *21*(1), 67-78. doi: 10.1007/s10648-008-9093-4

Reys, B., Dingman, S., Olson, T. A., Sutter, A., Teuscher, D., Chval, K., . . . Kasmer, L. (2006). *The Intended Mathematics Curriculum as Represented in State Level Curriculum Standards: Consensus or Confusion?* Greenwich, CT: Info Age Publishing.

Reys, R., Reys, B., Lapan, R., & Holliday, G. (2003). Assessing the impact of standards-based middle grades mathematics curriculum materials on student achievement. *Journal of Research in Mathematics Education, 34*(1), 74-95.

Richland, L. E., Begolli, K. N., Simms, N., Frausel, R. R., & Lyons, E. A. (2016). Supporting mathematical discussions: The roles of comparison and cognitive load. *Educational Psychology Review, 29*, 41-53. doi: 10.1007/s10648-016-9382-2

Riordan, J. E. & Noyce, P. E. (2001). The impact of two standards-based mathematics curricula on student achievement in mathematics. *Journal for Research in Mathematics Education, 32*(4), 368-398. Retrieved from https://www.jstor.org/stable/749700

Rossi Becker, J. & Rivera, F. (2005). Generalization strategies of beginning high school students. In Chick, H. L. & Vincent, J. L. (Eds.) *Proceedings of the 29th Conference of the International Group for the Psychology of Mathematics Education,* Vol. 4. pp. 121-128. Melbourne: PME.

Rumsey, D. J. (2002). Statistics literacy as a goal for introductory statistics courses. *Journal of Statistics Education, 10*(3), 1-12. doi: 10.1080/10691898.2002.11910678

Ruthven, K. (1994). Review article: Better judgement: Rethinking assessment in mathematics education. *Educational Studies in Mathematics, 27*(4), 433-450. Retrieved from https://www.jstor.org/stable/3482965

Sabena, C. (2018). Multimodality and the semiotic bundle lens: A constructive resonance with the theory of objectification. *PNA, 12*(4), 185-208.

Sáenz-Ludlow, A., & Kadunz, G. (2016). Constructing knowledge seen as a semiotic activity. In *Semiotics as a tool for learning mathematics* (pp. 1-21). Brill Sense.

Schaefer, S. L. (2019a). The evolution of NYS algebra standards and assessments: 2002 to present research component. Unpublished manuscript.

Schaefer, S. L. (2019b). The evolution of NYS algebra math assessments: 2002 to present. Paper presented at the Southern Tier Mathematics Conference, Binghamton, New York.

Schaefer, S. L. (2019c). How to use those old algebra regents exams. Paper presented at the Broome-Tioga BOCES teacher center, Binghamton, New York.

Schmidt, W. H., Wang, H. C., & McKnight, C. C. (2005). Curriculum coherence: An examination of US mathematics and science content standards from an international perspective. *Journal of Curriculum Studies, 37*(5), 525-559. https://doi.org/10.1080/0022027042000294682

Schneider, M., Merz, S., Stricker, J., De Smedt, B., Torbeyns, J., Verschaffel, L., & Luwel, K. (2018). Associations of number line estimation with mathematical competence: A meta-analysis. *Child Development, 89*(5), 1467-1484. doi: 10.1111/cdev.13068

Schoenfeld, A. H. (2015). Summative and formative assessments in mathematics supporting the goals of the common core standards. *Theory into Practice, 54*(3), 1-24. doi:10.1080/00405841.2015.1044346

Schreiber, C. (2013, January). Semiotic processes in chat-based problem-solving situations. *Educational Studies in Mathematics, 82*(1), 51-73. Retrieved from https://www.jstor.org/stable/23434839

Seeger, F. (2008). Intentionality and sign. In L. Radfod, G. Schubring, & F. Seeger (Eds.), *Semiotics in Mathematics Education: Epistemology, History, Classroom, and Culture* (pp. 1-18). Rotterdam, The Netherlands: Sense Publishers.

Sfard, A. (1991). On the dual nature of mathematical concepts: Reflections on processes and objects as different sides of the same coin. *Educational Studies in Mathematics, 22*(1), 1-36. Retrieved from https://www.jstor.org/stable/3482237

Sfard, A. (1992). Operational origins of mathematical objects and the quandary of reification – The case of function. In G. Harel & E. Dubinsky (Eds.), The concept of function: Aspects of epistemology and pedagogy (pp. 59–84). Washington, DC: The Mathematical Association of America.

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.

Shah, P. & Freedman, E. G. (2011). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Sciences, 3,* 560-578. doi: 10.1111/j.1756-8765-2009.01066.x

Shah, P. & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review, 14*(1), 47-69.

