ACHIEVEMENT, ASSESSMENT, AND LEARNING: A STUDY OF EMERGENT
BILINGUAL STUDENTS IN MAINSTREAM CONTENT CLASSROOMS

A Dissertation
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Linguistics

By

Aubrey Elizabeth Logan-Terry, M.S.

Washington, DC
October 31, 2011
ACHIEVEMENT, ASSESSMENT, AND LEARNING: A STUDY OF EMERGENT BILINGUAL STUDENTS IN MAINSTREAM CONTENT CLASSROOMS

Aubrey E. Logan-Terry, M.S.

Thesis Co-Advisors: Heidi E. Hamilton, Ph.D.; Jeffrey Connor-Linton, Ph.D.

ABSTRACT

This dissertation investigates and problematizes emergent bilingual students’ achievement gaps in an existing corpus of middle-school science assessment data. I first characterize achievement gaps across national data, state data, and the corpus of approximately 27,000 students as an analytic backdrop.

I then sample a subset of the 6th grade physics tests (n = 852) and employ computerized readability tools and multilevel modeling to determine whether linguistic complexity of item prompts predicts increased achievement gaps across emergent bilingual and native English speaking groups of students. Findings indicate that some measures of linguistic complexity are related to increased achievement gaps while others may serve to close achievement gaps. These results have important implications for the pressing debate surrounding language as a source of construct-irrelevant variance in content tests and test accommodations for emergent bilingual students (e.g. Abedi, 2006; Bailey, 2005; Farnsworth, 2008; Martiniello, 2008; Wolf & Leon, 2009).

Motivated by results from the previous quantitative analysis, I conduct a follow-on qualitative study of 12 emergent bilingual and 15 native English speaking students’ constructed responses in order to further explore sources of achievement gaps. Drawing on analytic tools from Interactional Sociolinguistics and schema
theory, I show how the knowledge schemas (Tannen & Wallat, 1993) of each of the interlocutors involved in the test interactions—test developers, students, and raters—play a crucial role in the construction of meaning during the test taking process. Furthermore, I point to evidence of important contextualization cues that are used by the respective interlocutors to frame aspects of the test interaction as relating to certain target ideas. I also identify specific elements that contribute to schema miscues for emergent bilingual students, including the sequence of test items, the particular lexical items provided in the text prompts, grammatical aspects of test prompts, and visual accompaniment of items.

Based on findings from the above analyses, I show that emergent bilingual students’ "achievement gaps" are in fact—at least to a certain extent—created by monolingual content testing practices. Thus, I argue that monolingual content tests serve as gatekeeping encounters (Erickson, 1975; Erickson & Schultz, 1982) for emergent bilingual students.
ACKNOWLEDGMENTS

First, I would like to thank my dissertation committee members: Heidi Hamilton, Jeff Connor-Linton, Ande Tyler, and Tim Farnsworth.

Heidi, you are a model researcher, mentor, and human being. Throughout every phase of my process here at Georgetown you have provided insight and encouragement. You have taught me the value of rigorous qualitative research, as well as the importance of applying research in cross-disciplinary contexts. It is not easy to bridge paradigms and address important real-world problems. Your unwavering support is the reason I have been able to finish this project and move on to new, exciting challenges. Thank you, and I hope to make you proud as I move forward with my career.

Jeff, thank you for agreeing to be a co-chair on this committee and for introducing me to quantitative methodologies. Your point of view has caused me to try to make this work as airtight as possible.

Ande, in my early years at Georgetown you showed me how to bring together my longstanding interest in second language acquisition/bilingualism with my emerging passion for discourse analysis. Throughout my Master’s degree and this dissertation process, you have been a source of inspiration and guidance, and for that I am very grateful.

Tim, your expertise in the areas of testing and quantitative methodology have greatly shaped my work. I am extremely grateful for the extensive feedback you have given me over the course of this project. You have truly gone above-and-beyond as an outside committee member, giving me invaluable critiques and
pushing me to be a much stronger researcher. I look forward to continuing to collaborate with you in the future.

This dissertation would also not have been possible without the generous support and collaboration of Laura Wright. You first introduced me to the SCALE-uP project and allowed me my first opportunities to work in the area of testing for emergent bilingual students. Working with you has been an absolute joy, and inspired much of my dissertation effort. I would also like to thank the other researchers on the original SCALE-uP project who have given me crucial access to data and assistance: Joel Kuipers, Curtis Pyke, Rob Ochsendorf, and Gail Viechnicki.

Other scholars have also served as mentors to me over the years here at Georgetown. Kendall King, thank you for socializing me into the academic world and teaching me much of what I know about how to be a member of this community. While working with Kendall, I also benefitted from guidance and collaboration with Alison Mackey. Thank you, Alison, for teaching me so many important lessons about research and academic life.

Deborah Tannen, your work and wisdom have long inspired me, and your continued advice and support make it possible for me to thrive as a teacher and researcher here at Georgetown. Another shining example of strong research and academic success is Elana Shohamy. Thank you, Elana, for sharing your passion in the areas of testing and policy, and for your continued friendship and collaboration.

I would not have been able to get through the dissertation process without the statistical help of Rusan Chen. Thank you for spending so many hours teaching me how to do things that once seemed impossible. Thanks are also due to others in
the Georgetown community who have provided me with advice, encouragement, and feedback at the most important of times. Anna Trester, you are an incredible scholar and colleague. Thank you for being there for me when I have needed it most. Heather Weger, Lyn Fogle, and Natalia Jacobsen, you have also helped me immensely throughout this process. Thank you.

While I view other graduate students past and present as a sort of family that has been integral in helping me succeed in our dissertation program (thanks to each of you for the study dates, commiseration, and many laughs along the way!), I am also eager to thank my family and friends for everything they have done to make this possible.

Mom and Dad, everything I do is a reflection of the many amazing opportunities and unconditional love you have given me throughout my life. Thank you for being the best parents anyone could ever wish for. Thanks also to my extended family who have provided me with so much love and support. The desire to make you proud has helped me achieve so many of my goals.

Stephanie Cramer has been my best friend for over a decade, and it is through many a tear-filled phone call that you have held my hand and urged me on when otherwise I might not have. Randy Althaus, I am so glad to have met you as a colleague and continue to enjoy your friendship, encouragement, and hilarity. Jamie Schissel, you are a wonderful friend and colleague—I know we will continue to do great things together (e.g. take over the world).

My lovely Southern Spanish Linguistics Ladies (SSLLs), Elizabeth Kissling, Rachael Albritten, Sarah Grey, and Ellen Johnson Serafini: thank you for being there
through the highs and the lows. Celebrating each of our victories throughout the process has been a highlight of graduate school for me.

Rebecca Damari, I am so incredibly fortunate to have benefitted from your insight, friendship, and support in the dissertation process. I cannot imagine how I would have ever survived this if not for you. And I am beyond ecstatic to be able to continue working with you on our new project—we have so many exciting, fun, and challenging things to come!

Finally, I would like to thank the love of my life, Brian C. Williams. You are the definition of a supportive partner, and I am the luckiest lady on earth to have found you. Thank you for riding the dissertation roller coaster with me. I can’t wait to spend the rest of my life with you (and I promise there will be fewer sleepless nights now that this dissertation is done!).
DEDICATION

For the many emergent bilingual students in U.S. schools,
whose academic achievement and potential deserve our attention.
# TABLE OF CONTENTS

Chapter 1: Introduction

1.1 Problem.......................................................................................................................... 1

1.2 SCALE-uP Data .............................................................................................................. 3

1.2.1 The Questions about Motion Middle School Physics Test .................................. 5

1.2.2 Emergent Bilingual and Native English Speaking Students ............................... 9

1.3 Theoretical Framework .................................................................................................. 10

1.4 Study Design .................................................................................................................. 14

Chapter 2: Achievement Gaps:

Setting the Analytic Scene .................................................................................................. 16

2.1 Introduction .................................................................................................................... 16

2.2 Method ............................................................................................................................ 18

2.3 Findings ........................................................................................................................... 21

2.3.1 The National and State Backdrop............................................................................. 21

2.3.2 SCALE-uP Achievement Gaps across the Corpus ................................................. 27

2.3.3 SCALE-uP Achievement Gaps in Physics Implementations ................................. 31

2.3.4 SCALE-uP Achievement Gaps in the Year 2 Physics Implementation ................. 36

2.3.5 SCALE-uP Achievement Gaps in the Eastmond Classroom ............................... 41

2.4 Conclusion ....................................................................................................................... 43
Chapter 3: Linguistic Characteristics of Assessment Items:

An Investigation of Achievement Gaps Using Computerized Readability Tools and Multilevel Modeling

3.1 Introduction

3.1.1 Problem

3.1.2 Operationalizing Linguistic Characteristics of Test Items

3.1.3 Incorporating Readability Literature

3.1.3.1 Coh-Metrix Measures

3.1.4 Summary of Findings from Past Literature

3.1.5 Research Questions and Hypotheses

3.2 Method

3.2.1 Sample

3.2.2 Analysis

3.2.2.1 Measurement of Linguistic Characteristics of Test Items

3.2.2.2 Multilevel Modeling

3.3 Findings

3.3.1 RQ1) Are Linguistic Characteristics of Test Items Related to Students’ Item-Level Scores?

3.3.1.1 Descriptive Statistics for Initial Predictor Variables (Linguistic Characteristics) and Creation of Variables for Exploratory MLM

3.3.1.2 Exploratory Univariate MLM Analyses of the Characteristics-Item-Level Scores Relationships

3.3.1.3 Confirmatory Multivariate MLM Analyses
3.3.1.4 Summary of Findings from the Exploratory and Confirmatory Multivariate MLM Analyses for Research Question One............. 95

3.3.2 RQ2) Are There Differences Between Students’ Item-Level Scores Across ESOL Groups? ................................................................. 96

3.3.2.1 Descriptive Statistics Across ESOL Groups ........................................... 96

3.3.2.2 Univariate MLM Investigating ESOL Group Status as a Predictor of Students’ Item-Level Scores ................................................................. 97

3.3.2.3 Summary of Findings for Research Question Two............................... 98

3.3.3 RQ3) Does the Relationship Between Linguistic Characteristics of Test Items and Students’ Item-Level Scores Vary Across ESOL Groups? ........................................................................................................ 98

3.3.3.1 Tests for Interactions between Each of the Predictor Variables and ESOL Group Status .................................................................................. 99

3.3.3.2 Differences between Odds Ratios for Main Effects of Linguistic Complexity Variables Across ESOL Groups.................................................. 106

3.3.3.3 Summary of Findings for Research Question Three............................... 112

3.4 Conclusion ............................................................................................................. 116

3.4.1 Summary ............................................................................................................. 116

3.4.2 Discussion and Implications ............................................................................. 117

3.4.2.1 Research Question One ........................................................................... 117

3.4.2.2 Research Question Two ........................................................................... 118

3.4.2.3 Research Question Three ......................................................................... 119

3.4.2.4 Further Significance of Study Findings ..................................................... 122
3.4.3 Limitations and Suggestions for Future Research........................................124

Chapter 4: Students’ Constructed Responses:

An Interactional Approach to the Study of Achievement Gaps...............................127

4.1 Introduction ........................................................................................................129
  4.1.1 Analytic Framework .......................................................................................136

4.2 Data ....................................................................................................................140
  4.2.1 The QAM Test ..............................................................................................141
  4.2.2 Eastmond Classroom Students ....................................................................143

4.3 Analysis ..............................................................................................................149
  4.3.1 Schema Activation Miscues in the “Rolling Ball” Items .........................151
  4.3.2 Overcoming Schema Activation Miscues in the “Rolling Ball” Items ....159
  4.3.3 Schema Activation Miscues in the “Gravity and a Tennis Ball” Item ....167
  4.3.4 Overcoming Schema Activation Miscues in the “Gravity and a Tennis Ball” Items .................................................................176
  4.3.5 Summary ....................................................................................................182

4.4 Conclusion ..........................................................................................................183

Chapter 5: Conclusion..............................................................................................188

5.1 Summary ............................................................................................................188

5.2 Implications .......................................................................................................190

5.3 Contributions ....................................................................................................192

5.4 Limitations and Future Research .....................................................................197
Appendix A: Questions about Motion (QAM) Test ................................................. 201
Appendix B: Interpretive Argument for QAM Assessment from the Original SCALE- uP Researchers ...........................................................................................................207
Appendix C: Rating Guide for the QAM Test.............................................................210
Appendix D: Initial Item Characteristics Scores for Questions About Motion Test......................................................................................................................227

References.................................................................................................................. 229
LIST OF FIGURES

Figure 1.1 Rolling Ball Part I (QAM Test Item #8 and #9) .......................................................... 6
Figure 2.1 Example SPSS Document .................................................................................................. 18
Figure 2.2 Average U.S. NAEP Science Scores, 4th Grade (ELL and non-ELL) ......................... 22
Figure 2.3 Significance of Differences in NAEP Science Scores .................................................. 22
Figure 2.4 Average U.S. NAEP Science Scores, 8th Grade (ELL and non-ELL) ..................... 23
Figure 2.5 Significance of Differences in NAEP Science Scores .................................................. 23
Figure 2.6 Average MD NAEP Science Scores, 4th Grade (ELL and non-ELL) ..................... 24
Figure 2.7 Significance of Differences in NAEP Science Scores .................................................. 25
Figure 2.8 Average MD NAEP Science Scores, 8th Grade (ELL and non-ELL) ..................... 26
Figure 2.9 Significance of Differences in NAEP Science Scores .................................................. 26
Figure 3.1 Overview of Characteristics Contributing to Text Difficulty/Comprehensibility .......................................................................................................................... 60
Figure 3.2 Target Ideas and Corresponding QAM Items ................................................................. 66
Figure 3.3 Example Item (QAM Item 1) and Non-Coh-Metrix Variables ................................. 73
Figure 3.4 Procedure for Analysis of Research Questions ............................................................ 80
Figure 4.1 QAM Item #9 (“Rolling Ball Part I”) ........................................................................ 131
Figure 4.2 Sequential, Triadic Participant Structure of Test Interactions .................................... 138
Figure 4.3 Target Ideas of QAM Items ......................................................................................... 142
Figure 4.4 Eastmond Classroom Seating Chart ............................................................................ 146
Figure 5.1 Summary of Key Findings as Related to Construction of Achievement
LIST OF TABLES

Table 1.1 Difficulty and Discrimination Indices of the QAM Test Items.............................. 8
Table 2.1 Motion and Forces Test Scores (4-year average) ......................................................... 32
Table 2.2 Motion and Forces Test Scores (Year 2) ................................................................. 37
Table 2.3 Average Gains (Eastmond Classroom) ................................................................. 42
Table 3.1 Overview of Linguistic Characteristics of Test Items in Previous Studies .......... 51
Table 3.2 Twenty Initial Item Characteristics Variables ......................................................... 68
Table 3.3 Example Scores for QAM Item 1 on Non-Coh-Metrix Predictor Variables . 74
Table 3.4 Descriptive Statistics for Initial Predictor Variables ............................................... 82
Table 3.5 Creation of Ordinal Variables to Reduce Skewness and Enhance MLM Interpretation........................................................................................................................................ 83
Table 3.6 Descriptive Statistics for Transformed Predictor Variables .................................. 84
Table 3.7 Results from the Exploratory Univariate MLM Analyses of the Characteristics–Item-Level Scores Relationships (N=8640 for All Variables) .................. 86
Table 3.8 Results from Confirmatory Multivariate MLM Analyses ........................................ 94
Table 3.9 Descriptive Statistics of Students’ Item-Level Scores Across ESOL Groups ......................................................................................................................................................... 96
Table 3.10 Differences between Responses across ESOL Groups in Univariate MLM ......................................................................................................................................................... 97
Table 3.11 Overview of Findings for Format Variables Research Question Three .. 100
Table 3.12 Overview of Findings for Syntax Variables Research Question Three.... 102
Table 3.13 Overview of Findings for Lexis Variables Research Question Three ................................................................. 103

Table 3.14 Overview of Findings for Discourse Variables Research Question Three ................................................................................................................. 104

Table 3.15 Differences between Odds Ratios Across ESOL Groups for Significant Format Predictor Variables................................................................................................. 107

Table 3.16 Differences between Odds Ratios Across ESOL Groups for Significant Syntax Predictor Variables............................................................................................................. 108

Table 3.17 Differences between Odds Ratios Across ESOL Groups for Significant Lexis Predictor Variables................................................................................................................. 110

Table 3.18 Differences between Odds Ratios Across ESOL Groups for Significant Discourse Predictor Variables.................................................................................................................. 111

Table 3.19 Summary of Findings for Research Question Three................................................................................................................. 114

Table 4.1 Emergent Bilingual Students’ Constructed Responses that Addressed a Different Idea than the Target Idea of the Item ........................................................................................................ 133

Table 4.2 Native English Speaking Students’ Constructed Responses that Addressed a Different Idea than the Target Idea of the Item .................................................................. 133

Table 4.3 Overview of Achievement Gaps Across the Corpus as Compared with the Eastmond Classroom .................................................................................................................................................. 144

Table 4.4 Emergent Bilingual Students’ Reported Demographic Information .................................................................................................................. 147

Table 4.5 Native English Speaking Students’ Reported Demographic Information ................................................................................................................. 148
CHAPTER 1

INTRODUCTION

1.1. PROBLEM

Achievement gaps between emergent bilingual students and native English speaking students are the subject of much recent research across various fields, including general education (e.g. August & Hakuta, 1997; Lynch, Kuipers, Pyke, & Szesze, 2005; National Research Council, 2002) and testing literature (e.g. Abedi, 2006; Abedi & Gándara, 2006; Koenig & Bachman, 2004; Bailey, 2005; Farnsworth, 2008; Hudicourt-Barnes, Noble, Roseberry, Suarez & Warren, 2008; Kieffer, Lesaux, Rivera, & Francis, 2009; Logan-Terry & Wright, 2010; Martiniello, 2008; Ockey, 2007; Rivera & Collum, 2006; Shohamy, 2011; Wright & Logan-Terry, in preparation; Wolf, Herman, Bachman, Bailey, & Griffin, 2008; Wolf & Leon, 2009).¹

Researchers recognize that the rapid increases in the population of emergent bilinguals in U.S. schools—currently numbering over five million (Colombo & Furbush, 2009)—make achievement gaps for this growing group of students a pressing concern. Emergent bilingual students consistently rank among the lowest-scoring groups of students on large-scale, English-language content assessments such as the National Assessment of Educational Progress (NAEP) and the Trends in International Mathematics and Science Study (TIMSS), averaging approximately

---

¹ In other literature, this population of students is referred to as “English Language Learners”, “ELLs,” “ELs,” etc. I choose to refer to these students here as “emergent bilinguals” as I believe this more accurately highlights the nature of their situation and is supported in recent work (e.g. Garcia, 2010).
thirty percent lower than their native speaking peers in math and science achievement over the past decade (U.S. Department of Education, 2009).