Sharma, S. V. (2006, October). High school students interpreting tables and graphs: Implications for research. *International Journal of Science and Mathematics Education, 4*(2), 241-268. doi: 10.1007/s10763-005-9005-8Sibol, S., & Watson, S. (2018). JMAP: Free Regents mathematics materials for teachers. Retrieved from http://www.jmap.org

Siegler, R. S., Duncan, G. J., Davis-Kean, P., Duckworth, K., Claessens, A., Engel, M., Susperreguy, M. I., & Chen, M. (2012, July). *Psychological Science, 23*(7), 691-697. Retrieved from https://www.jstor.org/stable/23262481

Silva, K. A. P. & Almedia, L. M. W. (2018). The exponential function meaning in mathematical modeling activities: A semiotic approach. *REDIMAT - Journal of Research in Mathematics Education, 7*(2), 195-215. doi: 10.4471/redimat.2018.2762

Silver, E. A., Clark, L. M., Ghousseini, H. N., Charalambous, C. Y., & Sealy, J. T. (2007, October 16). Where is the mathematics? Examining teachers' mathematical learning opportunities in practice-based professional learning task. *Journal of Mathematics Teacher Education, 10,* 261-277. doi: 10.1007/s10857-007-9039-7

Simpson, D., Frech, K., Garbade, L., & Schaefer, S. L. (2018). Critical reading in science and math literacy for 7-12. Broome-Tioga BOCES Study Group, Binghamton, NY, March 15, 2018.

Skemp, R. R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching, 77,* 20-26.

Smith, H. A. (1997). Peirce's sign and mathematics education: an introduction. *Philosophy of Mathematics Education Journal, 10.* Retrieved from webdoc.sub.gwdg.de/edoc/e/pome/pome10/art3.htm

Smith, J. B. (1996). Does an extra year make any difference? The impact of early access to algebra on long-term gains in mathematics attainment. *Educational Evaluation and Policy Analysis, 18*(2), 141-153. Retrieved from https://www.jstor.org/stable/1164553

Smith, K. J. (1965). The application of readability formulae to mathematics text materials: A note. *The Arithmetic Teacher, 12*(5), 354-355. Retrieved from http://www.jstor.org/stable/41186934

Smith, M. S. (2000, March). Balancing old and new: An experienced middle school teacher's learning in the context of mathematics instructional reform. *The Elementary School Journal, 100*(4), 351-375. Retrieved from https://www.jstor.org/stable/1002147

Sood, S. & Jitendra, A. K. (2007, March). A comparative analysis of number sense instruction in reform-based and traditional mathematics textbooks. *The Journal of Special Education, 41*(3), 145-157.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72-101.

Spielhagen, F. R. (2006). Closing the achievement gap in math: The long-term effects of eighth-grade algebra. *Journal of Advanced Academics, 18*(1), 34-59. Retrieved from https://files.eric.ed.gov/fulltext/EJ753970.pdf

Spoth, A. E. (2016). *An analysis of language difficulties in Algebra I (Common Core) assessments versus integrated algebra assessments.* (Master's thesis). Retrieved from https://dspace.sunyconnect.suny.edu/bitstream/handle/1951/70439/Spoth_Amy_MastersThesis_Spring2016.pdf?sequence=1&isAllowed=y

Stacey, K. & MacGregor, M. (1995). The influence of problem representation on algebraic equation writing and solution strategies. In L. Meira & D. Carraher (Eds.), Proceedings of the 19th International Conference for Psychology of Mathematics Education, 2, 90-97.

Steele, D. (2005, March). Using writing to access students' schemata knowledge for algebraic thinking. *School Science and Mathematics, 105*(3), 142-154. Retrieved from https://search.proquest.com/docview/195206115?accountid=9753

Steele, D. F. & Johanning, D. I. (2004). A schematic-theoretic view of problem solving and development of algebraic thinking. *Educational Studies in Mathematics, 57*(1), 65-90. Retrieved from https://www.jstor.org/stable/4150315

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal, 33*(2), 455-488. Retrieved from https://www.jstor.org/stable/1163292

Stein, M. K. & Smith, M. S. (1998). Mathematical tasks as a framework for reflection: From research to practice. *Mathematics Teaching in the Middle School, 3*(4), 268-275. Retrieved from https://www.nctm.org/Handlers/AttachmentHandler.ashx?attachmentID=wTjgEy0K1jw%3D

Stylianides, A. J. & Stylianides, G. J. (2007). Learning mathematics with understanding: A critical consideration of the learning principle in the principles and standards for school mathematics. *The Mathematics Enthusiast, 4*(1), 103-114. Retrieved from https://scholarworks.umt.edu/tme/vol4/iss1/8

Surya, E., Sabandar, J., Kusumah, Y. S., & Darhim, D. (2013). Improving of junior high school visual thinking representation ability in mathematical problem solving by CTL. *Journal on Mathematical Education, 4*(1), 113-126.

Swafford, J. O. & Langrall, C. W. (2000). Grade 6 students' preinstructional use of equations to describe and represent problem situations. *Journal for Research in Mathematics Education, 31*(1), 89-112.

Swee, F. N., & Lee, K. (2009). Model method: A visual tool to support algebra word problem solving at the primary level. In W. K. Yoong, L. P. Yee, B. Kaur, F. P. Yee & S. N. Fong (Eds.), *Mathematics education: The Singapore journey* (pp. 169-203). Singapore: World Scientific.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*, 295-312.

Tarasenkova, N. (2013). The quality of mathematical education in the context of semiotics. *American Journal of Educational Research, 1*(11), 464-471. doi: 10.12691/education-1-11-2

Teuscher, D., Kulinna, P. H, & Crooker, C. (2015). Writing to learn mathematics: An update. *The Mathematics Educator, 24*(2), 56-78.

Thompson, T. (2008, July). Mathematics teachers' interpretation of higher-order thinking in Bloom's taxonomy. *International Electronic Journal of Mathematics Education, 3*(2), 96-109. ISSN: 1306-3030

Tian, J. & Siegler, R. S. (2017). Fractions learning in children with mathematics difficulties. *Journal of Learning Disabilities, 50*(6), 614-620. doi: 10.1177/0022219416662032

Tinker, M. A. (1954). Readability of mathematics tables. *Journal of Applied Psychology, 38*(6), 436-442. http://dx.doi.org/10.1037/h0061760

Tversky, B. (2011). Visualizing thought. *Topics in Cognitive Science, 3*, 499-535. doi: 10.1111/j.1756-8765.2010.01113.x

Ubuz, B., Erbaş, A. K., Çetinkaya, B., & Özgeldi, M. (2010). Exploring the quality of the mathematical tasks in the new Turkish elementary school mathematics curriculum guidebook: The case of algebra. *ZDM Mathematics Education, 42*, 483-491. doi: 10.1007/s11858-010-0258-5

Usiskin, Z. (1995). Why is algebra important to learn? *American Educator, 19*(1), 30-37.

Usiskin, Z. (2004). Should all students study a significant amount of algebra? Keynote at the Dutch National Mathematics Days (NWD), 2004. [keynote also in Nieuw Archief voor Wiskunde, June 2004, 147-151]

Vale, I. & Cabrita, I. (2008). Learning through patterns: A powerful approach to algebraic thinking. In K. Kumpulainen & A. Toom, (Eds.), *The Proceedings of the 18th Annual Conference of the European Teacher Education Network*, (pp. 63-70). Helsinki, Finland: ETEN European Teacher Education Network.

Vamvakoussi, X. & Vosniadou, S. (2010). How many decimals are there between two fractions? Aspects of secondary school students' understanding of rational numbers and their notation. *Cognition and Instruction, 28*(2), 181-209. Retrieved from https://www.jstor.org/stable/27806348

Van Amerom, B. A. (2003). Focusing on informal strategies when linking arithmetic to early algebra. *Educational Studies in Mathematics, 54*(1), 63-75. Retrieved from https://www.jstor.org/stable/3483215

van den Bosch, R. M., Espin, C. A., Chung, S., & Saab, N. (2017). Data-based decision-making: Teachers' comprehension of curriculum-based measurement progress-monitoring graphs. *Learning Disabilities Research & Practice, 32*(1), 46-60. doi: 10.1111/ldrp.12122

van den Heuvel-Panhuizen, M. (1996). Assessment and realistic mathematics education. Ultrect, The Netherlands: Freudenthal Institute, Ultrecht University.

van den Heuvel-Panhuizen, M. (2005). The role of contexts in assessment problems in mathematics. *For the Learning of Mathematics, 25*(2), 2-9. Retrieved from http://www.jstor.org/stable/40248489

van Hiele, P. M. (1984). English summary. [The problem of insight in connection with school children's insight into the subject matter of geometry.] (Unpublished doctoral dissertation, University of Utrecht, 1957) In D. Fuys, D. Geddes, & R. Tischler (Eds.), English translation of selected writings of Dina van Hiele-Geldof and P. M. van Hiele (pp. 237-241). Brooklyn: Brooklyn College.

van Lier, L., & Walqui, A. (2012, January). Language and the common core state standards. Paper presented at the Understanding Language Conference, Stanford, CA. Retrieved from http://ell.stanford.edu/papers/language.

Venenciano, L. & Heck, R. (2016). Proposing and testing a model to explain traits of algebra preparedness. *Educational Studies in Mathematics, 92*, 21-35. doi: 10.1007/s10649-015-9672-5

Vickers, P., Faith, J., & Rossiter, N. (2013). Understanding visualization: A formal approach using category theory and semiotics. *IEEE Transactions on Visualization and Computer Graphics, 19*(6), 1048-1061. Retrieved from http://dx.doi.org/10.1109/TVCG.2012.294

von Gog, T. & Kester, L. (2012). A test of the testing effect: Acquiring problem solving skills from worked examples. *Cognitive Science, 36*, 1532-1541. doi: 10.1111/cogs.12002

von Gog, T. & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review, 22*(2), 155-174. Retrieved from https://www.jstor.org/stable/23364127

Walberg, H. J. & Tsai, S. (1983). Matthew effects in education. *American Educational Research Journal, 20*(3), 359-373. Retrieved from https://journals.sagepub.com/doi/pdf/10.3102/00028312020003359

Walker, C. M., Zhang, B., & Surber, J. (2008). Using a multidimensional differential item functioning framework to determine if reading ability affects student performance in mathematics. *Applied Measurement in Education, 21*, 162-181. doi: 10.1080/08957340801926201

Walkington, C., Clinton, V., & Howell, E. (2013). The associations between readability measures and problem solving in algebra. In Martinez, M. & Castro Superfine, A. (Eds.) (pp. 86-89) *Proceedings of the 35th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Chicago, IL: University of Illinois at Chicago.