Compounding the problem, the No Child Left Behind (NCLB) legislation—in place since 2001 and currently up for reauthorization—mandates that the content scores of even low English proficiency students be integrated into states’ accountability systems for schools and teachers, making achievement on large-scale content assessments extremely high stakes for students, teachers, and school systems (Menken, 2009).

Some researchers have argued that the measurement of achievement—predominantly conducted in English in the United States—may be contributing to achievement gaps (e.g. Abedi, 2004; Abedi & Lord, 2001; Cid, 2009; Farnsworth, 2006; Pollitt, Marriot, & Ahmed, 2000; Staley, 2005; Wright & Li, 2008). Many of these previous studies have examined the relationship between the linguistic complexity of content test items and differential performance of emergent bilingual and native English speaking students on the tests, an issue I further investigate in this dissertation work.

Much of the previous research has found that longer and more linguistically complex items tend to favor native English speaking students; however, the degree to which, and the ways in which, language functions as a source of construct-irrelevant difficulty (Messick, 1989) is still heavily debated. Indeed, contradictory findings in recent studies have caused researchers to question the degree to which factors other than language (e.g. legitimate differences in underlying content knowledge that are linked to socioeconomic issues, opportunities to learn in the
classroom, etc.) are contributing to achievement gaps between emergent bilingual and native English speaking students (e.g. Abedi & Herman, 2010; Koenig & Bachman, 2004; Elliot, 2008; Farnsworth, 2008, Gee, 2008; Ockey, 2007).

Based on this debate in the literature, the present dissertation study begins with the overarching research question, “Are there test-related sources of achievement gaps for emergent bilingual students in U.S. schools?”

1.2. SCALE-UP DATA

In order to address this question, I investigate achievement gaps in an existing corpus of middle school science data collected for a large-scale interdisciplinary research project Scaling up Curriculum for Achievement, Learning, and Equity Project (SCALE-up). The five-year (2001-2006), NSF-funded SCALE-up study which was housed at The George Washington University examined student achievement data and ethnographic classroom data as a means of researching the efficacy of three different research-based science curriculum units. By employing existing data from the SCALE-up corpus, I engage in a process of data discovery rather than the more traditional notion of data collection. As part of the data discovery process for this dissertation, I first examine which data are available in the SCALE-up corpus as well as which specific aspects of the corpus can be utilized in order to investigate achievement gaps for emergent bilingual students.

---

2 See http://www.gwu.edu/~scale-up/ for additional information about the original SCALE-up project.
Within the SCALE-uP corpus, achievement gaps between emergent bilingual and native English speaking students had already been identified as a potential area for future research (Lynch et al., 2005). The researchers’ original research questions dealt with the efficacy of inquiry-based science curricula that researchers hypothesized would lessen achievement gaps for minority students based on the Project 2061 standards. In a pilot study (Lynch et al., ibid), the original SCALE-uP researchers found that the inquiry-based curricula did indeed lessen achievement gaps for several minority subgroups of students, except for those students currently receiving English as a Second or Other Language (ESOL) services.

This dissertation study seeks to explore achievement gaps between emergent bilingual and native English speaking students in SCALE-uP data outside of the initial pilot study discussed in Lynch et al. (2005). In particular, I examine those data that serve as samples for the three analytic chapters of the dissertation: the four year implementation of the physics unit *Motion and Forces* overviewed in the background Chapter 2, a single year *Motion and Forces* implementation explored in the quantitative Chapter 3, and a specific 6th grade physics classroom at Eastmond Middle School in which this curriculum was enacted and is investigated in detail in the qualitative Chapter 4.

---

3 See http://www.project2061.org/ for more information on these standards.
1.2.1. The Questions about Motion Middle School Physics Test

My analysis centers around the Questions about Motion (QAM) 6th grade physics assessment that was utilized by the SCALE-uP researchers to measure achievement during the Motion and Forces curriculum implementations (please see Appendix A for reference to the full QAM assessment). I also discuss some of the key features of the assessment here in order to orient the reader, as it serves as the main basis for the analytic chapters in this dissertation.

The test features ten total items, with seven constructed response items and three selected response items. Figure 1.1 presents an example of a selected response and a constructed response from the QAM test:
Figure 1.1. Rolling Ball Part I (QAM Test Item #8 and #9)

A Rolling Ball

Part I. Your friend rolls a ball on the floor toward you. The ball continues to roll.

a. Is there a force constantly pushing or pulling the ball as it rolls toward you? Fill in the oval above your answer:

Yes
No

b. Explain your answer.

Students could receive full credit on the selected response portion of this item by selecting either yes or no, depending on how they explained their answers in part b. The answer that the test developers intended as the correct answer for this item was “no” for the selected response with an explanation that one force started the push, but does not constantly push on it as it moves; however, students could choose “yes” and explain that the force of friction is constantly acting on the ball. In this way, the item is meant to target the idea that an object at rest stays that way unless acted on by a force and an object in motion will continue to move unabated unless acted on by a force. This item proved particularly problematic for
both emergent bilingual and native English speaking students, and I explore the
factors contributing to this in both Chapter 3 and Chapter 4 of this dissertation.

Now let us consider some of the basic item characteristics of all the QAM items. Difficulty indices measure the percent correct on a given item. They range from one (indicating that 100% of students got the item correct) to zero (indicating that none of the students got the item correct). Thus, it is somewhat counterintuitive that a lower difficulty index indicates that an item is more difficult (Anastasi & Urbina, 1997).4

Discrimination indices also range from one to zero and calculate whether high-scoring students have a high probability of answering an item correctly and whether low-scoring students have a low probability of answering a question correctly. In the case of discrimination indices, the higher the discrimination index, the more discriminative the item is, which is more intuitive than the difficulty indices. In this way, discrimination can identify the degree to which items correctly differentiate among students (Crocker & Algina, 1986).

The following Table 1.1 presents the difficulty and discrimination indices and of the QAM items.

---

4 To avoid confusion, this is also referred to as “item facility” in some studies (e.g. Brown, 1988, 1989).
Table 1.1: Difficulty and Discrimination Indices of the QAM Test Items

<table>
<thead>
<tr>
<th>Item</th>
<th>Difficulty</th>
<th>Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td>QAM #1 (Pushing a Couch, selected response)</td>
<td>0.94</td>
<td>0.19</td>
</tr>
<tr>
<td>QAM #1 (Pushing a Couch, constructed response)</td>
<td>0.81</td>
<td>0.44</td>
</tr>
<tr>
<td>QAM #3 (A Full Truck, constructed response)</td>
<td>0.67</td>
<td>0.32</td>
</tr>
<tr>
<td>QAM #4 (Gravity and a Tennis Ball Part I.a., selected response)</td>
<td>0.76</td>
<td>0.39</td>
</tr>
<tr>
<td>QAM #5 (Gravity and a Tennis Ball Part I.b., constructed response)</td>
<td>0.48</td>
<td>0.62</td>
</tr>
<tr>
<td>QAM #6 (Gravity and a Tennis Ball Part II.a., constructed response)</td>
<td>0.66</td>
<td>0.64</td>
</tr>
<tr>
<td>QAM #7 (Gravity and a Tennis Ball Part II.b., constructed response)</td>
<td>0.74</td>
<td>0.48</td>
</tr>
<tr>
<td>QAM #8 (A Rolling Ball Part I.a., selected response)</td>
<td>0.31</td>
<td>0.22</td>
</tr>
<tr>
<td>QAM #9 (A Rolling Ball Part I.b., constructed response)</td>
<td>0.25</td>
<td>0.52</td>
</tr>
<tr>
<td>QAM #10 (A Rolling Ball Part II., constructed response)</td>
<td>0.54</td>
<td>0.61</td>
</tr>
</tbody>
</table>

In the table we can see that the QAM items provide varying degrees of differential information. For example, QAM #1 “Pushing a Couch, selected response” has a difficulty index very close to 1.0 and does not provide much differential information. In contrast, other items seemed to have performed better on the QAM assessment (e.g. QAM #2 “Pushing a Couch, constructed response” which has a discrimination index very close to 0.4). Characteristics of the QAM test items are further explored in the analytic Chapters 3 and 4 of this dissertation. In particular, item characteristics are discussed in terms of how they relate to the sources of achievement gaps for emergent bilingual students in the analytic chapters of this dissertation.
1.2.2. Emergent Bilingual and Native English Speaking Students

In selecting samples for the analyses in this dissertation, I focus on those students currently receiving ESOL services and those students who had never received ESOL services. Those students currently receiving ESOL services represent emergent bilingual students in this study, and those students never having received ESOL services represent native English speaking students. A third category of students, those who had previously received ESOL services, are excluded from the present analyses. Although this group of prior ESOL students is important, they represent a very heterogeneous group with some students having recently exited ESOL services along with others having been exited from ESOL services for half a decade. Therefore, I chose to conduct analyses based on the two ESOL groups that most clearly represent differences between emergent bilingual and native English speaking students.

The limitations of this ESOL status-based operationalization are numerous (e.g. “native English speaking students” may in fact speak an additional language at home yet not have been enrolled in ESOL services, “emergent bilingual students” no doubt vary in terms of their English proficiency, etc.). The limitations of the ESOL status-based categorization of students are also discussed in other chapters of the dissertation; however, this ESOL categorization is the only language status grouping variable available in the SCALE-uP corpus and is thus the most reliable measure to use in the present study.
An additional limitation inherent in analyzing the existing SCALE-uP data is the issue of generalizability and the representativeness of the student participants. SCALE-uP data were collected in Montgomery County, Maryland. As I show in the backdrop analytic Chapter 2 of this dissertation, students in Maryland state schools tend to outperform students from most other states in the U.S. on large-scale, standardized tests. There are also less severe achievement gap problems for emergent bilingual students in Maryland as compared with other states, as I show in the analyses in Chapter 2. In addition, the specific school system from which the SCALE-uP data were collected, Montgomery Country Public Schools, has been chosen by Pearson as a flagship school system in recent years.5

In sum, this dissertation investigates achievement gaps in the SCALE-uP corpus that was collected in the relatively anomalous context of Montgomery County, MD and, thus, the findings are not meant to be generalizable to other contexts. However, my work does endeavor to pilot innovative methodologies and shed light on test-related sources of achievement gaps that might be pertinent in other data sets.

1.3. THEORECTICAL FRAMEWORK

The chapters presented in this dissertation can be interpreted as studies of the validity of the QAM assessment for emergent bilingual students. By asking, "Are there test-related sources of achievement gaps for emergent bilingual students in

5 For additional information please see http://www.montgomeryschoolsmd.org/press/index.aspx?page=showrelease&id=2778
U.S. schools?” and investigating this question in relation to the SCALE-uP data, I also provide evidence as to the validity of the QAM test utilized in the original SCALE-uP study.

Indeed, the sort of evidence that I provide in my analyses (e.g. evidence or lack thereof of linguistic bias in test items) has long been a central component of test validity theory (e.g. Cronbach, 1971; Messick, 1989; Kane, 1992), and is a common feature in more recent work in educational assessment theory (e.g. Kane, 2006, 2010a, 2010b; Mislevy, Steinberg & Almond, 2003). In fact, Kane (2010b: 1) defines validity as “the degree to which a proposed interpretation is justified by evidence.” Researchers have provided evidence of the differential effects of linguistic characteristics of test items in order to add to the growing body of literature that argues test prompt language contributes to achievement gaps for emergent bilingual students (e.g. Abedi, 2006; Farnsworth, 2006; Martiniello, 2008; Pollitt, Marriot, & Ahmed, 2000; Staley, 2005; Wolf & Leon, 2009).

As McNamara (2006: 32) points out, “tests are procedures for gathering evidence. The word evidence suggests a parallel with the role of evidence in law, and empirical validation of test-score interpretation has parallels in legal processes.” Indeed, Michael Kane, one of the main proponents of an evidence-based approach to validity, has a background in the legal realm. Kane’s (1992, 2006, 2010a, 2010b) framework includes an interpretive argument that expresses the logic behind all steps in the testing process: from performances to uses and interpretations of test scores. This approach also builds upon Toulmin (2003 [1958])’s idea of warrants that allow for inferences, as these warrants and the resulting inferences must be
based on evidence in some form (e.g. statistical models, psychometric models, and general logic).

Kane’s (1992, 2006) model also includes a validity argument, which provides an assessment of the evidence for and against the interpretive argument of a test. This validity argument generally relies on various analyses and evidence from empirical and judgment-based sources. The work in this dissertation can be seen as contributing to the construction of a validity argument with regard to the Questions about Motion (QAM) test from the SCALE-uP corpus, in that it systematically investigates the interpretive argument of the QAM assessment by investigating test-related sources of achievement gaps between emergent bilingual and native English speaking students.6

In addition to these general claims that are not specific to content testing, the Bachman & Palmer (1996) model of testing research suggests scoring content and language separately, to the degree that it is possible. In an application of this model to content tests, Weigle & Jensen (1997) show how the Bachman and Palmer model can be successfully applied across contexts. Farnsworth (2008) has also more recently applied this framework fruitfully in the area of content testing for emergent bilingual students. Farnsworth (ibid: 33) investigates the interpretations made based on standardized math and science tests mandated under NCLB, thereby serving as evidence in the validity argument for these tests by shedding light on whether they are “equitable, impartial, relevant, and generalizable.”

---

6 Some aspects of the interpretive argument of the QAM assessment can be found in Appendix B and Appendix C. These show insight into the SCALE-uP researchers’ processes in creating the QAM assessment.
In contrast, some researchers have pointed out the limitations of adopting the Bachman & Palmer model for content tests. For example, Short (1993) argues that it is impossible to separate language and content features on assessments, and argues for alternative forms of assessment for language learners in content subjects (e.g. checklists, portfolios, interviews, performance-based tasks, etc.). Other researchers have also criticized the validity of current ways of assessing language learners’ content knowledge, arguing for more dynamic assessment given the dynamic nature of language and content (Poehner & Lantolf, 2005; Shohamy, 2006).

The validity of the QAM test also relates to another foundational theoretical concept that is discussed throughout this dissertation: gatekeeping (Erickson, 1975; Erickson & Schultz, 1982). Gatekeeping is defined as a linguistic event in which a member of a minority community requires the approval of the majority community. In this way, gatekeeping constitutes a linguistic basis of discrimination against minority populations. Other researchers have discussed gatekeeping in educational contexts (Wrigley, 1993), gatekeeping events for emergent bilingual students (Duff, 2001), and assessment as gatekeeping (Moore, 1996; Short, 1993; Tarone & Kuehn, 2000).

My dissertation effort seeks to understand whether there are language-related validity problems with the QAM test. In this way, I also endeavor to discover whether emergent bilingual students’ interactions with test items can be considered part of a crucial gatekeeping conversation with the educational institution. In other words, if emergent bilingual students’ performance on content tests is affected by linguistic difficulties, then this would provide evidence of validity issues on these
tests; in addition, this would provide evidence that these tests also constitute
gatekeeping encounters because linguistic issues contribute to limitations to access
to resources and opportunities for the minority population of emergent bilingual
students.

1.4. STUDY DESIGN

In order to address the questions and issues discussed above surrounding
achievement gaps, the validity of the QAM test, and gatekeeping encounters, this
dissertation employs a sequential model of mixed methods research (Creswell,
2008). Chapter 2 provides a large-scale, quantitative backdrop of achievement gaps
for emergent bilingual students across national and local levels. Chapter 3
quantitatively investigates the linguistic complexity of item prompts in the QAM
assessment as a source of achievement gaps. In the final analytic Chapter 4, I
provide qualitative analyses that shed further light on the discursive processes that
contribute to the patterns identified in the previous quantitative analyses.

As such, qualitatively-oriented researchers who read this dissertation may
find the first two analytic chapters to be lacking in exemplification and richness of
data. In turn, quantitatively-oriented researchers may find Chapter 4 to be lacking in
generalizability to larger data sets. Throughout the dissertation, I attempt to engage
with diverse paradigms and audiences, although I realize certain chapters and
claims may resonate with readers to varying degrees, depending on their respective
analytic priorities. Regardless of readers’ individual analytic orientations, I hope my work here serves provocative and interesting.
CHAPTER 2

ACHIEVEMENT GAPS:

Setting the Analytic Scene

2.1. INTRODUCTION

As explained in Chapter 1, the goal of my larger dissertation study is to investigate the issue of achievement gaps by exploring the sources of differences between students’ scores in the existing corpus of middle school science data from the SCALE-uP project described in Chapter 1, Section 1.2. The present chapter sets the analytic backdrop for the dissertation by detailing the achievement gaps for emergent bilingual students in the SCALE-uP corpus data that I analyze, in light of the national and state-level contexts surrounding the SCALE-uP corpus.

Within the SCALE-uP corpus, achievement gaps between emergent bilingual and native English speaking students have already been identified (Lynch et al., 2005). The researchers’ original research questions dealt with the efficacy of an inquiry-based chemistry curriculum, Chemistry That Acts (CTA). The researchers hypothesized that this type of curriculum unit would lessen achievement gaps for minority students (e.g. ethnic minorities, socioeconomically and linguistically diverse students, etc.) based on the Project 2061 standards. They found that CTA did indeed lessen achievement gaps for all subgroups of students, except for those

---

7 See http://www.project2061.org/ for more information on these standards.
students currently receiving English as a Second or Other Language (ESOL) services.

Lynch et al. (ibid: 942) conclude:

This suggests that the instructional characteristics of CTA supported the conceptual learning of the majority of students, and were more effective than the standard fare. The only exception was that the CTA and comparison group students currently eligible for ESOL services showed similar results for the Conservation of Matter Assessment. Perhaps the literacy demands of CTA are too great and need modification, or perhaps our assessment failed to capture the learning gains of these students. The goal orientation and engagement results indicate that most of these students viewed CTA about the same as the regular curriculum, with surprisingly high scores reported on these scales. These results require more exploration. (emphasis my own)

These conclusions were based on one year of the CTA implementation during the SCALE-uP study. The present analysis seeks to explore achievement gaps between emergent bilingual and native English speaking students across other subsets of the SCALE-uP data, particularly those that serve as data samples for the subsequent analytic chapters of the dissertation: the Physics unit *Motion and Forces* implementation explored in Chapter 3, as well as a specific 6th grade physics classroom at Eastmond Middle School in which this curriculum was enacted, and the specific students at Eastmond whose tests are investigated in Chapter 4. Based on the national trends and the findings of the original SCALE-uP researchers, I hypothesize that there will be significant differences between pre-test scores, post-
test scores, and students’ gains between emergent bilingual and native English speaking students across the SCALE-uP corpus.

2.2. METHOD

The existing SCALE-uP data are organized in SPSS documents, and an example of one student’s data can be found in Figure 2.1 below.