Walkington, C., Clinton, V., Ritter, S. N., & Nathan, M. J. (2015). How readability and topic incidence relate to performance on mathematics story problems in computer-based curricula. *Journal of Educational Psychology, 107*(4), 1051-1074. http://dx.doi.org/10.1037/edu0000036

Walkington, C., Clinton, V., & Shivraj, P. (2017). How readability factors are differentially associated with performance for students of different backgrounds when solving mathematics word problems. *American Educational Research Journal, 55*(2), 362-414. doi: 10.3102/0002831217737028

Wang, Y., Lin, L., Kuhl, P., & Hirisch, J. (2007). Mathematical and linguistic processing differs between native and second languages: An fMRI study. *Brain Imaging and Behavior, 1*, 68-82. doi: 10.1007/s11682-007-9007-y

Watson, R. S. (2010). *Stability and Change in New York State Regents Mathematics Examinations, 1866-2009: A Socio-Historical Analysis.* UMI Dissertation Publishing, UMI number 3433796.

Webb, N. L. (1999). Alignment of science and mathematics standards and assessments in four states. Research monograph #18, National Institute for Science Education and Council of Chief State School Officers. Madison, Wisconsin: Wisconsin Center for Education Research. Retrieved from http://www.wcer.wisc.edu/archive/nise/publications/Research_Monographs/vol18.pdf

Webb, N. (November, 2005). Depth-of-Knowledge levels for four content areas. Presentation to the Florida Education Research Association, 50th Annual Meeting, Miami, Florida

Webb, N. & Herman, J. (2006, June). Alignment of mathematics state-level standards and assessments: The role of reviewer agreement. CSE Report 685. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST).*

What is a quantile measure. (2019). Retrieved from https://www.quantiles.com/parents-students/understanding-quantile-measures/what-is-a-quantile-measure/

Wilson, A. A. (2011, March). A social semiotics framework for conceptualizing content area literacies. *Journal of Adolescent & Adult Literacy, 54*(6), 435-444. doi:10.1598/JAAL.54.6.5.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research, 20,* 557-585.

Wu, H. (2014). Potential impact of the common core mathematics standards on the American curriculum. In Y. LI, G. Lappan (eds.) *Mathematics Curriculum in School Education* (pp. 119-142), Advanced in Mathematics Education, Dordrecht: Springer.

Yerushalmy, M. (2005). Functions of interactive visual representations in interactive mathematical textbooks. *International Journal of Computers for Mathematical Learning, 10*(3), 217-249.

Yu, K., Fan, S., & Lin, K. (2014). Enhancing students' problem-solving skills through context-based learning. *International Journal of Science and Mathematics Education, 12,* 64-78.

Yung, H. I., & Paas, F. (2015). Effects of computer-based visual representation on mathematics learning and cognitive load. *Educational Technology and Society, 18*(4), 70-77.

Zawicki, J., Jabot, M., Falconer, K, MacIsaac, D. Henry, D., Fischer, R. (2003). A preliminary analysis of the June 2003 New York State Regents Examination in Physics. Perspectives on Science Education (April). Penn Yan, NY: New York State Science Education Leadership Association.

Zimmerman, B. J., & Dibenedetto, M. K. (2008). Mastery learning and assessment: Implications for students and teachers in an era of high-stakes testing. *Psychology in the Schools, 45*(3), 206-216. doi:101.1002/pits.20291

Appendix A

Directions for Readability Formulas

1. For each item, find the total number of words. Words cannot be a part of any other representation other than a verbal representation.
2. For each item, find the total number of sentences. Sentences start with a capital letter and end with a punctuation mark. For item stems that do not end with punctuation, the responses are considered as part of one sentence.
3. For each item, find the number of syllables present. This does not include any other part of a representation other than a verbal representation.

The Flesch-Kincaid grade level is found with the following formula:

$$0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllables}{total\ words}\right) - 15.59$$

The Flesch reading ease score (1979) is found with the following formula:

$$206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right)$$

Appendix B

Definition of Contextual Situations

Categories	*Definition*	*Examples*
1. Contextual	"Real world" applications that bring in outside settings or scenarios	Aug. 2018, *Current*, 34 A car was purchased for $25,000. Research shows that the car has an average yearly depreciation rate of 18.5%. Create a function that will determine the value, $V(t)$, of the car t years after purchase. Determine, to the *nearest cent*, how much the car will depreciate from year 3 to year 4. June 2018, *Current*, 7 On the main floor of the Kodak Hall at the Eastman Theater, the number of seats per row increases at a constant rate. Steven counts 31 seats in row 3 and 37 seats in row 6. How many seats are there in row 20? Jan. 2009, *Old*, 33 The "Little People" day care center has a rectangular, fenced play area behind their building. The play area is 30 meters long and 20 meters wide. Find, to the *nearest meter*, the length of a pathway that runs along the diagonal of the play area.
0. Non-Contextual	Items that do not include outside settings or scenarios	Jan. 2017, *Current*, 27 Solve the inequality below: $1.8 - 0.4y \geq 2.2 - 2y$ June 2008, *Recent*, 20 What is the slope of the line that passes through the points (–6,1) and (4,–4)? (1) -2 (2) 2 (3) -1/2 (4) 1/2 August 2006, *Old*, 11 Which transformation does not always result in an image that is congruent to the original figure? (1) dilation (2) reflection (3) rotation (4) translation

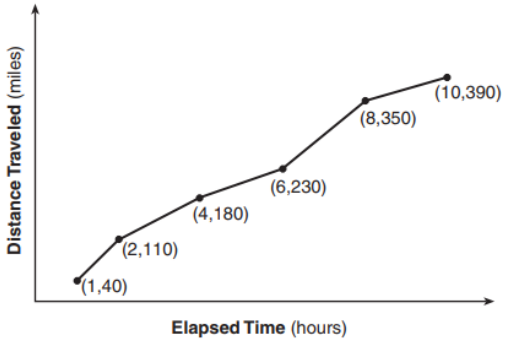
Appendix C

Cognitive Demand Complexity Framework

Levels	Explanation	Examples
Level 0 Very Low *Memorization Tasks*	• Requires only recall of information without computation or only reproduction of facts or a definition • Involves immediate recognition of the answer being correct or incorrect • Is unambiguous, requires exact reproduction of previously seen material • Includes no connections to the meaning of the underlying concepts, rules or definitions being used • Involves no explanation and no "real world" context	January 2009, *Old*, 23 Which statement is an example of a biconditional statement? (1) If Craig has money, he buys a car. (2) Craig buys a car if and only if he has money. (3) Craig has money or he buys a car. (4) Craig has money and he buys a car. January 2009, *Old*, 24 Which property of real numbers is illustrated by the equation 52 + (27 + 36) = (52 + 27) + 36? (1) commutative property (2) associative property (3) distributive property (4) identity problem of addition
Level 1 Low *Procedural Tasks without Connections*	• Requires only memorized algorithmic steps to solve the tasks by carrying out a routine explanation or well rehearsed process where the use of the procedure is either specifically called for or immediately evident from previous experience or instruction • Involves recognition of the procedures to be used to complete the task • Focuses on the procedure and solving to get the correct answer rather than developing or demonstrating mathematical understanding • Includes no connections to the meaning of the underlying concepts, rules or definitions being used • Involves no explanation or an explanation that only focuses on describing the steps of procedure that was used like a recipe	June 2014, *Current*, 27 Given 2x + ax – 7 > -12, determine the largest integer of *a* when x = -1. January 2009, *Old*, #4 Subtract $2x^2 - 5x + 8$ from $6x^2 + 3x - 2$ and express the answer as a trinomial.

Appendix C

Cognitive Demand Complexity Framework (Continued)

Levels	Explanation	Examples
Level 2 High *Procedures with connections*	• Requires some degree of conceptual understanding of the problem situation along with the algorithmic thinking; requires knowing how and why can use procedures • May or may not involve recognition of the procedures to be used to complete the task, but even if recognized procedures are involved, they cannot be followed mindlessly • Focus is on developing or demonstrating mathematical understanding • Requires connections to the underlying concepts often involving multiple representation types being used for mathematical ideas • Requires an explanation that goes beyond procedures to include conceptual meaning, interpreting a situation, or explaining a "real world" application	June 2014 Current Exam #18 **18** The Jamison family kept a log of the distance they traveled during a trip, as represented by the graph below.  During which interval was their average speed the greatest? (1) the first hour to the second hour (2) the second hour to the fourth hour (3) the sixth hour to the eighth hour (4) the eighth hour to the tenth hour

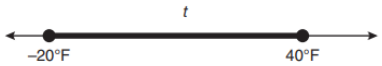
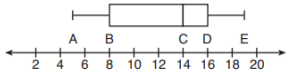
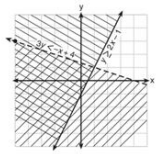
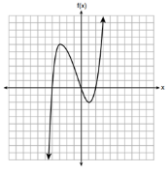
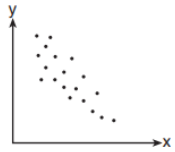
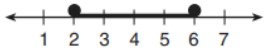
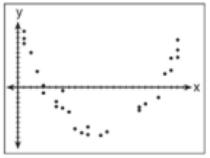
Appendix C

Cognitive Demand Complexity Framework (Continued)

Levels	Explanation	Examples
Level 3 Very High *Doing Mathematics*	• Requires non-algorithmic or complex thinking to solve a novel task that has not been rehearsed and whose pathway is not clear • Involve no recognition of the procedures to be used to complete the task, but rather involves students exploring the nature of mathematical relationships and contexts • Focus is on developing or demonstrating mathematical understanding through analysis of a complex task • Requires analyze of task to examine any constraints that might be involved with the use of certain representations or strategies • Requires an explanation that goes beyond procedures to include conceptual meaning, interpreting a situation, or explaining a "real world" application in ways that are novel and challenging	January 2017, *Current*, 26 A typical marathon is 26.2 miles. Allan averages 12 kilometers per hour when running in marathons. Determine how long it would take Allan to complete a marathon, to the nearest tenth of an hour. Justify your answer.