*Figure 2.1. Example SPSS Document*

<table>
<thead>
<tr>
<th>GWID</th>
<th>gender</th>
<th>ethnic</th>
<th>farms</th>
<th>esol</th>
<th>scieGPA</th>
<th>grlv</th>
<th>Pre</th>
<th>Post</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>52045</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>3.00</td>
<td>6</td>
<td>33.99</td>
<td>31.20</td>
<td>-2.79</td>
</tr>
</tbody>
</table>

As seen in Figure 2.1, the SCALE-UP researchers gave students an anonymous ID number and categorized them according to: gender (1 = female, 0 = male), ethnicity (1 = ‘Asian American,’ 2 = ‘African American,’ 3 = ‘White,’ 4 = ‘Hispanic’), free and reduced meals program status (FARMS; a proxy for socio economic status; 0 = never having received FARMS, 1 = previously having received FARMS, 2 = currently receiving FARMS). In addition, the researchers recorded students’ grade point average in science class before the implementation of the curriculum unit (science GPA) as well as the students’ grade level (6 for the 6th grad students receiving the physics curriculum unit, 7 for the 7th grade students receiving the climate/earth science curriculum unit, and 8 for the 8th grade students receiving the chemistry curriculum unit). As for ESOL status, the students were labeled as
'ESOL0' (never having received ESOL services), 'ESOL1' (previously having received ESOL services), or 'ESOL2' (currently receiving ESOL services).

However, here and throughout the rest of the dissertation I choose to only study those students currently receiving ESOL services as the "emergent bilingual category" and only those students never having received ESOL services as the native English speaker category."\(^8\) A third category of students, those who had previously received ESOL services, was excluded from analysis throughout my study. Although this group of prior ESOL students is important, they represent a very heterogeneous group with some students having recently exited ESOL services along with others having been exited from ESOL services for half a decade. Much of the previous literature would suggest that these ESOL1 (previously having received ESOL services) students should be treated as emergent bilinguals because it takes decades for immigrant students to become proficient in the dominant language, despite having been exited from ESOL or other language services (e.g. Levin & Shohamy, 2008). This is one of the limitations of using existing data, as I do not have access to proficiency scores of any kind for each of the students. Therefore, I have chosen to explore achievement gaps in the SCALE-uP corpus across the two ESOL groups that most clearly represent differences between emergent bilingual and native English speaking students.

\(^8\) The limitations of equating native English speaking or bilingual status with enrollment in ESOL services are also discussed elsewhere in the dissertation; however, this categorization is the only language status grouping variable available in the SCALE-uP corpus and is thus the most reliable measure to use in the present study.
In this chapter, I first discuss issues surrounding achievement gaps at the national and state level in the U.S., drawing on analysis from the "data explorer" tool on the website for the U.S. Department of Education Institute for Education Sciences National Center for Education Statistics in Section 2.3.1. Then, in Section 2.3.2, I turn to achievement gaps across the entire SCALE-uP corpus, primarily drawing on year-end reports from the original SCALE-uP researchers, which include the chemistry, climate and physics curriculum implementations. In Section 2.3.3, I analyze the four years of physics implementation data that serves as the sample for my Chapter 3 analysis. I first present the descriptive statistics for these data, then present the results of a series of t-tests on the raw scores and weighted pretests and the posttests to test for significant differences across students’ mean scores. Finally, I report the results of repeated measures/mixed ANOVAs, showing that there were significant differences between the gains from pre- to posttest. All of these serve to provide a backdrop picture of the achievement gaps for emergent bilinguals across the entire Motion and Forces curriculum implementation (2002-2006).

Section 2.3.4 presents a similar analysis as the previous section (t-tests to test for significant differences across means on pretests and posttests, as well as repeated measures ANOVAs to test significant differences between gains); however, in this section I focus only on the Year 2 Motion and Forces implementation, as this was the year that I sample from for the analyses in Chapter 3 and Chapter 4, because

---

9 The data explorer tool is available at: http://nces.ed.gov/
both the achievement and ethnographic data are most available and interesting for this year of implementation.\textsuperscript{10}

Sections 2.3.5 deals with achievement gaps in the Eastmond classroom used for the qualitative study of students’ constructed responses in Chapter 4. This section also discusses issues surrounding representativeness of the samples I have chosen for the analyses in this dissertation, in light of the national, state, and corpus-wide comparison data provided in previous sections. All of these findings are then discussed in the conclusion as a backdrop and impetus for the analyses in subsequent analytic chapters of the dissertation.

2.3. FINDINGS

2.3.1. The National and State Backdrop

The most recent data available on national achievement gaps for emergent bilingual students are from the Science portion of the 2009 National Assessment of Educational Progress (NAEP). The following Figures 2.2 and 2.3 show the average scores on NAEP science for 4th grade emergent bilingual students (ELL) and their peers (non-ELL) in U.S. public schools, as well as tests for significance between these two means.

\textsuperscript{10} Further explanation of the decision to focus on Year 2 of \textit{Motion and Forces} as well as the Eastmond classroom is discussed in the motivations within the subsequent analytic chapters 3 and 4.
As the above figures show, the 4th grade emergent bilingual students in U.S. public schools had overall lower average scores on the NAEP - Science assessment (114) than their native English speaking peers (153). These differences are also
shown to be significant using the Institute for Education Sciences tests for statistical significance.\(^{11}\)

The following Figures 2.4 and 2.5 show the average scores on NAEP science for 8th grade emergent bilingual students (ELL) and their peers (non-ELL) in U.S. public schools, as well as tests for significance between these two means.

**Figure 2.4. Average U.S. NAEP Science Scores, 8^{th} Grade (ELL and non-ELL)**

<table>
<thead>
<tr>
<th>Year</th>
<th>Jurisdiction</th>
<th>Average scale score</th>
<th>Standard Error</th>
<th>Average scale score</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>National public</td>
<td>103 (1.0)</td>
<td>111 (2.3)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: The NAEP Science scale ranges from 0 to 300. Some apparent differences between estimates may not be statistically significant.


**Figure 2.5. Significance of Differences in NAEP Science Scores**

<table>
<thead>
<tr>
<th>ELL (103)</th>
<th>Not ELL (151)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diff = -48</td>
<td>P-value = 0.0000</td>
</tr>
</tbody>
</table>

LEGEND:

< Significantly lower.

> Significantly higher.

\(^{11}\) It is unclear which statistical tests are used in the IES "data explorer" tool to determine these significance values.
As the above figures show, the 8th grade emergent bilingual students in U.S. public schools had overall lower average scores on the NAEP - Science assessment (103) than their native English speaking peers (151). These differences are also shown to be significant using the Institute for Education Sciences tests for statistical significance. In total, these four figures show that achievement gaps between emergent bilingual students and native English speaking students are significant at the national level.

Moving to the state-level data, the following Figures 2.6 and 2.7 show the average scores on NAEP science for 4th grade emergent bilingual students (ELL) and their peers (non-ELL) in Maryland public schools, as well as tests for significance between these two means.

*Figure 2.6. Average MD NAEP Science Scores, 4th Grade (ELL and non-ELL)*
As the above figures show, the 4th grade emergent bilingual students in Maryland public schools had overall lower average scores on the NAEP - Science assessment (130) than their native English speaking peers (151). These differences are also shown to be significant using the Institute for Education Sciences tests for statistical significance. These figures also show that the achievement gaps for 4th grade emergent bilingual students are less severe in Maryland than in the national average presented above.

The following Figures 2.8 and 2.9 show that, unfortunately, the reporting standards were not met for 8th grade emergent bilingual students for the 2009 NAEP - Science assessment in Maryland.
Figure 2.8. Average MD NAEP Science Scores, 8th Grade (ELL and non-ELL)

![Table Image]

<table>
<thead>
<tr>
<th>Year</th>
<th>Jurisdiction</th>
<th>ELL</th>
<th>Average scale score</th>
<th>Standard Error</th>
<th>Not ELL (149)</th>
<th>Average scale score</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>(1)</td>
<td>149</td>
<td>(1.1)</td>
<td></td>
</tr>
</tbody>
</table>

* Not applicable

NOTE: The NAEP science scale ranges from 0 to 300. Some apparent differences between estimates may not be statistically significant.


Figure 2.9. Significance of Differences in NAEP Science Scores

![Diagram Image]

Based on the national trends for both grades 4 and 8, we can hypothesize that Maryland’s data for the Grade 8 NAEP Science assessment would reflect similar
trends: more severe achievement gaps for the older students, but overall lower gaps in the Maryland context as compared with the national average. This has implications for my dissertation study, as perhaps the achievement gaps that I explore in the SCALE-uP corpus—6th grade science students in Maryland public schools—are somewhat less severe than they would be in a different state context within the U.S.\(^{12}\)

### 2.3.2. SCALE-uP Achievement Gaps across the Corpus

As mentioned above, in the report from the pilot year of the SCALE-uP study (Lynch et. al, 2005), achievement gaps for emergent bilinguals were prominent. Throughout the subsequent years of the study, emergent bilingual students continued to have overall lower scores than their native English speaking peers; however, their gains using the treatment curriculum were sometimes comparable to the native English speaking students. Rethinem, Pyke, & Lynch (2008) summarize the overall process and findings of the study in the following way:

The results of the first implementation study of a curriculum unit developed in the USA called Chemistry That Applies (CTA [State of Michigan, 1993]) conducted by SCALE-uP in 2001-02 supported the effectiveness of the unit and the

---

\(^{12}\) In fact, as mentioned in the Introductory Chapter 1 of this dissertation, Montgomery Country Public Schools have been chosen by Pearson as a flagship school system. For additional information please see http://www.montgomeryschoolsmd.org/press/index.aspx?page=showrelease&id=2778
utility of ANCOVA for detecting group differences. The first CTA study and its replication the following year met the expectations that a curriculum unit with a high rating for instructional quality (American Association for the Advancement of Science [AAAS], 2001) would result in positive outcomes overall and significant differences between favor of the new curriculum unit. The ANCOVA analyses showed that students’ understanding of the target ideas as measured by the post-test assessment was higher with the new unit and analyses of disaggregated student demographic data (gender, ethnicity/race, socioeconomic status [SES], English language literacy and special education status) showed that subgroups of students learning with the new condition nearly always outscored their peers in the comparison condition (Lynch et al., 2005, 2007).

In 2004, SCALE-uP conducted quasi-experiments on two additional curriculum units: Motion and Forces (M&F [Harvard-Smithsonian Center for Astrophysics, 2001]) and Real Reasons for the Seasons (Seasons [Lawrence Hall of Science, 2000]). These two units did not rate as highly in instructional quality as the first unit (CTA), although these materials do have instructional characteristics more promising than traditional US science textbooks (Ochsendorf et al., 2006;
O’Donnell et al., 2006). Unlike the first unit investigated, in
which overall effects and effects when the data were
disaggregated mirrored each other, the ANCOVA results for
these units showed different profiles of effects for
disaggregated data.

In the Year 2 implementation of *Motion and Forces*—the year from which the
data for my dissertation were sampled—a 2 X 3 between-groups Analysis of
Covariance (ANCOVA) was conducted to evaluate the interaction of curriculum
condition and ESOL Status on M&F posttest scores. Independent variables were
curriculum condition (treatment and comparison) and ESOL status (Never ESOL,
Prior ESOL, and Now ESOL). The covariate was the study sample pretest M&F score
(mean = 41.52). ANCOVA indicated no statistically significant interaction between
curriculum condition and ESOL. However, there was a statistically significant main
effect for ESOL, $F(2, 2165) = 32.19, p < .05$. These results suggest that the magnitude
of the effect of the curriculum unit for each subgroup was similar. Follow-up tests
were performed to assess pairwise differences among the three subgroups for the
main effect of ESOL. These tests indicated that the Never ESOL (native English
speaking students in my study) subgroup mean was significantly higher than the
Prior ESOL and Now ESOL (emergent bilingual students in my study) subgroup
means across both conditions.

For Year 3 of the *Motion and Forces* implementation, the SCALE-uP
researchers conducted a 2 X 3 between-groups Analysis of Covariance (ANCOVA) to
evaluate the interaction of curriculum condition and ESOL Status on M&F posttest scores. Independent variables were curriculum condition (treatment and comparison) and ESOL status (Never ESOL, Prior ESOL, and Now ESOL). The covariate was the study sample pretest M&F score (mean = 38.04). ANCOVA indicated no statistically significant interaction between curriculum condition and ESOL. However, there was a statistically significant main effect for ESOL, $F(2, 2247) = 17.754, p < .05$. These results suggest that the magnitude of the effect of the curriculum unit for each ESOL subgroup was similar. Follow-up tests were performed to assess pairwise differences among the three subgroups for the main effect of ESOL. These tests indicated that the Never ESOL subgroup mean was significantly higher than the Prior ESOL and Now ESOL subgroup means and that the Prior ESOL subgroup mean was significantly higher than the Now ESOL subgroup mean.

In Year 4 of the Motion and Forces (M&F) physics implementation, the researchers employed a 2 X 3 between-groups Analysis of Covariance (ANCOVA) to evaluate the interaction of curriculum condition and ESOL Status on M&F scores. Independent variables were curriculum condition (treatment and comparison) and ESOL status (Never ESOL, Prior ESOL, and Now ESOL). The ANCOVA indicated no statistically significant interaction between curriculum condition and ESOL status. However, there was a statistically significant main effect for ESOL, $F(2, 1755) = 68.67, p < .05$. These results suggest that the magnitude of the effect of the

---

13 There were actually two separate implementations of Motion and Forces conducted in Montgomery Country during this academic year; thus, despite covering only 3 academic years of results, we have four sets of implementation results.
curriculum unit for each ESOL subgroup was similar. Follow-up tests were performed to assess pairwise differences among the three subgroups for the main effect of ESOL. These tests indicated that the Never ESOL subgroup mean was significantly higher than the Prior ESOL and Now ESOL subgroup means and that the Prior ESOL subgroup mean was significantly higher than the Now ESOL subgroup mean.

2.3.3. SCALE-uP Achievement Gaps in Physics Implementations (Four Years)

At this stage in the analytic backdrop chapter, I focus solely on the physics data, and modify the original SCALE-uP SPSS documents to include only data from ESOL 0 (native English speaking) and ESOL 2 (emergent bilingual) students. Also, I include both the treatment and comparison condition data, as the effect of the Motion and Forces curriculum on achievement gaps is not the focus of my study, although it was of course the focus of the larger SCALE-uP study. Another difference in my analysis here is that I separate out students’ raw scores from their weighted scores and analyze each separately, unlike the original SCALE-uP researchers.14

In this section, I analyze the four years of physics implementation data that serve as the sample for my Chapter 3 analysis. I first present the descriptive statistics for these data, then present the results of a series of $t$-tests on the raw

---

14 In the original SCALE-uP study, student responses to the constructed response items were judged by trained raters; raw scores (sum of correct responses) were determined and a panel of experts and GWU researchers engaged in an analytic judgment process for standard setting (Plake & Hambleton, 2001). Then, the expert panel decided to weight Target Idea #1 as 60% of the total score, Target Ideas #2 and #3 as 15% each, and Target Idea #4 as only 10%.
score and weighted pretests and the posttests testing for significant differences across students’ mean scores. Finally, I report the results of repeated measures/mixed ANOVAs to determine whether there were significant differences between the gains from pre- to posttest. All of these serve to provide a backdrop picture of the achievement gaps for emergent bilinguals across the entire Motion and Forces curriculum implementation (2002-2006).

The following Table 2.1 presents descriptive statistics for the raw scores and weighted scores for the pretest and posttest data for both ESOL groups of students. In the case of some years of the SCALE-uP implementation, the larger database (which was composed in the last year of the SCALE-uP study) did not include raw scores for all students, as the researchers simply worked off of the weighted scores for analyses based on the combined database. Thus, I limit this descriptive table and subsequent parametric tests to just those students with both the raw and weighted scores available in the larger corpus.

Table 2.1. Motion and Forces Test Scores (4-year average)

<table>
<thead>
<tr>
<th>ESOL Level</th>
<th>Score</th>
<th>Mean</th>
<th>SD</th>
<th>N size</th>
<th>Range</th>
<th>Skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ESOL 0 (Native English Speakers)</strong></td>
<td>Raw Score Pretest</td>
<td>4.19</td>
<td>2.0</td>
<td>7281</td>
<td>10</td>
<td>.27</td>
</tr>
<tr>
<td></td>
<td>Weighted Score Pretest</td>
<td>38.72</td>
<td>19.4</td>
<td>7281</td>
<td>100</td>
<td>.46</td>
</tr>
<tr>
<td></td>
<td>Raw Score Posttest</td>
<td>5.37</td>
<td>2.2</td>
<td>7281</td>
<td>10</td>
<td>-.09</td>
</tr>
<tr>
<td></td>
<td>Weighted Score Posttest</td>
<td>55.18</td>
<td>23.4</td>
<td>7281</td>
<td>100</td>
<td>-.06</td>
</tr>
<tr>
<td><strong>ESOL2 (Emergent Bilinguals)</strong></td>
<td>Raw Score Pretest</td>
<td>3.18</td>
<td>1.9</td>
<td>484</td>
<td>10</td>
<td>.55</td>
</tr>
<tr>
<td></td>
<td>Weighted Score Pretest</td>
<td>27.25</td>
<td>16.4</td>
<td>484</td>
<td>97.20</td>
<td>.85</td>
</tr>
<tr>
<td></td>
<td>Raw Score Posttest</td>
<td>3.88</td>
<td>2.2</td>
<td>484</td>
<td>10</td>
<td>.36</td>
</tr>
<tr>
<td></td>
<td>Weighted Score Posttest</td>
<td>37.55</td>
<td>23.2</td>
<td>484</td>
<td>100</td>
<td>.52</td>
</tr>
</tbody>
</table>

| ESOL0 (NS) Raw Score Gain | 1.18 |
| ESOL0 (NS) Weighted Score Gain | 16.46 |
Table 2.1 shows that native English speaking students had higher averages across all pretests, posttests, and gain scores. I now present a series of t-tests to determine whether the differences between these mean scores are statistically significant.

Independent t-test on raw score pretest

On average, native English speaking students have higher raw pretest scores (M = 4.19, SE = .024) than emergent bilingual students (M = 3.18, SE = .087). This difference is significant, t(558) = 11.145, p < .01. It represents a large effect size $r = 0.427$ ($d = .944$).15

Independent t-test on weighted score pretest

On average, native English speaking students have higher weighted pretest scores (M = 38.72, SE = .225) than emergent bilingual students (M = 27.15, SE = .733). This difference is significant, t(194) = 8.309, p < .01. It represents a large effect size $r = .526$ ($d = 1.236$).

Independent t-test on raw score posttest

On average, native English speaking students have higher raw posttest scores (M = 5.37, SE = .023) than emergent bilingual students (M = 3.88, SE = .093). This difference is significant, t(558) = 11.145, p < .01. It represents a large effect size $r = 0.427$ ($d = .944$).