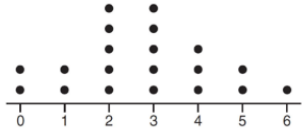
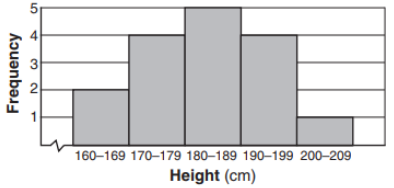
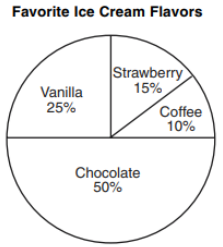
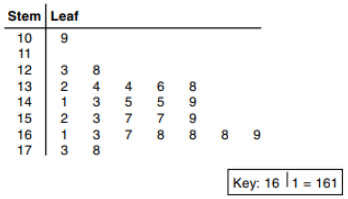
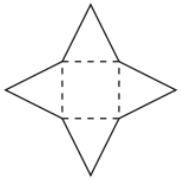
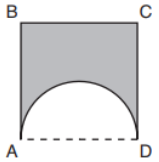
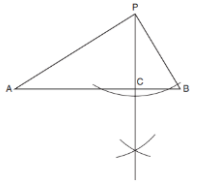
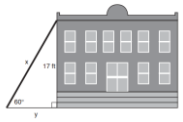
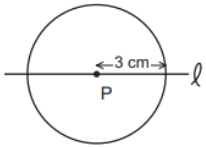
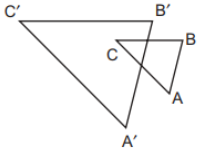
Appendix D

Conceptual Categories and Subcategories for Multiple Representations

Category	Subcategory	Explanation	Examples
V. Visual	V-G Graphs	Plots of single points on one-dimensional lines; points and curves in two dimensional planes (that may or may not be functions); points, curves, and surfaces in three-dimensional spaces (that may or may not be functions)	June 2008, *Recent*, #33  Jan. 2015, *Recent*, #8  Jan. 2009, *Old*, #22  Aug. 2018, *Current*, #5  June 2015, *Recent*, #12  Jan. 2003, *Old*, #12  Jan. 2016, *Current*, #24 

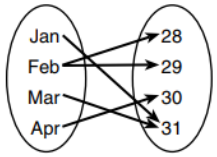
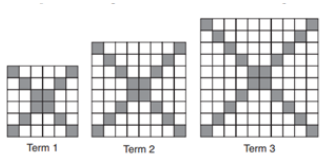
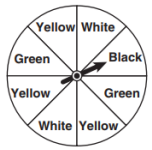
Appendix D

Conceptual Categories and Subcategories for Multiple Representations (Continued)

Category	Subcategory	Explanation	Examples		
V. Visual	V-C Charts	Charts represent comparative quantities in one or two dimensions such as pie charts, dot plots, bar charts, histograms, and stem and leaf plots	June 2016, *Current*, 20 	Jan 2005, *Old*, 4 	
			Jan. 2003, *Old*, 25 	June 2005, *Old*, 9 	
V. Visual	V-F Geometric Figures	Geometric figures (e.g., circles, triangles) and lines and/or shapes with geometric relationships (e.g., parallel lines with a transversal)	August 2002, *Recent*, 15 	Jan 2014, *Recent*, 7 	Jan. 2004, *Old*, 20 
			Aug. 2002, *Old*, 31 	June 2002, *Old*, 12 	June 2002, *Old*, 16 

Appendix D

Conceptual Categories and Subcategories for Multiple Representations (Continued)

Category	Subcategory	Explanation	Examples	
V. Visual	V-D Diagram	Diagrams represent **mathematical** relationships and patterns (e.g. diagram of 1-1 function mapping domain values to range values; pattern blocks representing a function and/or sequence)	June 2017, *Current*, 9 	June 2014, *Current*, 24 
V. Visual	V-P Picture	Iconic pictures illustrate situations that would be seen in life that do not in and of themselves represent mathematical concepts or relationships (e.g., pictures of spinners and hats)	June 2008, *Recent*, 2 	Aug. 2010, *Recent*, 38 

Appendix D

Conceptual Categories and Subcategories for Multiple Representations (Continued)