15 I use as a general baseline for relative effect sizes Cohen (1988)'s definition of "small, $d = .2$," "medium, $d = .5$," and "large, $d = .8$."

<table>
<thead>
<tr>
<th></th>
<th>ESOL2 (EB) Raw Score Gain</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.70</td>
</tr>
<tr>
<td>ESOL2 (EB)</td>
<td></td>
<td>10.30</td>
</tr>
<tr>
<td>Weighted</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
difference is significant, \( t(9115) = 15.952, p < .01 \). It represents a small/medium effect size \( r = 0.165 \) \((d = .334)\).

**Independent t-test on weighted score posttest**

On average, native English speaking students have higher raw posttest scores \((M = 55.18, SE = .250)\) than emergent bilingual students \((M = 37.95, SE = .955)\). This difference is significant, \( t(9302) = 17.754, p < .01 \). It represents a small/medium effect size \( r = 0.181 \) \((d = .368)\)

**Repeated measures ANOVAs to test for significance of gains (raw scores)**

Field (2009: 348) discusses the superiority of repeated measures ANOVA over use of a \( t \)-test on average gain scores (minimization of familywise error). Thus, I run a two by three repeated measures/mixed ANOVA with time (pre- and posttest) as the within-subjects measure and ESOL status as the between-subjects measure to test for significant differences in the gain scores.

There is a significant main effect of time on students’ raw scores, \( F(1, 7338) = 401.686, p < .01 \). This represents a medium effect size, \( r = .228 \) \((d = .468)\). In other words, students do significantly better on the posttest than on the pretests.

There is also a significant interaction effect between time and ESOL group status, \( F(1, 7338) = 12.712, p < .01 \); however, it represents a much smaller effect size, \( r = .045 \) \((d = .090)\). In other words, there are significant differences between score gains across ESOL groups, but this only accounts for a very small amount of the variation in the sample.
Repeated measures ANOVAs to test for significance of gains (weighted scores)

There is a significant main effect of time on students' weighted scores, $F(1, 7605) = 570.970, p < .01$. This represents a medium effect size, $r = .265 (d = .549)$. In other words, students do significantly better on the posttest than on the pretests.

There is also a significant interaction effect between time and ESOL group status, $F(1, 7605) = 20.358, p < .01$; however, it represents a much smaller effect size, $r = .056 (d = .110)$. In other words, there are significant differences between score gains across ESOL groups, but this only accounts for a very small amount of the variation in the sample.

In sum, the t-tests and repeated measures/mixed ANOVAs demonstrate statistically significant achievement gaps across the four years of *Motion and Forces* implementation in the SCALE-uP corpus. The independent t-tests conducted on the raw scores and weighted scores pretest data demonstrated significant main effects, with large effect sizes. The independent t-tests conducted on the raw scores and weighted scores posttest data demonstrates significant main effects, with small/medium effect sizes.

The differences in the magnitude of effect sizes across the t-tests for pretest and posttests may be due to the fact that these four years of *Exploring Motion and Forces* data contain both treatment and comparison group students; therefore, efficacy of the treatment curriculum condition could have reduced the effect size of
the significant differences in means for the posttest data (half of which was collected after the treatment curriculum, which was designed to reduce achievement gaps).

In summary of the ANOVA findings for the four years of physics data: the repeated measures ANOVAs reveal a significant main effect of time on students’ raw scores as well their weighted scores, both of which having a medium effect size. In other words, students do significantly better on the posttest than on the pretests. As for the achievement gaps in the ANOVA data, there is also a significant interaction effect between time and ESOL group status for both raw and weighted scores; however, in both cases they represent a much smaller effect size. In other words, there are significant differences between score gains across ESOL groups, but this only accounts for a very small amount of the variation in the sample.

2.3.4. SCALE-uP Achievement Gaps in the Year 2 Physics Implementation

At this stage in the analytic backdrop chapter, I focus solely on the physics data from the Year 2 implementation (because this is the year from which I sample for the qualitative analysis in Chapter 4, as the qualitative data available for this year are superior, as discussed in the introduction to Chapter 4), and modify the original SCALE-uP SPSS documents to include only data from ESOL 0 (native English speaking) and ESOL 2 (emergent bilingual) students. Also, I include both the treatment and comparison condition data, as the effects of the Motion and Forces curriculum on achievement gaps were not the focus of my study, although they were of course the focus of the larger SCALE-uP study. Another difference between the
original SCALE-uP analyses described above and my analyses here is that I separate out students' raw scores from their weighted scores and analyze each separately, unlike the original SCALE-uP researchers.\textsuperscript{16}

In this section, much like the previous section that dealt with all four years of the Motion and Forces implementation, I first present descriptive statistics, and then present $t$-tests to test for significant differences across means on pretests and posttests, as well as repeated measures ANOVAs to test significant differences between gains of just the Year 2 Motion and Forces implementation. These analyses serve as analytic backdrop for Chapters 3 and 4 of the dissertation by describing the achievement gap problem that motivates those investigations.

The following Table 2.2 presents descriptive statistics for the raw scores and weighted scores for the pretest and posttest data for both ESOL groups of students.

\textit{Table 2.2. Motion and Forces Test Scores (Year 2)}

<table>
<thead>
<tr>
<th>ESOL Level</th>
<th>Score</th>
<th>Mean</th>
<th>SD</th>
<th>N size</th>
<th>Range</th>
<th>Skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESOL 0 (Native English Speakers)</td>
<td>Raw Score Pretest</td>
<td>5.33</td>
<td>1.9</td>
<td>1800</td>
<td>10</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td>Weighted Score Pretest</td>
<td>43.37</td>
<td>19.15</td>
<td>1800</td>
<td>95.69</td>
<td>.32</td>
</tr>
<tr>
<td></td>
<td>Raw Score Posttest</td>
<td>6.37</td>
<td>1.9</td>
<td>1800</td>
<td>10</td>
<td>-.31</td>
</tr>
<tr>
<td></td>
<td>Weighted Score Posttest</td>
<td>58.47</td>
<td>21.5</td>
<td>1800</td>
<td>95.69</td>
<td>-.22</td>
</tr>
<tr>
<td>ESOL 2 (Emergent Bilinguals)</td>
<td>Raw Score Pretest</td>
<td>3.80</td>
<td>1.9</td>
<td>160</td>
<td>10</td>
<td>-.02</td>
</tr>
<tr>
<td></td>
<td>Weighted Score Pretest</td>
<td>31.26</td>
<td>17.5</td>
<td>160</td>
<td>95.69</td>
<td>.60</td>
</tr>
<tr>
<td></td>
<td>Raw Score Posttest</td>
<td>4.74</td>
<td>2.2</td>
<td>160</td>
<td>10</td>
<td>.09</td>
</tr>
<tr>
<td></td>
<td>Weighted Score Posttest</td>
<td>40.53</td>
<td>23.6</td>
<td>160</td>
<td>95.69</td>
<td>.42</td>
</tr>
</tbody>
</table>

\textsuperscript{16} As mentioned earlier, in analytic Chapter 3 and Chapter 4 I focus on students’ item-level scores, so it is important to include these raw scores in my backdrop chapter.
The above table shows that native English speaking students have higher averages across all pretests, posttests, and gain scores. I now present a series of t-tests to determine whether the differences between these mean scores are statistically significant.

**Independent t-test on raw score pretest**

On average, native English speaking students have higher raw pretest scores \((M = 5.33, SE = .045)\) than emergent bilingual students \((M = 3.80, SE = .152)\). This difference is significant \(t(1958) = 9.677, p < .01\). It represents a medium effect size \(r = .214 (d = .437)\).\(^{17}\)

**Independent t-test on weighted score pretest**

On average, native English speaking students have higher weighted pretest scores \((M = 43.37, SE = .452)\) than emergent bilingual students \((M = 31.26, SE = .1.387)\). This difference is significant \(t(194) = 8.309, p < .01\). It represents a large effect size \(r = .512 (d = 1.193)\).

---

\(^{17}\) I use as a general baseline for relative effect sizes Cohen (1988)'s definition of “small, \(d = .2\),” “medium, \(d = .5\),” and “large, \(d = .8\).”
Independent t-test on raw score posttest

On average, native English speaking students have higher raw posttest scores (M = 6.37, SE = .045) than emergent bilingual students (M = 4.74, SE = .172). This difference is significant t(185) = 9.144, p < .01. It represents a large effect size r = 0.559 (d = 1.348).

Independent t-test on weighted score posttest

On average, native English speaking students have higher raw posttest scores (M = 58.47, SE = .507) than emergent bilingual students (M = 40.53, SE = 1.846). This difference is significant t(1961) = 10.112, p < .01. It represents a medium effect size r = .223 (d = .457).

Repeated measures ANOVAs to test for significance of gains (raw scores)

There is a significant main effect of time on students’ raw scores, F(1, 1861) = 128.536, p < .01. This represents a medium effect size, r = .255 (d = .527). In other words, students do significantly better on the posttests than on the pretests.

There is not a significant interaction effect between time and ESOL group status, F(1, 1861) = .054, p > .001. In other words, there are not significant differences between raw score gains across ESOL groups in the Year 2 Motion and Forces implementation, unlike the average across all four years of implementation.
Repeated measures ANOVAs to test for significance of gains (weighted scores)

There is a significant main effect of time on students' weighted scores, $F(1, 1861) = 158.678, p < .01$. This represents a medium effect size, $r = .281$ ($d = .586$). In other words, students do significantly better on the posttest than on the pretests.

There is also a significant interaction effect between time and ESOL group status, $F(1, 1861) = 5.913, p < .05$; however, it represents a much smaller effect size, $r = .056$ ($d = .110$). In other words, there are significant differences between weighted score gains across ESOL groups, but this only accounts for a very small amount of the variation in the sample.

In sum, the $t$-tests and repeated measures/mixed ANOVAs demonstrate statistically significant achievement gaps across Year 2 Motion and Forces implementation in the SCALE-uP corpus for most of the tests I conducted. The independent $t$-tests conducted on the raw scores and weighted scores pretest data demonstrate significant main effects, with medium and large effect sizes, respectively. The independent $t$-tests conducted on the raw scores and weighted scores posttest data demonstrate significant main effects, with large and medium effect sizes, respectively.

The lack of differences in the magnitude of effect sizes across the $t$-tests for pretest and posttests (as we saw across the four years of physics data) may be due to the fact that the Motion and Forces curriculum treatment was less effective during Year 2 than in the overall average across all years of implementation (see Rethinam et al., 2008 for further discussion of fidelity of implementation issues during Year 2);
therefore, efficacy of the treatment curriculum condition may not have been a factor in reducing the effect size of the significant differences in means for the posttest data (half of which was collected after the treatment curriculum, which was designed to reduce achievement gaps).

In summary of the ANOVA findings for the four years of physics data: the repeated measures ANOVAs reveal a significant main effect of time on students’ raw scores as well their weighted scores, both of which have a medium effect size. In other words, students do significantly better on the posttest than on the pretests. As for the achievement gaps in the ANOVA data, there is not a significant interaction effect between time and ESOL group status for the raw scores data. There is, however, a significant interaction effect between time and ESOL group status for the weighted scores data. Much like in the four year data set, though, this interaction represents a much smaller effect size than the main effects of time. In other words, there are significant differences between score gains across ESOL groups for the weighted scores data, but this only accounts for a very small amount of the variation in the sample.

2.3.5. SCALE-uP Achievement Gaps in the Eastmond Classroom

In this section, I present the descriptive statistics for the Eastmond classroom from the Motion and Forces implementation, as this is the classroom that I chose for the qualitative study in Chapter 4 of the dissertation. Again, I focus on just the emergent bilingual (ESOL 2) and native English speaking (ESOL 0) students and
consider both the raw and weighted scores in my analyses of students’ assessment data.

This section also discusses issues surrounding representativeness of the samples I have chosen for the analyses in this dissertation, in light of the national, state, and corpus-wide comparison data provided in previous sections of this backdrop Chapter 2.

The following Table 2.3 compares the average gains of emergent bilingual students and native English speaking students in the Eastmond classroom, as set against the backdrop of the Year 2 Motion and Forces implementation averages discussed above.

<table>
<thead>
<tr>
<th></th>
<th>Emergent Bilinguals</th>
<th>Native English Speaking Students</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comparison</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>27.9</td>
<td>45.1</td>
</tr>
<tr>
<td>Posttest</td>
<td>34.1</td>
<td>57.8</td>
</tr>
<tr>
<td><strong>Avg. Gains</strong></td>
<td><strong>6.2</strong></td>
<td><strong>12.7</strong></td>
</tr>
<tr>
<td><strong>Treatment</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>34.8</td>
<td>41.9</td>
</tr>
<tr>
<td>Posttest</td>
<td>45.9</td>
<td>59.0</td>
</tr>
<tr>
<td><strong>Avg. Gains</strong></td>
<td><strong>11.1</strong></td>
<td><strong>17.2</strong></td>
</tr>
<tr>
<td><strong>Eastmond</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>39.7</td>
<td>34.7</td>
</tr>
<tr>
<td>Posttest</td>
<td>43.8</td>
<td>52.1</td>
</tr>
<tr>
<td><strong>Avg. Gains</strong></td>
<td><strong>4.0</strong></td>
<td><strong>17.3</strong></td>
</tr>
</tbody>
</table>

The table shows that the achievement gaps in the Eastmond classroom are even more pronounced than across the larger SCALE-uP corpus (emergent bilingual students averaged only an approximately 4% gain while native English speaking students averaged approximately 17%). This makes the Eastmond classroom a
particularly interesting site for the study found in Chapter 4 of this dissertation, as it more closely resembles the national backdrop and the trends across the SCALE-uP corpus than many of the other classrooms in the Year 2 Motion and Forces implementation (i.e. it has more extreme achievement gaps than the average across Year 2).

2.4. CONCLUSION

In sum, ESOL group status is generally found to be a significant predictor of differences between achievement between emergent bilingual and native English speaking students. These are measured using various descriptive and parametric statistics throughout this chapter. While the effect sizes of these achievement gaps differ across cases, ESOL status seems to account for a relatively large portion of the variation in much of the SCALE-uP data.

These analyses serve as a backdrop and impetus for the subsequent, more targeted analyses of sources of achievement gaps in the SCALE-uP corpus by posing a general research question: What are the sources of these achievement gaps in the SCALE-uP corpus? To answer this, Chapter 3 presents an investigation of linguistic characteristics of assessment items as sources of achievement gaps through the use of computerized readability tools and multilevel modeling; Chapter 4 furthers the investigation of test-related contributions to achievement gaps by examining students’ constructed responses using a qualitative, interactional framework.
CHAPTER 3

LINGUISTIC CHARACTERISTICS OF ASSESSMENT ITEMS

An Investigation of Achievement Gaps Using Computerized Readability Tools and Multilevel Modeling

Achievement gaps between emergent bilingual students and their native English speaking peers present an important challenge for researchers, as explained in the introductory chapter (Chapter 1, Section 1.1). These national trends are mirrored in the achievement gaps found in my initial analysis of the SCALE-uP corpus that is reported in the preceding chapter (Chapter 2, Section 2.3–2.4). Those findings serve as a backdrop for the current study of linguistic characteristics of test items as potential sources of achievement gaps for emergent bilingual students.

For this analysis, I employ a computerized readability tool (Coh-Metrix) and an increasingly popular statistical procedure, Multilevel Modeling (MLM), to test the relationship between linguistic characteristics of test items and students’ item-level scores. The vast majority of the characteristics investigated in this study are related to the complexity of test item prompt language; however, some of the variables are not as transparently linguistic (e.g. selected versus constructed response type, although I would argue the salient distinction within this variable is still linguistic in that it relies on whether or not students are required to produce written language). With this caveat, I choose to refer to the set of variables in this study as linguistic characteristics of test items rather than a more general term such as item characteristics that might more clearly encompass formatting variables, etc.
I also use MLM to investigate whether the relationship between linguistic characteristics of test items and students' item-level scores differs across emergent bilingual and native English speaking student groups. My results indicate that some linguistic characteristics of test items are predictive of students' item-level scores, and these relationships vary across emergent bilingual and native English speaking students for certain format, syntax, and discourse variables. This provides some evidence that native English speaking content tests conflate the assessment of English language proficiency and content knowledge, thereby exacerbating achievement gaps for emergent bilingual students. However, the lack of significant results for many of the variables tested in this analysis also points to the fact that much of the variation in students’ item-level scores is not explained by linguistic aspects of the test item prompts.

In the following introduction to the chapter (Section 3.1), I outline the general problem and previous studies related to linguistic characteristics of test items and readability that have motivated my research questions and hypotheses. The next section of this chapter (Section 3.2) provides details about the methodology employed, including the data sample and analysis procedures. Results of the MLM for each research question are presented and briefly discussed in the findings section (Section 3.3). Finally, the conclusion (Section 3.4) provides summary along with discussion, implications, limitations, and suggestions for future research.
3.1. INTRODUCTION

3.1.1. Problem

Researchers have argued that the measurement of achievement itself—native English speaking content testing—may contribute to achievement gaps because it conflates students’ language proficiency and content knowledge (e.g. Abedi, 2006; Farnsworth, 2006; Martiniello, 2008; Pollitt, Marriot, & Ahmed, 2000; Staley, 2005; Wolf & Leon, 2009). In other words, if a student is asked to demonstrate content knowledge in a second language (L2), then her/his L2 proficiency may serve as a source of construct-irrelevant difficulty in that content assessment (Messick, 1989). Indeed, the validity of native English speaking content tests has long been questioned in national reports because language barriers may prevent emergent bilingual students from demonstrating the full range of their content knowledge (e.g. August & Hakuta, 1997; National Research Council, 2002).

Reflecting an assumption that these language barriers increase along with increased difficulty in test language, recent research has focused on the linguistic characteristics of test items as potential sources of bias against emergent bilingual students. Much of this research has found that longer and more linguistically complex items tend to favor native English speaking students. A prominent example of this can be found in the work of Jamal Abedi and his colleagues, who have conducted a series of pioneering studies of linguistic characteristics of test items across content subjects and student age levels. This research has provided strong
evidence that linguistic complexity of content test items negatively affects the performance of emergent bilingual students (Abedi, 2004; Abedi, 2006; Abedi, Courtney, & Leon, 2001; Abedi, Courtney, Mirocha, & Goldberg 2005; Abedi, Leon, & Mirocha, 2003; Abedi & Lord, 2001; Abedi, Lord, & Plummer, 1997).

Other researchers have extended this line of inquiry, providing additional support for the argument that linguistic characteristics of test items are related to differences between content test performance between emergent bilingual and native English speaking students (e.g. Farnsworth, 2006; Martiniello, 2008; Pollitt, Marriot, & Ahmed, 2000; Staley, 2005; Wolf & Leon, 2009). The vast majority of these studies have focused on math and/or science tests for elementary and/or middle school students, with consistent findings of a relationship between emergent bilingual students’ performance and the general amount of language used in test items (Martiniello, 2008; Staley, 2005; Wolf & Leon, 2009), usage of particular English forms in item prompts (Farnsworth, 2006; Martiniello, 2008; Pollitt, Marriot, & Ahmed, 2000; Wolf & Leon, 2009), and the inclusion of visual representations and graphics (Martiniello, 2008; Logan-Terry & Wright, 2010; Wolf & Leon, 2009).

In contrast, some studies of linguistic characteristics of test items and student performance have had limited and/or mixed results. For example, Bailey (2005), a comprehensive study of math and science items for middle school students, found minimal correlations between measures of linguistic complexity and differential student outcomes between emergent bilingual and native English speaking students. Out of many measures of linguistic complexity employed by
Bailey (ibid; details discussed in the next section of this introduction), only discourse-level demands emerged as a salient predictor of student outcome differences, and this relationship was only found in the context of math items. Similarly, Cid (2009), a study of linguistic complexity and high-stakes reading comprehension tests for high school students, provided only very limited support for the explanatory properties of the linguistic features, despite investigating a range of item characteristic variables.

The issue of linguistic complexity as a source of achievement gaps for emergent bilingual students has also been addressed in literature related to accommodations, or changes to tests that attempt to address the needs of ESOL students without altering the construct being measured (Rivera & Collum, 2006). Reflecting an assumption that linguistic complexity is a source of difficulty for emergent bilingual students, linguistic simplification has become one of the most popular test accommodations for ESOL students (Kopriva, 2008). However, large-scale studies of linguistic simplification have produced only limited findings of the relative success of the accommodation (e.g. Hudicourt-Barnes, Noble, Roseberry, Suarez & Warren, 2008; Wolf, Herman, Bachman, Bailey, & Griffin, 2008).

In addition, a recent meta-analysis of test accommodations research conducted by Kieffer, Lesaux, Rivera, & Francis (2009) found linguistic simplification to be completely ineffective in reducing achievement gaps. This brings into question the findings discussed above about the relationship between linguistic complexity and achievement gaps (e.g. the prominent work of Abedi and colleagues). Indeed, of the seven total test accommodations studied by Kieffer et al.
(ibid), only the provision of English dictionaries or glossaries was found to have a significant effect on achievement gaps, and the effect size for the provision-of-English-dictionaries/glossaries variable was very small.

Other recent studies have also yielded results that directly contradict claims of a relationship between linguistic complexity and differences in student performance between emergent bilingual and native English speaking student groups. For instance, Farnsworth (2008) investigated Academic English Language (AEL) as a factor influencing performance on standardized elementary and middle school math and science tests, finding that AEL did not explain differences in performance. Due to these recent contradictory and compelling results, Farnsworth (ibid) and other researchers (e.g. Abedi & Gándara, 2006; Abedi & Herman, 2010; Kieffer et al., 2009; Koenig & Bachman, 2004; Elliot, 2008; Ockey, 2007) have argued that linguistic aspects of test items may actually be construct-relevant (i.e. that there is a necessary role for academic language skills in content assessments), and a much larger proportion of differences in performance between groups may be due to factors other than linguistic characteristics of test items (e.g. underlying differences in knowledge that are linked to differences in socioeconomic status and opportunities to learn in the classroom for emergent bilingual and native English speaking students).

In sum, while some research in this area provides evidence that differences in performance between emergent bilingual and native English speaking students are related to linguistic characteristics of test items, there is much recent debate in the field as to whether this holds true. This debate casts doubt on the argument that
language serves as a source of construct-irrelevant variance in monolingual content tests for emergent bilingual students, thereby contributing to achievement gaps. The present study aims to shed new light on this issue by investigating linguistic characteristics of test items and students' item-level scores in the SCALE-uP corpus described in Chapter 1, Section 1.2. I do this by incorporating readability literature in the operationalization of linguistic characteristics of tests items, as described in the next section.

3.1.2. Operationalizing Linguistic Characteristics of Test Items

The operationalization of linguistic characteristics of test items may account for some of the discrepancies in the findings of previous researchers, as this is one of the most important and challenging aspects of conducting item characteristic research. Each study has a distinct way of describing linguistic characteristics of test items, and even the terminology used across studies is inconsistent. For example, Abedi (2006) and Martiniello (2008) utilize the term “linguistic complexity” while other researchers point to “language demands” (Wolf & Leon, 2009) or “Academic English Language” (Farnsworth, 2006, 2008). Still others opt for more general wording, as with Cid (2009)’s use of “linguistic features” when describing item characteristics that may affect differential performance between emergent bilingual and native English speaking students. These terminology differences aside, there are some general patterns across previous studies that prove useful in operationalizing linguistic characteristics of test items for the current analysis. The following Table
3.1 provides an overview of some of the operationalizations of linguistic characteristics of test items from previous studies.

Table 3.1. Overview of Linguistic Characteristics of Test Items in Previous Studies

<table>
<thead>
<tr>
<th>Linguistic Area of Measures</th>
<th>Example Measures of Linguistic Characteristics of Test Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
<td>Length, selected vs. constructed response, visuals/layout</td>
</tr>
<tr>
<td>Syntax</td>
<td>Passive voice, multiple/complex clause structure, conditionals, negation, modals, relative pronouns, fragments, modifiers</td>
</tr>
<tr>
<td>Lexis</td>
<td>Length of words; frequency/commonality of words; cultural references/idiomatic expressions; abstract, ambiguous, polysemous, and/or false cognate vocabulary; academic vocabulary</td>
</tr>
<tr>
<td>Discourse</td>
<td>Synthesis across sentences, lexical referents, substitution</td>
</tr>
</tbody>
</table>

As outlined in Table 3.1, previous researchers have generally divided linguistic characteristics of test items into categories related to item format, syntax, lexis, and discourse. In most studies, researcher have given holistic scores based on their own and trained raters’ assessments of these characteristics rather than taking direct measurements of the characteristics for each item (e.g. Bailey, 2005; Farnsworth, 2008; Martiniello, 2008). However, researchers have long pointed out the difficulty of such coding, encouraging the use of more precise measurements when possible (e.g. Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

Some of the most basic item characteristics in past studies have been item length, format, and layout. Researchers often take into account the number of words and/or the length of sentences in item prompts (e.g. Abedi, 2006; Martiniello, 2008; Wolf & Leon, 2009). In some cases, it has been appropriate to categorize items as
selected or constructed response, although this is less applicable in large-scale standardized testing that is primarily multiple choice in nature (Martiniello, 2008). Researchers have also recently analyzed the presence and type of visual elements in the item, arguing that they may aid students' comprehension of item prompts (e.g. Martiniello, 2008; Wolf & Leon, 2009).

Syntactic complexity and lexical complexity are also prominent in past studies, and the operationalization of these categories varies quite a bit. In assessing syntactic complexity, for instance, researchers have investigated the use of: passive voice (Abedi & Lord, 2001; Farnsworth, 2008; Martiniello, 2008; Wolf & Leon, 2009); multiple/complex clause structures (Abedi & Lord, 2001; Bailey, 2005; Cid, 2009; Farnsworth, 2008; Martiniello, 2008; Wolf & Leon, 2009); negation (Abedi, 2006; Cid, 2009; Martiniello, 2008); modals and relative pronouns (Farnsworth, 2008; Wolf & Leon, 2009); fragments (Bailey, 2005); and modifiers (Martiniello, 2008).

Likewise, lexical complexity measures have some commonalities and some differences across studies. Items are often rated according to: average length of words (Abedi, 2006); average frequency of lexical items (Abedi, 2006; August, Carlo, Dressler, & Snow, 2005; Bailey, 2005; Bailey, Huang, Shin, Farnsworth & Butler, 2007; Martiniello, 2008); cultural references and/or idiomatic expressions (Bernhardt, 2005; Martiniello, 2008), and abstract, ambiguous, polysemous, and/or false cognate vocabulary (Cid, 2009; Martiniello, 2008). As mentioned in the previous section, these studies of syntactic and lexical complexity have produced
mixed results with regard to whether language disadvantages emergent bilingual students.

The treatment of academic terminology is a particularly important aspect of item lexical complexity. As discussed in the introductory chapter (Chapter 1, Section 1.3), academic language has been defined in various ways throughout both assessment and classroom literature. Many studies of linguistic characteristics of test items and student outcomes have simply characterized the quantity of academic language (i.e. school-specific, not used in everyday life) in a holistic rating of items, regardless of whether the academic language is general or discipline-specific (e.g. Bailey, 2005; Farnsworth, 2008). Other studies have distinguished between general academic vocabulary and context-specific and/or technical vocabulary (e.g. Martiniello, 2008; Wolf & Leon, 2009), arguing that the latter is construct-relevant in content testing. These differences between operationalization of academic language may explain some of the differences between findings, as those studies that included discipline-specific vocabulary in the estimation of academic language were also those with the least support of an academic language–outcome relationship (Bailey, 2005; Farnsworth, 2008).

A final category found in many of the item characteristics studies is that of discourse-level complexity. This category has had mixed results—both confirming and refuting language as a source of construct-irrelevant variance—in past literature, and has been defined in various ways including: synthesis across sentences (Bailey, 2005), inclusion of lexical referents (Farnsworth, 2008; Wolf & Leon, 2009), and substitution (Wolf & Leon, 2009). Operationalizing this discourse
category of item characteristics is particularly problematic and can benefit greatly from incorporating insights from readability literature, as discussed in the next section.

3.1.3. Incorporating Readability Literature

Many of the linguistic characteristics of test items found in the testing literature reflect an assumption that increased linguistic complexity leads to increased difficulty in comprehension. This assumption has roots in traditional readability literature. For instance, the Flesch readability scales—developed in 1948 and revised in 1975—continue to be used throughout education research and practice. The Flesch scales emphasize sentence length and lexical complexity (e.g. number of syllables and the frequency of lexical items) as sources of reading difficulty, and traditional L2 readability scales generally follow the same pattern (Greenfield, 1999). The assumption that complexity leads to difficulty is also apparent in popular ways of describing texts found in strands of second language acquisition (SLA) and bilingualism research, including analysis of mean length of utterance (MLU), the ratio of clauses to T-units, total number of grammatical verb forms, modality, voice, type-token ratio/D, etc. (e.g. Ellis, 2004; King & Logan-Terry, 2008).

However, readability researchers have pointed out various problems with assuming that increased linguistic complexity leads to increased difficulty in comprehension (Davison & Green, 1988; Davison & Kantor, 1982; Koda, 2005;
McNamara, Kintsch, Butler-Songer, & Kintsch, 1996), particularly for emergent bilinguals (Crossley, Greenfield, & McNamara, 2008). These readability studies have argued that certain aspects of linguistic complexity may be necessary for comprehension for both native English speaking and emergent bilingual students. For example, traditional readability formulas may penalize a text for containing longer words with more syllables; however, morphologically rich lexical items have been shown to enhance readability in some cases (e.g. Randall, 1988). Other elements that may increase item complexity yet may also be important components of discourse cohesion and comprehension include: connectives (Halliday & Hasan, 1976; Kemper, 1988); repetition (Brown, 1998; Halliday & Hasan, 1976); context-specific details (Charrow, 1988); concreteness (Gibbs, 2002), and explicit information connecting events (Kemper, 1988). These elements are also important components of oral discourse, and researchers have long argued that the relationship between oral and written discourse is underestimated (e.g. Chafe, 1982; Tannen, 1989, 2007). Together these findings indicate that item characteristics studies should not assume linguistic complexity leads to difficulty for both emergent bilingual and native English speaking student groups.

Insights from the readability and oral discourse literature also highlight why visual elements may be especially salient in the testing context, as discussed in some of the item characteristics studies (e.g. Martiniello, 2008; Wolf & Leon, 2009). Researchers have shown the importance of visual details accompanying written texts (Rumelhart, 1994). From the Interactional Sociolinguistic perspective that is explained more thoroughly and utilized in Chapter 4 (see also Gumperz, 1982, 1999;
Hamilton, 1994; Schiffrin, 1994; Tannen, 2004), speakers use contextualization cues (e.g. pitch, tone, gestures, body posture, etc.) to signal how interaction is meant to be interpreted in spoken discourse. Upon reading an item prompt, test takers interpret what the test makers are asking; however, there are few contextualization cues to help signal meaning interpretation in this written context. Therefore, visual accompaniments to item prompts may be able to provide contextualization cues that activate students’ knowledge schemas and may thus be important in determining how both emergent bilingual and native English speaking test takers’ schemas are activated (see also Logan-Terry & Wright, 2010 as well as Wright & Logan-Terry, in preparation for further discussion of this). This idea is also further explored in the subsequent qualitative Chapter 4.

In addition to oral discourse literature, modern readability research also takes into account psycholinguistic and cognitive models of language processing. For example, Koda (2005) proposes that reading operates at lexical, syntactic, semantic, and discoursal levels of processing. Others have argued that reading processing occurs at three levels: decoding, syntactic parsing, and meaning construction (Just & Carpenter, 1987; Rayner & Pollatsek, 1994). Across these reading theories, automaticity remains an important component, often being linked to the usage and frequency of vocabulary in both spoken and written discourse. This emphasis on automaticity and frequency is also reflected in recent work in the field of SLA (e.g. Ellis & Collins, 2009; Ellis & Larsen-Freeman, 2009), and the popularity of corpus-based approaches to language (e.g. Biber, 1986). These insights from the
readability literature are incorporated in the computerized Coh-Metrix approach that I employ in the present study, outlined in the following section.

3.1.3.1. Coh-Metrix Measures

Over the past decade, readability researchers have worked to incorporate these language processing models in the development of an online computational tool, Coh-Metrix, to help determine text readability (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, Louwerse, & Graesser, 2002). This tool measures cohesion and text difficulty at the various levels outlined by psycholinguistic and cognitive models of processing. Coh-Metrix also draws on insights from work in the areas of computational linguistics, corpus linguistics, information extraction, information retrieval, and discourse processing, including the use of semantic lexicons, pattern classifiers, part of speech taggers, syntactic parsers, and semantic interpreters. This tool has been shown to be very useful in assessing text readability for native English speaking students (Graesser et al., 2004; McNamara et al., 2002).

Crossley, Greenfield, and McNamara (2008) also tested Coh-Metrix output as an improved means of measuring English text readability for emergent bilingual students. In particular, the researchers investigated three variables computed by Coh-Metrix: lexical frequency (calculated using the CELEX database; Baayen, Piepenbrock, & Gulikers, 1993), syntactic similarity (the uniformity and consistency of parallel syntactic constructions), and content word overlap (how often content words overlap between two adjacent sentences). Crossley et al. (ibid) argue these
variables have greater construct validity because they broadly correspond to the
three levels of processing discussed above—decoding, syntactic parsing, and
meaning construction. The researchers’ hypotheses that higher scores for these
variables would predict higher readability for emergent bilinguals were confirmed
in the study. This indicates that these complexity measures may also be valuable for
studies of the relationship between linguistic characteristics of test items and the
performance of emergent bilingual and native English speaking students.

In addition to the three variables investigated in the Crossley et al. (2008)
study, Coh-Metrix also calculates other measures of cohesion that may be useful,
such as the situational dimensions of causation (especially important for science
tests that assess knowledge of causation), intentionality, temporality, and spatiality.
Coh-Metrix also offers output related to Latent Semantic Analysis (an alternative to
the content word overlap variable for discourse-level variables), logical operators
(Boolean and some conditionals), lexical concreteness (based on the MRC
psycholinguistics database; Coltheart, 1981), hypernyms (from WordNet), and other
syntactic complexity measures (e.g. mean number of modifiers per noun phrase,
mean number of higher level constituents per sentences, number of words
appearing before main verb, and pronominal reference). As other Coh-Metrix
output variables have been found to be important predictors of readability for
emergent bilingual students (Crossley et al., ibid), item characteristic studies would
benefit from incorporating these other Coh-Metrix measures, as well. However, it is
also important to note that Coh-Metrix was designed for longer texts than those

18 The Coh-Metrix tool and additional rationale for each of the readability measures
are available at http://cohmetrix.memphis.edu/cohmetrixpr/index.html
found in most test item prompts; thus, it may be necessary to modify computerized readability tools to suit item characteristics studies. This limitation of the usage of Coh-Metrix is also discussed in Section 3.4.3 of this chapter.

3.1.4. Summary of Findings from Past Literature

In sum, researchers have operationalized linguistic characteristics of test items in various ways, often including categories for format, syntax, lexis, and discourse. These operationalizations reflect an assumption that linguistic complexity in these areas leads to decreased comprehension; however, insights from readability literature demonstrate the importance of considering that some linguistic characteristics of test items may be related to increased comprehensibility. Coh-Metrix, an online tool for assessing text readability and cohesion, takes into account linguistic features that may make texts more difficult as well as linguistic features that may make texts more comprehensible. Thus, Coh-Metrix has potential utility in item characteristic studies as it brings together insights from diverse strands of research and has been shown to predict readability for both emergent bilingual and native English speaking students.

Figure 3.1 provides an overview of findings from the item characteristic literature, as well as the readability literature. It summarizes the features that have been argued to make text more difficult (left-hand column) and the features that have been argued to make text more comprehensible (right-hand column).
Figure 3.1. Overview of Characteristics Contributing to Text Difficulty/Comprehensibility

Difficult

- Length (number of words, number of sentences)
- Syntactic complexity (sentence length, modifiers per noun phrase, words appearing before the main verb, logical operators)
- Lexical complexity (low frequency lexical items, abstractness/hypernymy, low concreteness, academic language)

Comprehensible

- Discourse/cohesion (connectives, syntactic and lexical repetition, coreference, anaphora, situational dimensions of causation, intention, temporal, spatial)
- Visual elements (pictures, graphics, layout)
3.1.5. Research Questions and Hypotheses

Given the debate over linguistic characteristics of test items and the advances in readability literature, I ask the following research questions:

1) Are linguistic characteristics of test items related to students’ item-level scores?
2) Are there differences between students’ item-level scores across emergent bilingual and native English speaking groups of students?
3) Does the relationship between linguistic characteristics of test items and students’ item-level scores vary across emergent bilingual and native English speaking students?

For research question one, I hypothesize that some linguistic characteristics of test items will have a negative relationship with students’ correct responses (those features listed in the left-hand column in Figure 3.1 at the end of Section 3.1.3. above), while some characteristics will have a positive relationship with students’ correct responses (those features listed in the right-hand column in Figure 3.1 at the end of Section 3.1.3. above).