Category	Subcategory	Explanation	Examples
S. Symbolic	S-A Algebra	Expressions, equations, inequalities, and systems of equations involving algebraic notation, arithmetic operators, numerals, and algebraic symbols; includes items written in function notation (e.g., coefficients, variables, exponents)	June 2015, *Recent*, 6 $81 - 16x^2$ June 2015, *Recent*, 11 $2a + 3b = 12$ $a = \frac{1}{2}b - 6$ June 2015, *Recent*, 6 $(9 - 4x)(9 + 4x)$ Aug. 2018, *Current*, 10 $1.25r + 2g \leq 50$ $\boldsymbol{r \geq 10}$ $g > 12$ Aug. 2018, *Current*, 6 $f(x) = x^2$ Aug. 2018, *Current*, 15 $g(x) = \begin{cases} 5x, & x \leq 3 \\ x^2 + 4, & x > 3 \end{cases}$ Aug. 2018, *Current*, 32 $a_1 = 0$ $a_n = 2(a_{n-1})^2 - 1$, for $n > 1$
S. Symbolic	S-S Sets	Groups of elements written in set notation or as a list using brackets or symbols used to denote elements, sets, or subsets (e.g., A $\subseteq$ B, a$\notin$B, {1, 2, 3})	June 2015, *Recent*, 1 $A = \{1, 2, 3, 4, 5, 6, 7, 8\}$ June 2014, *Recent*, 26 $M \cup N$ June 2014, *Recent*, 26 M = {green, red, yellow, black}
S. Symbolic	S-I Intervals	A subset of real numbers on the number line written with square brackets, parenthesis, or equivalent symbols (e.g., 2 < x < 5, (3, 5])	June 2015, *Recent*, 29 [6,10) June 2014, *Current*, 9 $x \geq 2$ June 2014, *Current*, 9 $2.5 < x < 5.5$

Appendix D

Conceptual Categories and Subcategories for Multiple Representations (Continued)

Category	Subcategory	Explanation	Examples		
S. Symbolic	S-P Probability Statistics	Use of letters and specific markings in mathematics that refer to the likelihood of events and/or the analysis of numerical data (e.g. n!, ${}_nC_r$, $P_{(n,r)}$, Q_2, σ)	Jun. 2006, *Old*, 5 $\frac{8!}{4!}$	Aug. 2007, *Old*, 4 ${}_{13}P_3$	Aug. 2007, *Old*, 4 ${}_{13}C_3$
			Aug. 2018, *Current*, 31 $\bar{x}, \sigma_x, n,$ min, med, max, Q_1, Q_3		
S. Symbolic	S-G Geometric notation	Use of symbols and specific markings that denote geometric figures and relationships (e.g., ΔXYZ, a⊥b, c‖d, ∠PQR)	June 2015, *Recent*, 23 m∠F = 90	June 2015, *Recent*, 32 $\overline{BC}$	June 2009, *Old*, 31 ΔLMN
			Jan. 2003, *Old*, 9 line *C*	Jan. 2003, *Old*, 16 *BC* = 15	Jan. 2003, *Old*, 2 Triangle *A'B'C'*
			Jan. 2003, *Old*, 16 tan ∠*C*		
S. Symbolic	S-Q Sequences and Series	Use of symbols and specific markings that denote sequences, series, and summation (e.g., A = $\{a_i\}$ = $\{2i + 1\}$, $\sum_1^8(1 - 3n)$)	June 2014, *Current*, 24 $a_n = 4n + 2$	August 2014, *Current*, 24 $f(1) = 3\ and\ f(n) = -2f(n-1) + 1$	

Appendix D

Conceptual Categories and Subcategories for Multiple Representations (Continued)

Categories	*Subcategory*	*Definition*	*Examples*		
N. Numeric	N-N Numeral	Quantities that may be real (rational or irrational) or complex; may involve root symbols or plus/minus symbols; repeating decimals may involve bars over the repeating sequence of end digits, fractions may involve the fraction bar, monetary values may involve a dollar or cent sign; other applications may involve other units (e.g., 55 mph); certain irrational constants (i.e., e and π) that frequently occur may be represented by a letter	June 2014, *Current*, 10 $2 \pm 2\sqrt{5}$ June 2015, *Recent*, 15 $\frac{43}{50}$ Aug. 2018, *Current*, 12 $\frac{4\ quarts}{1\ gallon} \cdot \frac{2\ pints}{1\ quart} \cdot \frac{2\ cups}{1\ pint} \cdot \frac{\frac{1}{4}\ cup}{4\ tablespoons} \cdot \frac{3\ teaspoons}{1\ tablespoon}$	June 2014, *Current*, 5 8.25 Jan. 2003, *Old*, 4 $\sqrt{3}, \pi, 3.2, 3\frac{1}{3}$	Jan. 2003, *Old*, 19 $\frac{6.3 \times 10^8}{3 \times 10^4}$ Aug. 2002, *Old*, 20 IV
N. Numeric	N-P Points	Corresponding quantities that typically ordered pairs or triples written in the traditional n-tuple form with parentheses	Jan. 2009, *Old*, 14 (3.5, -2)		