Based on national trends and the achievement gaps found in my previous analysis, for research question two I hypothesize that students’ item-level scores will vary significantly across emergent bilingual and native English speaking groups. The findings for this second research question primarily serve as a precondition for investigating the third research question.
As for the third research question, dealing with how the item characteristic–student response relationship may vary across emergent bilingual and native English speaking students, past research results are less clear. Some of the item characteristic literature suggest a possible significant interaction between measures of lexical and syntactic complexity and membership in emergent bilingual or native English speaking groups (e.g. Abedi, 2006; Martiniello, 2008; Wolf & Leon, 2009). In contrast, other item characteristic research might indicate that there will not be significant interactions between complexity measures and student language group membership (e.g. Bailey, 2005; Cid, 2009; Farnsworth, 2008). Readability studies with emergent bilingual students also weigh in on this debate, indicating that the relationship between certain linguistic characteristics of test items calculated by the Coh-Metrix tool and students’ outcomes might be more extreme for emergent bilingual students (e.g. lexical frequency, syntactic similarity, and content word overlap; Crossley et al., 2008).

In sum, I hypothesize that there will be both negative and positive relationships between linguistic characteristics of test items and students’ item-level scores, and the slope of these relationships will be more extreme in the case of emergent bilingual students for some item characteristic variables (e.g. item length, syntactic similarity, lexical frequency, and content word overlap).
3.2. METHOD

3.2.1. Sample

A subset of the SCALE-uP corpus—discussed in Chapter 1, Section 1.2—was selected for the analysis of linguistic characteristics of test items and students’ item-level scores. I selected the sixth grade physics assessments, Questions About Motion (QAM) and the responses of 864 students for this analysis, as the physics assessment and classroom data were the most consistent across the SCALE-uP corpus.\(^{19}\) As mentioned in the Introduction, the QAM test was created by the research team as a means of measuring students’ learning over the course of the Motion and Forces curriculum unit; however, I chose to only analyze students’ pretest responses to simplify analysis and interpretation for the current chapter. The research questions for this chapter relate to the influence of the tests themselves rather than the effects of the curriculum units, and I argue that these pretests more accurately reflect the influence of linguistic characteristics of test items on students’ item-level scores, as opposed to differences between learning during the curriculum unit.\(^{20}\) Because I was only analyzing pretests, it was appropriate to look at students in both the experimental and control conditions of

\(^{19}\) Consistency was determined by the degree to which data were available for all students in the assessment and ethnographic data sets from the original SCALE-uP corpus.

\(^{20}\) It would be interesting to see the effects of the linguistic characteristics of test items in the posttest data, but would require a separate analysis that accounts for differences across curriculum treatment groups. This sort of analysis is outside the scope of the current analytic chapter.
the original SCALE-uP study. The treatment—implementation of the curriculum unit—had not yet been administered at the time of the pretest, and both the treatment and control groups consisted of a random sample of students.

Another decision I made in selecting cases for this analysis was to focus on only those students currently receiving English as a Second Language (ESOL) services and those students who had never received ESOL services. Those students currently receiving ESOL services represent emergent bilingual students in this study, and those students never having received ESOL services represent native English speaking students.21 A third category of students, those who had previously received ESOL services, was excluded from the present analysis. Although this group of prior ESOL students is important, they represent a very heterogeneous group with some students having recently exited ESOL services along with others having been exited from ESOL services for half a decade. Therefore, I chose to examine linguistic characteristics of test items and student outcomes across the two ESOL groups that most clearly represent differences between emergent bilingual and native English speaking students.

After deleting those students with missing pretest and/or ESOL information, there were 432 emergent bilingual students—those currently receiving ESOL services—and 6,215 native English speaking students in the sample—those never having received ESOL services. I decided to do a random listwise deletion of data

21 The limitations of this operationalization are also discussed throughout other chapters of this dissertation, particularly the Introduction Chapter 1; however, this categorization is the only language status grouping variable available in the SCALE-uP corpus and is thus the most reliable measure to use in the present study.
from the native English speaking student group so that group sizes would be equal.\textsuperscript{22} This is how I arrived at the 864 participants in the current study.

As discussed in the Introduction Chapter 1, the QAM test instrument features ten items, and can be found in Appendix A. I have chosen to use the binary distinction between correct and incorrect answers on these items as the student outcome variable in this study—the outcome variable. Reports about the development and implementation of QAM from the SCALE-uP team indicate that all the codes given to constructed responses were eventually coded as correct or incorrect and used as such in calculating students’ overall scores on the tests. Thus, in calculating their item-level scores, I used this same binary distinction. While this decision loses some information about students’ ratings, it serves to simplify the interpretation of the MLM analysis and is more ecologically-valid in keeping with the original design of the test. As such, analysis of the specific coding instructions given to the coders is outside the scope of this analytic chapter, which analyzes the linguistic characteristics of the test item prompts as they relate to students’ item-level scores.

The QAM assessment was developed based on the Advancement of Science (AAAS) Project 2061 guidelines.\textsuperscript{23} Figure 3.2 also shows the four target ideas and their corresponding QAM items. I treated the predictor variable of target idea as

\textsuperscript{22} Randomly deleting native English speaking students’ data simplified data management and sped up some of the statistical analyses in SPSS without hurting the power of the analysis (which is tied to the smallest group size). The statistical procedure used in this analysis is robust to differences in group size, so this methodological was a decision driven by convenience rather than necessity.

\textsuperscript{23} Additional information about AAAS Project 2061 can be found at http://www.project2061.org/
ordinal rather than categorical because the target ideas are ordered in increasing difficulty and build upon one another, as seen in Figure 3.2. The four target ideas in the QAM assessment are not presented successively in the QAM assessment items (e.g. QAM Item #1 deals with Target Idea #2, which builds upon Target Idea #1 that is subsequently tested in QAM Items #4, 5, 6, and 10).

Figure 3.2. Target Ideas and Corresponding QAM Items

1) Changes in speed or direction are caused by forces.
   • target idea for QAM items 4, 5, 6, and 10

2) The greater the force is, the greater the change in motion will be.
   • target idea for QAM items 1, 2, and 7

3) The more massive an object is, the less effect a given force will have.
   • target idea for QAM item 3

4) An object at rest stays that way unless acted on by a force and an object in motion will continue to move unabated unless acted on by a force.
   • target idea for QAM items 8 and 9

24 The four target ideas in the QAM assessment are not presented successively in the QAM assessment items (e.g. QAM Item #1 deals with Target Idea #2, which builds upon Target Idea #1 that is subsequently tested in QAM Items #4, 5, 6, and 10); however, since their underlying content is ordinal in nature I treat them as such in this analysis. There are limitations to this, as students are not exposed to the target ideas in that same order as they interact with the assessment items, but this is the way that the test developers ordered the target concepts; thus, I treat them as such in the present analysis.
3.2.2. Analysis

In this section, I outline the measurement of the linguistic characteristics of test items (Section 3.2.2.1.) and then the multilevel modeling statistical approach taken in the analysis (Section 3.2.2.2.).

3.2.2.1. Measurement of Linguistic Characteristics of Test Items (Predictor Variables)

I measure 20 initial item characteristics—the predictor variables—based on the testing and readability research presented in the introductory section. Table 3.2 shows an overview of the 20 initial item characteristics and their basic definitions. In this table and throughout the rest of the chapter, the names of those variables that are hypothesized to have a negative relationship with students’ item-level scores are marked with a minus ( - ) sign while the names of those variables that are hypothesized to have a positive relationship with students’ scores are marked with a plus ( + ) sign.
<table>
<thead>
<tr>
<th>Linguistic Area of Predictor Variable</th>
<th>Linguistic Characteristic (Predictor Variable)</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
<td><strong>Target Idea</strong></td>
<td>One of four target ideas that each of the test items addresses</td>
</tr>
<tr>
<td></td>
<td><strong>Selected or Constructed</strong></td>
<td>Whether the item is selected among three multiple choice items or requires a constructed response from a student</td>
</tr>
<tr>
<td></td>
<td><strong>Item Length (# words)</strong></td>
<td>Number of words in the item prompt</td>
</tr>
<tr>
<td></td>
<td><strong>Item Length (# sentences)</strong></td>
<td>Number of sentences in the item prompt</td>
</tr>
<tr>
<td></td>
<td><strong>Distance Text Stem to Prompt</strong></td>
<td>Amount of distance (measured in terms of other questions, pictures, etc.) between the original set up of an item and the eventual item question</td>
</tr>
<tr>
<td></td>
<td><strong>Distance Picture to Question</strong></td>
<td>Amount of distance (measured in terms of other questions) between the picture association with an item and the eventual item question</td>
</tr>
<tr>
<td><strong>Syntax</strong></td>
<td><strong>Sentence Length</strong></td>
<td>Mean number of words per sentence in the item prompt</td>
</tr>
<tr>
<td></td>
<td><strong>Modifiers per Noun Phrase</strong></td>
<td>Mean number of modifiers per noun-phrase (optional adjectives, adverbs, or determiners that modify the head noun) in the item prompt</td>
</tr>
<tr>
<td></td>
<td><strong>Words before Main Verb</strong></td>
<td>Mean number of words before the main verb of the main clause in sentences in the item prompt—argued to be taxing on working memory</td>
</tr>
<tr>
<td></td>
<td><strong>Logical Operators</strong></td>
<td>Incidence of logical operators including Boolean operators and, or, not, if, then” — argued to contribute to text difficulty</td>
</tr>
<tr>
<td><strong>Lexis</strong></td>
<td><strong>Syllables per Word</strong></td>
<td>Mean number of syllables per content word, a ratio measure</td>
</tr>
<tr>
<td></td>
<td><strong>Frequency of Lexical Items</strong></td>
<td>Mean raw frequency of all of the content words (nouns, adverbs, adjectives, main verbs, and other categories with rich conceptual content) in the item prompt</td>
</tr>
<tr>
<td></td>
<td><strong>Noun Hypernymy</strong></td>
<td>Mean hypernym (multiplicity of meaning) value of nouns in the item prompt—argued to contribute to text difficulty</td>
</tr>
<tr>
<td></td>
<td><strong>Verb Hypernymy</strong></td>
<td>Mean hypernym (multiplicity of meaning) value of main verbs in the item prompt—argued to contribute to text difficulty</td>
</tr>
<tr>
<td></td>
<td><strong>Concreteness</strong></td>
<td>Mean concreteness value of all content words in the item prompt that match a word in the MRC database</td>
</tr>
<tr>
<td><strong>Discourse</strong></td>
<td><strong>Connectives</strong></td>
<td>Incidence of all connectives in the item prompt (e.g. positive: and, after, because; negative: but, until, although; additives: also, moreover, however; but; causal: because, so, consequently, although, nevertheless; logical: or, actually, if; temporal: after, before, when, until)</td>
</tr>
<tr>
<td></td>
<td><strong>Syntactic Repetition</strong></td>
<td>Proportion of intersection tree nodes between all adjacent sentences (algorithms build an intersection tree between two syntactic trees, one for each of the two sentences being compared)—a measure of cohesion</td>
</tr>
</tbody>
</table>
|                                      | **Content Word**                              | The proportion of content words in adjacent sentences that
<table>
<thead>
<tr>
<th>Repetition</th>
<th>share common content words in the item prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Causal Cohesion</em></td>
<td>Ratio of causal particles to causal verbs in the item prompts—cohesion suffers when the text has many causal verbs (signifying events and actions) but few causal particles that signal how the events and actions are connected</td>
</tr>
<tr>
<td><em>Intentional Cohesion</em></td>
<td>Ratio of intentional particles (e.g., <em>in order to, so that, for the purpose of, by means of, by, wanted to</em>) to the incidence of causal content (intentional actions/events)</td>
</tr>
</tbody>
</table>

Some of these Coh-Metrix measures are more transparent than others. As an example of one of the more complex Coh-Metrix measurements, *Concreteness* is measured using the MRC Psycholinguistics Database (Coltheart, 1981). This database gives scalar values to words on the basis of particular characteristics. As Graesser et al. (2004) explain: “The MRC Psycholinguistics Database contains 150,837 words and provides information of up to 26 different linguistic properties of these words. Most MRC indices are based on psycholinguistic experiments conducted by different researchers, so the coverage of words differs among the indices. Coh-Metrix 2.0 uses the MRC concreteness ratings for a large sample of content words. Concreteness measures how concrete a word is, based on human ratings. High numbers lean toward concrete and low numbers to abstract. Values vary between 100 and 700.”

Similarly, *Connectives* is a Coh-Metrix variable that features complex calculations. It is based on theoretical assumptions that connectives are one important class of signaling device for cohesion relations in text (Halliday & Hasan, 1976). Based on these assumptions, Graesser et al. (2004) make the following classifications for connectives in Coh-Metrix: “Connectives are classified on two dimensions in Coh-Metrix 2.0. On one dimension, the extension of the situation
described by the text is determined. Positive connectives extend events, whereas negative connectives cease to extend the expected events (Sanders, Spooren, & Noordman, 1992). Negative relations are synonymous with adversative relations, as defined in Halliday and Hasan (1976)."

Appendix D gives an overview of the item scores across these 20 variables. I use the online computational Coh-Metrix tool to calculate the majority of these measurements. As mentioned in the introduction to this chapter (Section 3.1.3), this tool draws on computational linguistics, corpus linguistics, information extraction, information retrieval, and discourse processing, including the use of semantic lexicons, pattern classifiers, part of speech taggers, syntactic parsers, and semantic interpreters in order to measure text cohesion and text difficulty.

As also mentioned in the introduction to this chapter, some of the most important measurements provided by Coh-Metrix for this study of emergent bilinguals include: lexical frequency (calculated using the CELEX database; Baayen, Piepenbrock, & Gulikers, 1993), syntactic similarity (the uniformity and consistency of parallel syntactic constructions), and content word overlap (how often content words overlap between two adjacent sentences) because they have been shown to be predictive of reading comprehension for emergent bilinguals (Crossley et al., 2008). Some other examples of key Coh-Metrix measurements include a number of syntactic indices that measure syntactic complexity. Coh-Metrix uses the Charniak syntactic parser for syntactic analyses, in which over 50 parts of speech are segregated into content and function words. Then, Coh-Metrix computes the number of noun-phrase (NP) constituents or number of verb-phrase (VP)
constituents per 1000 words. For my study, I rely on Coh-Mterix’s outputs with regard to the mean number of modifiers per noun-phrase. As explained in the Coh-Metrix manual (Graesser et al., 2002),25 “... a modifier is an optional element that describes the property of a head of a phrase. Modifiers per NP refer to adjectives, adverbs, or determiners that modify the head noun. For example, the noun-phrase the lovely, little girl has three modifiers: the, lovely and little.” Another Coh-Metrix measurement that I use in this study is the number of words that appear before the main verb of the main clause in the sentences of a text.

As for the lexical measurements available in Coh-Metrix, I rely most heavily on the frequencies generated using the Celex database, including the Log frequency of content words. According to Coh-Metrix, “Content words are nouns, adverbs, adjectives, main verbs, and other categories with rich conceptual content. Taking the log of the frequencies rather than the raw scores is compatible with research on reading time (Haberlandt & Graesser, 1985; Just & Carpenter, 1980).” Another useful Coh-Metrix lexical measure is that of verb hypernymy, which is one way of measuring the abstractness of the words using WordNet (Fellbaum, 1998). As described on the Coh-Metrix website:

WordNet is an online lexical reference system, the design of which is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized

into semantic fields of underlying lexical concepts. Some sets of words are functionally synonymous because they have the same or a very similar meaning. There are also relations between synonym sets. In particular, a hypernym metric is the number of levels in a conceptual taxonomic hierarchy above (super ordinate to) a word. For example, chair (as a seat) has 7 hypernym levels: seat -> furniture -> furnishings -> instrumentality -> artifact -> object -> entity.

Finally, I use Coh-Metrix to calculate some discourse-level measures, including sentence syntactic similarity and content word overlap. As explained by Coh-Metrix and in the introduction above, “The sentence syntax similarity indices compare the syntactic tree structures of sentences. The algorithms build an intersection tree between two syntactic trees, one for each of the two sentences being compared. An index of syntactic similarity between two sentences is the proportion of nodes in the two tree structures that are intersecting nodes.” As for content word overlap, this is simply the proportion of content words in adjacent sentences that share common content words (Graesser et al., 2002).

Coh-Metrix does not, however, measure visuals and layout variables; thus, I measure these for each item outside of Coh-Metrix. I measure whether there are visuals accompanying the item, the distance of the visuals from the item question, and the distance of the item question from the initial item prompt stem. Figure 3.3 shows an example QAM item with several of the visual and layout item characteristics identified and measured.
In this item, the Target Idea is idea number two as outlined in Figure 3.2—the greater the force is, the greater the change in motion will be. Therefore it receives a score of two for target idea. It is a selected response item, as students must choose among the three answers: it increases (correct answer), it decreases (incorrect answer) or it stays the same (incorrect answer). In this study, selected response items are scored as zero and constructed response items are scored as one for the Selected or Constructed Response variable. In terms of layout, there is a visual accompaniment to the text, and it is enmeshed with the item prompt stem (circled in green in Figure 3.3) and the item question (circled in orange in Figure 3.3). Thus, this item receives a score of zero for the Distance from Picture to Question variable.
In contrast, other items feature greater distance between the question and the visual element of the item (e.g. QAM items 6 and 7), and receive a score of 1 if the question is separated from the picture by one other question, two if the question and picture are separated by two other questions, and so on. In addition, the item question is separated from the item prompt stem by only the picture, as opposed to being separated by another question. Each element separating the item prompt stem from the item question (e.g. visuals, other questions, etc.) causes the item's score on the Distance from Stem to Question variable to go up by one. The Non-Coh-Metrix scores for this example item are also outlined in the below example Table 3.3.

Table 3.3. Example Scores for QAM Item 1 on Non-Coh-Metrix Predictor Variables

<table>
<thead>
<tr>
<th>ITEM</th>
<th>Selected or Constructed</th>
<th>Target Idea</th>
<th>Distance Picture to Prompt</th>
<th>Distance Stem to Prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td>QAM1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

The next step in preparing the predictor variables—calculated in Coh-Metrix and measured by me for visuals and layout—for the MLM analysis is to check the descriptive statistics (e.g. mean, standard deviation, range, and skewness) of the initial 20 linguistic characteristics of test items. These initial descriptive statistics can be found in Table 3.4 in the findings section (3.3.1.1 below). I check these scores for adequate variation in scores and normality issues. Some predictor variables are eliminated from the analysis at this phase (e.g. Inclusion of Visuals as all items
included a visual and thus had the same score on this measure). In other cases in which the variation in scores and/or normality of distribution of the variable measurements were marginal, I attempt various transformations such as taking the Log of the variable, taking the Square Root of the variable, or creating ordinal variables out of the continuous scores from the Coh-Metrix output in order to enhance the modeling potential of the variables. This practice has some advantages (e.g. enhancing interpretation of the MLM) and disadvantages (e.g. loss of some power in the case of creation of ordinal variables; Heck, Thomas, & Tabata, 2010). Details on variable transformations are also explained in the finding section (3.3.1.1. below), along with descriptive statistics for the transformed variables in Table 3.6. This is how I arrive at the 18 predictor variables (linguistic characteristics of test items) that are used in the exploratory phase of the MLM for RQ1 found in Section 3.3.1.2 in the findings below.

3.2.2.2. Multilevel Modeling

Much of the past item characteristic research has dealt with large-scale, standardized tests and relied on analytic procedures such as Differential Item Functioning (DIF; e.g. Bailey, 2005; Martiniello, 2008), confirmatory factor analysis (e.g. Farnsworth 2008), Analysis of Variance (ANOVA; e.g. Staley, 2005), and Rasch analysis (e.g. Pollitt et al., 2000). The QAM test is different from the tests that were analyzed in past studies because it is a classroom test that features only ten items. Therefore, it is inappropriate to utilize some of the statistical procedures that were
prominent in previous studies because they treat each item as a unit of analysis, as opposed to taking into account each individual student’s answers for each individual item on the test. Furthermore, aggregating student scores in order to create indices of differences in difficulty between emergent bilingual and native English speaking students—a popular technique in past research—is problematic because it sacrifices power and variance in the analysis (Heck, Thomas & Tabata, 2010).

Thus, I choose an innovative use of Multilevel Modeling (MLM) to answer my research questions for this chapter. MLM methodology has been growing in popularity over the past few decades in psychological, medical, and educational research, and it is beginning to appear in linguistics research. Indeed, MLM studies have recently been published in journals such as Language Testing (Barkaoui, 2010), International Journal of Bilingual Education and Bilingualism (Usborne, Caouette, Qumaaluk, & Taylor, 2009), and International Journal of Multilingualism (Joaristi, Lizasoain, Lukas, & Santiago, 2009). This type of analysis is also referred to as hierarchical linear modeling in education literature, (generalized linear) mixed model regression in psychology research, and random intercept modeling in medical studies.

MLM’s primary strength is that it allows the researcher to take into account both fixed and random effects. In other words, MLM allows us to consider several levels of cluster variables as random effect blocks, in addition to the fixed effects of a regular multiple regression. Other advantages of MLM include the lack of an assumption of homogeneity of regression slopes, being robust to missing data, and—most importantly—avoiding the assumption of independence of observations
via calculations of the random effect blocks (Fields, 2009; see also Hayes, 2006; Heck, Thomas, & Tabata, 2010; Singer & Willlet, 2003 for additional information on the advantages of MLM). However, in order conduct MLM, I first restructure the existing SCALE-uP data such that, instead of each row representing a person (as with a traditional repeated measures design), each row represents the interaction between an individual student and a test item (see Field, 2009: 763-767 for additional details on restructuring data in SPSS for MLM purposes). After the SCALE-uP data restructuring, there are 8,640 units for the analysis because each of the 864 student participants interacts with each of the ten QAM test items once.

My use of MLM differs from many of the other studies in education literature that use student, classroom, and school as cluster variables (e.g. Rethinam, Pyke, & Lynch, 2008) because I employ a type of generalized (non-linear, due to the binary outcome variable) repeated measures (longitudinal) MLM technique that is often used to measure individual growth curves (e.g. Baayen, Davidson, & Bates, 2008; Prado & Ullman, 2009). As Hayes (2006) points out, repeated measures/longitudinal MLM—the type of MLM employed in the present analysis—has many advantages over repeated measures analysis of variance, for example:

... there is no requirement that measurement begins and ends at the same time for every individual in the study, the measurement periods

---

26 Traditionally, MLM has been conducted in SAS, Stata, R, or HLM for SSI software, but the ability to conduct a “generalized linear mixed model” analysis (i.e. MLM with a binary outcome variable as needed for the present study) in SPSS became available in Dec. 2010 with Version 19; thus, SPSS is the software I use for the present analysis. Additional information about MLM functionality in SPSS can be found at http://www.spss.com/software/statistics/advanced-statistics/index.htm
do not have to be equally spaced, the number of measurement periods can vary between individuals, it is relatively easy to test models of the factors that predict how much or how quickly a person changes, and models of change can be estimated that do not presume that change is linear over time. It can also be used to model long sequences of measurements collected dozens or even hundreds of times on the same person in different situations or in response to different stimuli ... (ibid: 406)27

For this chapter, I design a two-level MLM analysis with the repeated measures (the individual test items) nested within individual students, which allows me to see whether relationships between the predictor/independent variables (item characteristics) and the outcome/dependent variable (students’ scores) vary across ESOL groups (Heck, Thomas, & Tabata, 2010). In other words, each of the students (the level-2 unit) is exposed to ten test items (the level-1 unit) that vary across different dimensions that represent predictor variables (the measures of test item linguistic complexity). The binary outcome variable is the score of right (1) or wrong (0). Generalized MLM then allows me to build models of the outcome

variable as a function of features of the test items (level-1 variables, e.g. syntactic complexity, item length, etc.) as well as characteristics of the students (the level-2 variable—ESOL status).

For research question one (Are characteristics of test items related to students’ item responses?), I estimate the level-1 variables (linguistic characteristics of test items) as random effects in order to investigate whether these features are related to students’ item-level scores. In order to answer research question two (Are there differences between students’ item-level scores across emergent bilingual and native English speaking groups of students?), I estimate the level-2 variable (ESOL status) as a random effect in order to investigate whether this student feature is related to students’ item-level scores. For research question 3 (Does the relationship between linguistic characteristics of test items and students’ item-level scores vary across emergent bilingual and native English speaking students?), I include cross-level interactions and comparisons of significantly different odds ratios to assess whether the level-1 variable relationships (linguistic complexity–scores) are related systematically to the level-2 variable of ESOL status.

At various stages of the analysis, I employ both exploratory (univariate modeling to test the main effects of each predictor variable individually) and confirmatory (multivariate modeling based on hypotheses from results of the exploratory, univariate models) MLM analyses to investigate my research questions. Figure 3.4 provides an overview and some details of the types of analyses employed for each research question.
Figure 3.4. Procedure for Analysis of Research Questions

RQ1) Are linguistic characteristics of test items related to students’ item-level scores?
- Exploratory univariate MLM analyses for each predictor variable (linguistic item characteristic)
- Confirmatory multivariate MLM analyses (building bigger models based on results and hypotheses from the exploratory modeling)

RQ2) Are there differences in students’ item-level scores across emergent bilingual and monolingual groups of students?
- Univariate analysis with ESOL status as predictor variable

RQ3) Does the relationship between linguistic characteristics of test items and students’ item-level scores vary across emergent bilingual and monolingual students?
- Tests for interactions between item characteristics and ESOL status in univariate MLM analyses
- Comparison of effect sizes of linguistic characteristic-student scores relationships across ESOL groups
3.3 FINDINGS

3.3.1. RQ1) Are Linguistic Characteristics of Test Items Related to Students’ Item-Level Scores?

For this research question, I hypothesized that some linguistic characteristics of test items would have a negative relationship with students’ correct responses (those features listed in the left-hand column in Figure 3.1 in Section 3.1.3. above); whereas, some characteristics would have a positive relationship with students’ correct responses (those features listed in the right-hand column in Figure 3.1 in Section 3.1.3 above).

The next sub-section (3.3.1.1 describes the use of descriptive statistics and transformations to arrive at the 18 predictor variables—the linguistic characteristics of test items—used in the initial exploratory univariate MLM. The following sub-sections present findings from the exploratory (Section 3.3.1.2) and the confirmatory (Section 3.3.1.3) MLM analyses, along with brief discussions of the results—elaborated discussions of the results and implications can be found in the concluding Section 3.4.

3.3.1.1. Descriptive Statistics for Initial Predictor Variables (Linguistic Characteristics) and Creation of Variables for Exploratory MLM
In order to run the MLM analyses, I first check the descriptive statistics of the initial 20 predictor variables from Coh-Metrix and that I measured—the linguistic characteristics of test items found in Appendix D—for normality (e.g. skewness) and adequate variation. Table 3.4 below provides an overview of some basic descriptive statistics for the initial predictor variables (linguistic characteristics of test items) across the 8640 cases. The variables are organized according to the linguistic area/category (format, syntax, lexis, discourse), following much of the literature discussion in the introductory Section 3.1.

Table 3.4. Descriptive Statistics for Initial Predictor Variables

<table>
<thead>
<tr>
<th>Linguistic Area of Predictor Variable</th>
<th>Linguistic Characteristic (Predictor Variable)</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
<th>Skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
<td>-Target Idea</td>
<td>2.07</td>
<td>1.20</td>
<td>3</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>-Selected or Constructed</td>
<td>0.71</td>
<td>0.46</td>
<td>1</td>
<td>-0.91</td>
</tr>
<tr>
<td></td>
<td>-Item Length (# words)</td>
<td>64.03</td>
<td>12.78</td>
<td>50</td>
<td>-0.47</td>
</tr>
<tr>
<td></td>
<td>-Item Length (# sentences)</td>
<td>7.46</td>
<td>1.22</td>
<td>4</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>-Distance Text Stem to Prompt</td>
<td>1.18</td>
<td>0.62</td>
<td>2</td>
<td>-0.14</td>
</tr>
<tr>
<td></td>
<td>-Distance Picture to Question</td>
<td>0.87</td>
<td>1.01</td>
<td>3</td>
<td>0.91</td>
</tr>
<tr>
<td>Syntax</td>
<td>-Sentence Length</td>
<td>8.68</td>
<td>1.77</td>
<td>5.12</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>-Modifiers per Noun Phrase</td>
<td>0.90</td>
<td>0.12</td>
<td>0.41</td>
<td>(1.78)</td>
</tr>
<tr>
<td></td>
<td>-Words before Main Verb</td>
<td>1.43</td>
<td>0.54</td>
<td>1.76</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>-Logical Operators</td>
<td>38.60</td>
<td>15.42</td>
<td>53.57</td>
<td>(-1.15)</td>
</tr>
<tr>
<td>Lexis</td>
<td>-Syllables per Word</td>
<td>1.28</td>
<td>0.05</td>
<td>0.18</td>
<td>(-1.07)</td>
</tr>
<tr>
<td></td>
<td>+Frequency of Lexical Items</td>
<td>1015.35</td>
<td>1079.91</td>
<td>4179.14</td>
<td>(1.84)</td>
</tr>
<tr>
<td></td>
<td>-Noun Hyponym</td>
<td>5.43</td>
<td>1.14</td>
<td>3.53</td>
<td>(1.47)</td>
</tr>
<tr>
<td></td>
<td>-Verb Hyponym</td>
<td>1.70</td>
<td>0.26</td>
<td>0.72</td>
<td>0.16</td>
</tr>
</tbody>
</table>
In the above Table 3.4 and in subsequent descriptive statistics tables, I indicate excessive skewness (above one or below negative one) with brackets. In cases when the normality of distribution of the variable measurements is marginal, I employ various transformations such as taking the Log of the variable, taking the Square Root of the variable, or creating ordinal variables out of the continuous scores from the Coh-Metrix output in order to enhance the modeling potential of the variables (Heck, Thomas, & Tabata, 2010). Table 3.5 below shows the values that I use to create ordinal variables out of the Coh-Metrix linguistic characteristics scores for some variables.

**Table 3.5. Creation of Ordinal Variables to Reduce Skewness and Enhance MLM Interpretation**

<table>
<thead>
<tr>
<th></th>
<th>Modifiers per Noun Phrase</th>
<th>Logical Operators</th>
<th>Syntactic Repetition</th>
<th>Content Word Repetition</th>
</tr>
</thead>
<tbody>
<tr>
<td>High (2)</td>
<td>.90-1.20</td>
<td>50.00-53.57</td>
<td>.54-.56</td>
<td>.24-.28</td>
</tr>
<tr>
<td>Mid (1)</td>
<td>.81-.89</td>
<td>15.15-49.99</td>
<td>.25-.55</td>
<td>.20-.24</td>
</tr>
<tr>
<td>Low (0)</td>
<td>.79-.80</td>
<td>.00-.15.14</td>
<td>.20-.24</td>
<td>.08-.19</td>
</tr>
</tbody>
</table>

Table 3.6 below shows that some of the transformations of the predictor variables were helpful in enhancing the variables’ suitability for MLM.
Table 3.6. Descriptive Statistics for Transformed Predictor Variables

<table>
<thead>
<tr>
<th>Linguistic Area of Predictor Variable</th>
<th>Linguistic Characteristic (Predictor Variable)</th>
<th>Mean</th>
<th>SD</th>
<th>N size</th>
<th>Range</th>
<th>Skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntax</td>
<td>-LogModifiers per Noun Phrase</td>
<td>-0.05</td>
<td>0.05</td>
<td>8640</td>
<td>0.18</td>
<td>(1.57)</td>
</tr>
<tr>
<td></td>
<td>-SqRtModifiers per Noun Phrase</td>
<td>0.95</td>
<td>0.06</td>
<td>8640</td>
<td>0.21</td>
<td>(1.68)</td>
</tr>
<tr>
<td></td>
<td>-OrdinalModifiers per Noun Phrase</td>
<td>1.27</td>
<td>0.75</td>
<td>8640</td>
<td>2</td>
<td>-0.50</td>
</tr>
<tr>
<td></td>
<td>-LogLogical Operators</td>
<td>1.59</td>
<td>0.16</td>
<td>8640</td>
<td>0.55</td>
<td>(-1.40)</td>
</tr>
<tr>
<td></td>
<td>-SqRtLogical Operators</td>
<td>5.93</td>
<td>1.87</td>
<td>8640</td>
<td>7.32</td>
<td>(-2.10)</td>
</tr>
<tr>
<td></td>
<td>-OrdinalLogical Operators</td>
<td>1.14</td>
<td>0.72</td>
<td>8640</td>
<td>2</td>
<td>-0.23</td>
</tr>
<tr>
<td>Lexis</td>
<td>-LogSyllables per Word</td>
<td>0.11</td>
<td>0.02</td>
<td>8640</td>
<td>0.06</td>
<td>(-1.13)</td>
</tr>
<tr>
<td></td>
<td>-SqRtSyllables per Word</td>
<td>1.13</td>
<td>0.02</td>
<td>8640</td>
<td>0.08</td>
<td>(-1.10)</td>
</tr>
<tr>
<td></td>
<td>+LogFrequency of Lexical Items</td>
<td>2.21</td>
<td>0.15</td>
<td>8640</td>
<td>0.45</td>
<td>-0.61</td>
</tr>
<tr>
<td></td>
<td>-LogNoun Hypernym</td>
<td>0.73</td>
<td>1.07</td>
<td>8640</td>
<td>0.25</td>
<td>(1.38)</td>
</tr>
<tr>
<td></td>
<td>-SqRtNoun Hypernym</td>
<td>2.32</td>
<td>0.23</td>
<td>8640</td>
<td>0.71</td>
<td>(1.42)</td>
</tr>
<tr>
<td>Discourse</td>
<td>+LogSyntactic Repetition</td>
<td>-1.07</td>
<td>0.25</td>
<td>8640</td>
<td>0.89</td>
<td>(1.16)</td>
</tr>
<tr>
<td></td>
<td>+SqRtSyntactic Repetition</td>
<td>0.30</td>
<td>0.10</td>
<td>8640</td>
<td>0.36</td>
<td>(1.64)</td>
</tr>
<tr>
<td></td>
<td>+OrdinalSyntactic Repetition</td>
<td>0.68</td>
<td>0.70</td>
<td>8640</td>
<td>2</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>+OrdinalContent Word Repetition</td>
<td>0.74</td>
<td>0.77</td>
<td>8640</td>
<td>2</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>+LogCausal Cohesion</td>
<td>-0.61</td>
<td>0.19</td>
<td>8640</td>
<td>0.53</td>
<td>(1.83)</td>
</tr>
<tr>
<td></td>
<td>+SqRtCausal Cohesion</td>
<td>0.21</td>
<td>0.26</td>
<td>8640</td>
<td>0.82</td>
<td>0.79</td>
</tr>
</tbody>
</table>

In the above Table 3.6 we see that -OrdinalModifiers per Noun Phrase (Skewness = -0.50), -OrdinalLogical Operators (Skewness = -0.23), +LogFrequency of
Lexical Items (Skewness = -0.61), +OrdinalSyntactic Repetition (Skewness = 0.54), +OrdinalContent Word Repetition (Skewness = 0.48), and +SqRtCausal Cohesion (Skewness = 0.79) emerge from the variable transformations with reduced skewness and are thus more suitable for plugging into the exploratory MLM to answer research question one (Are characteristics of test items related to students’ item responses?). These processes yield 18 predictor variables (linguistic characteristics of test items) to test in the exploratory univariate MLM analyses described in the next section.