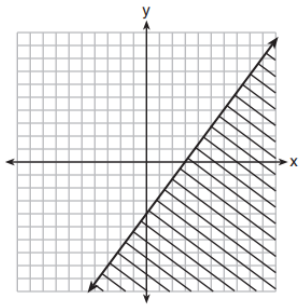
Appendix D

Conceptual Categories and Subcategories for Multiple Representations (Continued)

<table>
<tr><th>Category</th><th>Subcategory</th><th>Explanation</th><th>Examples</th></tr>
<tr><td>N.
Numeric</td><td>N-T
Tables and
Matrices</td><td>Sets of ordered n-tuples that are arranged in rectangular arrays using rows and columns to depict corresponding relationships</td><td>June 2015, Recent, 2
<table><tr><th>Candidate's Name</th><th>Number of Supporters</th></tr><tr><td>Ashley</td><td>30</td></tr><tr><td>Britney</td><td>28</td></tr><tr><td>Lyshon</td><td>14</td></tr><tr><td>Walker</td><td>28</td></tr></table>
Aug. 2002, Old, 22
<table><tr><th>⊕</th><th>1</th><th>2</th><th>3</th><th>4</th></tr><tr><td>1</td><td>2</td><td>3</td><td>4</td><td>1</td></tr><tr><td>2</td><td>3</td><td>4</td><td>1</td><td>2</td></tr><tr><td>3</td><td>4</td><td>1</td><td>2</td><td>3</td></tr><tr><td>4</td><td>1</td><td>2</td><td>3</td><td>4</td></tr></table>
June 2016, Current, 6
<table><tr><th>x</th><th>f(x)</th></tr><tr><td>1</td><td>12</td></tr><tr><td>2</td><td>19</td></tr><tr><td>3</td><td>26</td></tr><tr><td>4</td><td>33</td></tr></table>
June 2016, Current, 15
<table><tr><th>Age</th><th>For</th><th>Against</th><th>No Opinion</th></tr><tr><td>21–40</td><td>30</td><td>12</td><td>8</td></tr><tr><td>41–60</td><td>20</td><td>40</td><td>15</td></tr><tr><td>Over 60</td><td>25</td><td>35</td><td>15</td></tr></table>
Aug. 2018, Current, 31
<table><tr><th></th><th>$\bar{x}$</th><th>σ_x</th><th>n</th><th>min</th><th>Q_1</th><th>med</th><th>Q_3</th><th>max</th></tr><tr><td>4th Period</td><td>77.75</td><td>10.79</td><td>20</td><td>58</td><td>69</td><td>76.5</td><td>87.5</td><td>96</td></tr><tr><td>6th Period</td><td>78.4</td><td>9.83</td><td>20</td><td>59</td><td>71.5</td><td>78</td><td>88</td><td>96</td></tr></table></td></tr>
<tr><td>W.
Written</td><td>-</td><td>Written explanation or description involving words, may include numerals or numbers, but unless they are pertinent to the mathematics at hand, they are not included (e.g., Two girls travel x miles north, turn and travel y miles east. How far did they travel? would not receive this coding since two isn't mathematically pertinent information.)</td><td>June 2015, Recent, 22
For a class of students, which data set could be classified as qualitative?
(1) political opinions
(2) heights
(3) weights
(4) ages</td></tr>
</table>

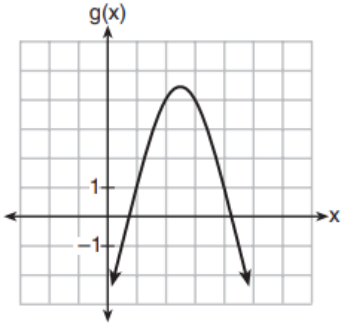
Appendix E

Representations Relationships Framework

Category	Explanation	Examples
RT. Representation Treatment	item involves staying in the same representation but moving through different versions of that representation	June 2014, *Current*, 8 Which equation has the same solution as $x^2 - 6x - 12 = 0$? (1) $(x + 3)^2 = 21$ (2) $(x - 3)^2 = 21$ (3) $(x + 3)^2 = 3$ (4) $(x - 3)^2 = 3$
RC. Representation Conversion	item involves moving from one representation to another	June 2015, *Recent*, 5 **5** Which inequality is shown in the graph below?  (1) $y \leq \frac{4}{3}x + 3$ (2) $y \geq \frac{4}{3}x + 3$ (3) $y \leq \frac{4}{3}x - 4$ (4) $y \geq \frac{4}{3}x - 4$

Appendix E

Representations Relationships Framework (Continued)

Category	Explanation	Examples
RS. Representation Synergy	item involved using more than one representation to work towards the solution	June 2015, *Current*, 14 **14** Which quadratic function has the largest maximum? $h(x) = (3 - x)(2 + x)$ (1) $k(x) = -5x^2 - 12x + 4$ (3) (2)  (4)

x	f(x)
−1	−3
0	5
1	9
2	9
3	5
4	−3