3.3.1.2. Exploratory Univariate MLM Analyses of the Characteristics–Item-Level Scores Relationships

The next step in answering this research question involves using MLM in SPSS Version 19 as a means of estimating the level-1 predictor variables (linguistic complexity measures) as random effects in order to investigate whether these features are predictive of students’ item-level scores. Table 3.7 contains an overview of the results from these initial, exploratory univariate analyses of the characteristics–item-level scores relationships.28

---

28 Although these findings are presented in a single table, they represent the results of a series of separate, univariate analyses.
Table 3.7. Results from the Exploratory Univariate MLM Analyses of the Characteristics–Item-Level Scores Relationships (N=8640 for All Variables)

<table>
<thead>
<tr>
<th>Linguistic Area of Predictor Variable</th>
<th>Linguistic Characteristic (Predictor Variable)</th>
<th>B (SE)</th>
<th>P</th>
<th>Lower</th>
<th>Odds Ratio</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
<td>-Target Idea</td>
<td>-0.142 (0.019)</td>
<td>.000</td>
<td>0.836</td>
<td>0.868</td>
<td>0.900</td>
</tr>
<tr>
<td></td>
<td>-Selected or Constructed</td>
<td>-1.065 (0.050)</td>
<td>.000</td>
<td>0.312</td>
<td>0.345</td>
<td>0.380</td>
</tr>
<tr>
<td></td>
<td>-Item Length (# words)</td>
<td>0.029 (0.002)</td>
<td>.000</td>
<td>1.026</td>
<td>1.030</td>
<td>1.034</td>
</tr>
<tr>
<td></td>
<td>-Item Length (# sentences)</td>
<td>0.015 (0.018)</td>
<td>(.422)</td>
<td>0.979</td>
<td>1.015</td>
<td>1.052</td>
</tr>
<tr>
<td></td>
<td>-Distance Text Stem to Prompt</td>
<td>-0.364 (0.037)</td>
<td>.000</td>
<td>0.647</td>
<td>0.695</td>
<td>0.747</td>
</tr>
<tr>
<td></td>
<td>-Distance Picture to Question</td>
<td>-0.029 (.188)</td>
<td></td>
<td>0.930</td>
<td>0.971</td>
<td>1.014</td>
</tr>
<tr>
<td>Syntax</td>
<td>-Sentence Length (# words)</td>
<td>0.219 (0.013)</td>
<td>.000</td>
<td>1.213</td>
<td>1.245</td>
<td>1.277</td>
</tr>
<tr>
<td></td>
<td>-OrdinalModifiers per Noun Phrase</td>
<td>-1.031 (0.034)</td>
<td>.000</td>
<td>0.334</td>
<td>0.357</td>
<td>0.381</td>
</tr>
<tr>
<td></td>
<td>-Words before Main Verb</td>
<td>-0.152 (0.042)</td>
<td>.000</td>
<td>0.791</td>
<td>0.859</td>
<td>0.932</td>
</tr>
<tr>
<td></td>
<td>-OrdinalLogical Operators</td>
<td>-0.965 (0.035)</td>
<td>.000</td>
<td>0.356</td>
<td>0.381</td>
<td>0.408</td>
</tr>
<tr>
<td>Lexis</td>
<td>+LogFrequency of Lexical Items</td>
<td>-0.247 (0.052)</td>
<td>.000</td>
<td>0.705</td>
<td>0.781</td>
<td>0.865</td>
</tr>
<tr>
<td></td>
<td>-Verb Hypernymy</td>
<td>-0.869 (0.089)</td>
<td>.000</td>
<td>0.352</td>
<td>0.420</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>+Concreteness</td>
<td>0.005 (0.001)</td>
<td>.000</td>
<td>1.003</td>
<td>1.005</td>
<td>1.007</td>
</tr>
<tr>
<td>Discourse</td>
<td>+Connectives</td>
<td>-0.050 (0.002)</td>
<td>.000</td>
<td>0.948</td>
<td>0.951</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td>+OrdinalSyntactic Repetition</td>
<td>0.059 (0.032)</td>
<td>(.067)</td>
<td>0.996</td>
<td>1.061</td>
<td>1.130</td>
</tr>
<tr>
<td></td>
<td>+OrdinalContent Word Repetition</td>
<td>1.023 (0.033)</td>
<td>.000</td>
<td>2.608</td>
<td>2.781</td>
<td>2.965</td>
</tr>
<tr>
<td></td>
<td>+Intentional Cohesion</td>
<td>0.023 (0.003)</td>
<td>.000</td>
<td>1.017</td>
<td>1.023</td>
<td>1.030</td>
</tr>
<tr>
<td></td>
<td>+SqRtCausal Cohesion</td>
<td>-1.010 (0.088)</td>
<td>.000</td>
<td>0.307</td>
<td>0.364</td>
<td>0.433</td>
</tr>
</tbody>
</table>
Strength of Relationships

In Table 3.7 we see that the majority of the relationships between the predictor variables (linguistic characteristics of test items) and students' item-level scores (correct answers) are significant at the $p < .01$ level. Of the 18 variables tested in the exploratory univariate analyses, the only variables that are not demonstrating a significant relationship with students' item-level scores are: *-Item Length as Number of Sentences* ($p = 0.422$), *-Distance Picture to Prompt* ($p = 0.188$), and *+OrdinalSyntactic Repetition* (proportion of intersection tree nodes between all adjacent sentences; $p = 0.067$).

Given the large sample size in this analysis, significant $p$ values are not surprising; rather, the most important results in Table 3.7, are the Odds Ratios (ORs) that encompass the effect size of each relationship. An OR of 1 indicates no relationship; thus, ORs increasingly greater than 1.0 indicate increasingly positive relationships and ORs decreasing away from 1.0 toward zero indicate increasingly negative relationships. For example, the ORs for *-Selected or Constructed* (OR = 0.345), *-OrdinalModifiers per Noun Phrase* (mean number of modifiers that modify the head noun; OR = 0.357), *-OrdinalLogical Operators* (incidence of logical operators including Boolean operators; OR = 0.381), *-Noun Hyponymy* (mean hypernym/multiplicity of meaning value of main verbs; OR = 0.420), and *+SqRtCausal Cohesion* (ratio of causal particles to causal verbs; OR = 0.364) all represent relatively large effect sizes for the negative relationship between these characteristics and
students’ correct answers. In other words, as the measures for these variables increase to their maximums, students are approximately one third as likely to get a correct answer as compared with the scenario in which the measure for that characteristic is at its minimum. Similarly, the OR for +OrdinalContent Word Repetition (proportion of content words in adjacent sentences that share common content words; OR = 2.781) displays a relatively large effect size for the positive relationship between increasing values of this variable and students’ correct answers. In this case, students are nearly three times as likely to get a correct answer as content word overlap (how often content words overlap between two adjacent sentences) increases to its maximum value.

Other variables also demonstrate significant relationships; however, the effect sizes (ORs) of these relationships range from moderate to very small. For example, -Distance Text Stem to Prompt (OR = 0.695), -Sentence Length (OR = 1.245), and +LogFrequency of Lexical Items (mean frequency of content words; OR = 0.781) have approximately half the effect size of the variables listed above with ORs in the range of 0.420 to 0.345. -Target Idea (OR = 0.868) as a predictor variable has a relatively small effect size; in fact, smaller than all but four of the 18 linguistic characteristics variables, although -Number of Modifiers per Noun Phrase (mean number of words before the main verb of the main clause; OR = 0.859) has approximately the same effect size as the -Target Idea variable. Some of the weakest

---

29 Here is a useful reference for interpretation of Odds Ratios from Heck, Thomas & Tabata (2010): Cohen’s $d$ value of 0.2 (small effect size) roughly corresponds to an OR of 1.3, Cohen’s $d$ value of 0.5 (medium effect size) roughly corresponds to an OR of 1.9, and Cohen’s $d$ value of 0.8 (large effect size) roughly corresponds to an OR of 2.7.
relationships found in the exploratory univariate analyses are -Item Length as Number of Words (OR = 1.030), +Concreteness (concreteness value of all content words that match a word in the MRC database; OR = 1.005), +Connectives (Incidence of all connectives; OR = 0.951), and +Intentional Cohesion (ratio of intentional particles to the incidence of causal content; OR = 1.023). The effect sizes of these four variables are negligible, as they hover very close to 1.0.

Directionality of Relationships

For 10 of the 18 variables tested in the exploratory univariate MLM, I find relationships between the linguistic characteristics of test items and students’ item-level scores in the predicted direction of my hypotheses. Specifically, the format variables of -Selected or Constructed (binary values of zero as a selected response and one as a constructed response; OR = 0.345) and -Distance Text Stem to Prompt (amount of distance between the original set up of an item and the eventual item question; OR = 0.695) have negative relationships with students’ scores, as predicted by past literature and my hypotheses. The syntactic variables of -OrdinalLogical Operators (incidence of logical operators including Boolean operators; OR = 0.381), -OrdinalModifiers per Noun Phrase (mean number of modifiers per noun-phrase that modify the head noun; OR = 0.357), and -Modifiers per Noun Phrase (mean number of words before the main verb of the main clause; mean number of words before the main verb of the main clause in sentences; OR = 0.859) also have predicted negative relationships with students’ scores. In other
words, as these measures of format and syntactic complexity increase, students’
item-level scores decreased.

Similarly, the lexical variable of *-Noun Hypernymy (mean hypernym/multiplicity of meaning value of main verbs; mean hypernym/multiplicity of meaning value of main verbs; OR = 0.420) have a negative relationship with students’ scores with a relatively large effect size, as predicted by previous studies using Coh-Metrix to assess readability of texts (Graesser et al., 2002). As for the discourse variables, the only relationship in the predicted direction with a large effect size is *+OrdinalContent Word Repetition* (proportion of content words in adjacent sentences that share common content words; OR = 2.781), which is in line with previous research that has argued that repetition of content words is integral for the readability of texts for both native English speaking and emergent bilingual students (Crossley et al., 2008). The other discourse variables with predicted positive relationships with students’ item-level scores, *+Concreteness* (concreteness value of all content words in the item prompt that match a word in the MRC database; OR = 1.005) and *+Intentional Cohesion* (ratio of intentional particles to the incidence of causal content; OR = 1.023), have only very small effect sizes and thus represent less interesting findings.

While the directionality of the above relationships are predicted and help to confirm my hypotheses for this research question, there are also some surprises at this phase in the analysis. For example, the format variable *-Item Length as Number of Words* (OR = 1.030) and the syntactic variable *-Sentence Length* (mean number of words per sentence; OR = 1.245) have positive relationships with students’ item-
level scores; whereas, these variables were predicted to have a negative relationship with student’ scores by past research and my hypotheses. Also, the discourse variable +Connectives (Incidence of all connectives; OR = 0.951)—argued to be an important component of text cohesion (e.g. Halliday & Hasan, 1976; Kemper, 1988)—have a negative relationship with students’ scores, although the OR for this relationship is very small, much like the OR for the surprising relationship of -Item Length as Number of Words—students’ scores.

Two linguistic characteristics of test items have surprising relationships with students’ item-level scores, given past literature and my hypotheses. The discourse variable +SqRtCausal Cohesion (ratio of causal particles to causal verbs; OR = 0.364) was predicted to have a positive relationship with students’ scores (i.e., the more Causal Cohesion in an item, the better students were predicted to perform on that item), particularly since the expression of causal relationships is a key aspect of science texts. However, this variable demonstrates a negative relationship with students’ scores and this relationship has a relatively large effect size. Also, the negative relationship between the average +LogFrequency of Lexical Items (mean frequency of content words; OR = 0.781) and students’ item-level scores is surprising (even though the effect size is relatively small), given that lexical frequency has been supported as an important variable across many strands of research such as testing, readability, and SLA/bilingualism, as outlined in the introduction to this chapter. Possible reasons for these surprising results are discussed in the next Section 3.4.

One of the most interesting findings from the exploratory univariate analyses is that the -Target Idea (OR = 0.868) predictor variable—while demonstrating a
negative relationship with students’ item-level scores as hypothesized—has a relatively small effect size. In other words, the linguistic characteristics of test items are more predictive of students’ scores than the difficulty assumed/imputed by the test makers of underlying content and target ideas of the test items. The implications of this finding are great in that they show the importance of language of the tests in explaining variation in students’ scores and are further discussed in Section 3.4.

It is also important to note that certain variables from each of the four linguistic areas/categories (format, syntax, lexis, and discourse) emerge as the strongest predictors of students’ scores in each area/category. Specifically, the format variables of -Selected or Constructed (OR = 0.345) and -Distance Text Stem to Prompt (OR = 0.695), the syntax variables of -OrdinalLogical Operators (incidence of logical operators including Boolean operators; OR = 0.381) and -OrdinalModifiers per Noun Phrase (mean number of modifiers that modify the head noun; OR = 0.357), the lexis variable of -Noun Hypernymy (mean hypernym/multiplicity of meaning value of main verbs; OR = 0.420), and the discourse variable of +OrdinalContent Word Repetition (proportion of content words in adjacent sentences that share common content words; OR = 2.781) are strong predictors of students’ item-level scores in the expected direction based on past literature and hypotheses. The results of these variables from each of the four linguistic areas/categories in the exploratory MLM help determine which variables to build into the confirmatory multivariate MLM for the next phase of the analysis.
Based on the previous literature, my hypotheses, and the above findings from the exploratory univariate MLM analyses, I conduct confirmatory MLM analyses to see if there is overlap in the variation accounted for by these linguistic characteristics of test items. After attempts with many iterations of variables from the linguistic categories in the exploratory study, I find that the strongest predictor variables in the univariate analyses also yield the strongest multivariate models. In particular, I try out each variable that emerges as significant for each of the four categories in the univariate analyses (format-four variables, syntax-four variables, lexis-three variables, and discourse-four variables) with collocations of the significant variables from each of the other categories in order to arrive at the best multivariate model.

Table 3.8 provides a summary of the multivariate MLM with that emerges with the best fit, as measured using the information criterion provided in the SPSS output.
Table 3.8. Results from Confirmatory Multivariate MLM Analyses

<table>
<thead>
<tr>
<th>Linguistic Area of Predictor Variable</th>
<th>Linguistic Characteristic (Predictor Variable)</th>
<th>B (SE)</th>
<th>P</th>
<th>Lower</th>
<th>Odds Ratio</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
<td>-Selected or Constructed</td>
<td>-1.293 (0.059)</td>
<td>.000</td>
<td>0.245</td>
<td>0.274</td>
<td>0.308</td>
</tr>
<tr>
<td>Syntax</td>
<td>-OrdinalLogical Operators</td>
<td>-0.504 (0.050)</td>
<td>.000</td>
<td>0.548</td>
<td>0.604</td>
<td>0.666</td>
</tr>
<tr>
<td>Lexis</td>
<td>-Verb Hypernymy</td>
<td>-1.398 (0.109)</td>
<td>.000</td>
<td>0.199</td>
<td>0.247</td>
<td>0.306</td>
</tr>
<tr>
<td>Discourse</td>
<td>+OrdinalContent Word Repetition</td>
<td>0.718 (0.047)</td>
<td>.000</td>
<td>1.872</td>
<td>2.051</td>
<td>2.247</td>
</tr>
</tbody>
</table>

Here again we find linguistic characteristics of test items in each of the linguistic areas/categories that account for variation in students' item-level scores: the format variable -Selected or Constructed (OR = 0.274), the syntax variable -OrdinalLogical Operators (incidence of logical operators including Boolean operators; OR = 0.604), the lexis variable -Noun Hypernymy (mean hypernym/multiplicity of meaning value of main verbs; OR = 0.247), and the discourse variable +OrdinalContent Word Repetition (proportion of content words in adjacent sentences that share common content words; OR = 2.051). The ORs for these variables in the multivariate MLM are similar to the ORs found in their respective univariate analyses; this provides evidence that they generally account for different portions of the variation in students' item-level scores in this SCALE-uP sample.
Section 3.3.1.4. Summary of Findings from the Exploratory and Confirmatory Multivariate MLM Analyses for Research Question One: Are linguistic characteristics of test items related to students’ item-level scores?

In sum, my hypotheses for research question one are essentially confirmed—linguistic characteristics of test items are predictive of students’ item-level scores in both negative and positive directions, depending on the variables, with several variables from each linguistic area/category emerging as the strongest predictors:

- Format - Selected or Constructed Response Type and Distance Text Stem to Prompt;
- Syntax - Ordinal Logical Operators and Ordinal Modifiers per Noun Phrase;
- Lexis - Verb Hypernym;
- Discourse - Ordinal Content Word Repetition.

However, some of the variables that emerge as significant with the biggest effect sizes in the univariate and multivariate analyses are not those with the strongest support in previous studies, and some variables with strong support from past literature are not predictive in the expected direction (e.g. $+LogLexical\ Frequency$ and $+SqrtTemporal\ Cohesion$ having negative relationships with students’ scores). Further discussion of these findings and implications related to the relationship between linguistic characteristics of test items and students’ item-level scores can be found in the conclusion of this analytic chapter (Section 3.4 below).
3.3.2. RQ2) Are There Differences between Students’ Item-Level Scores Across ESOL Groups?

Based on national trends and the achievement gaps found in my previous analysis (Dissertation Chapter 2), I hypothesized that students’ item-level scores would vary significantly across emergent bilingual and native English speaking groups. In the next sub-section (Section 3.3.2.1) I present descriptive statistics (Table 3.9) showing differences between students’ item-level scores across ESOL groups, and then I present results from univariate MLM (Section 3.3.2.2) to further answer the question of whether ESOL status is a significant predictor of students’ scores.

Section 3.3.2.1. Descriptive Statistics Across ESOL Groups

Table 3.9 provides an overview of the descriptive statistics for students’ item-level scores across ESOL groups.

Table 3.9. Descriptive Statistics of Students’ Item-Level Scores Across ESOL Groups

<table>
<thead>
<tr>
<th>ESOL GROUP</th>
<th>Mean Correct</th>
<th>SD</th>
<th>N size</th>
<th>Range</th>
<th>Skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESOL Never (Native English Speaking)</td>
<td>0.49</td>
<td>0.50</td>
<td>4320</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>ESOL Current (Emergent Bilingual)</td>
<td>0.34</td>
<td>0.48</td>
<td>4320</td>
<td>1</td>
<td>0.66</td>
</tr>
</tbody>
</table>
As we see in the table, native English speaking students achieve correct answers approximately half of the time, while emergent bilingual students achieve correct answers only approximately one third of the time.

Section 3.3.2.2. Univariate MLM Investigating ESOL Group Status as a Predictor of Students’ Item-Level Scores

The below table 3.10 shows the results of a univariate MLM investigating the main effect of ESOL status on students’ item-level scores.

Table 3.10. Differences between Responses across ESOL Groups in Univariate MLM

<table>
<thead>
<tr>
<th>Predictor Variable</th>
<th>B (SE)</th>
<th>P</th>
<th>Lower Odds Ratio</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESOL status</td>
<td>-0.575 (0.053)</td>
<td>.000</td>
<td>0.507</td>
<td>0.624</td>
</tr>
</tbody>
</table>

Table 3.10 reiterates the pattern in the descriptive statistics: an OR of 0.563 for ESOL status indicates that emergent bilingual students are approximately half as likely to get correct answers as native English speaking students. This represents a strong relationship, with significance at the \( p < .01 \) level.
Section 3.3.2.3. Summary of Findings for Research Question Two: Are there differences between students’ item-level scores across ESOL groups?

In sum, Table 3.9 and the results of the univariate analysis found in Table 3.10 all confirm my hypothesis for this research question: students’ item-level scores do differ across ESOL groups, with emergent bilingual students having significantly fewer correct responses than their native English speaking peers.

It is also important to note that the results of this Research Question Two combined with those from Research Question One help motivate comparisons of ORs across groups found in the findings section for Research Question Three, (Section 3.3.3 below) which addresses the issue of differences between the relationship between linguistic characteristics of test items and students’ item-level scores across ESL groups. This third research question can help shed light on the sources of achievement gaps for emergent bilingual students, the key issue at the heart of my dissertation study.

3.3.3. RQ3) Does the Relationship between Linguistic Characteristics of Test Items and Students’ Item-Level Scores Vary across ESOL Groups?

As mentioned above, my hypotheses for this third research question were less clear due to conflicting results in previous studies, some of which argue that linguistic complexity of test items penalize emergent bilingual students (e.g. Abedi, 2006; Martiniello, 2008; Wolf & Leon, 2009) and some of which argue that linguistic
aspects of test items are construct-relevant and a much larger proportion of
differences in performance between groups may be due to factors other than
linguistic characteristics of test items (e.g. Abedi & Herman, 2010; Kieffer et al.,
2009; Koenig & Bachman, 2004; Farnsworth, 2008).

However, given the additional findings from readability studies with emergent bilingual students showing that certain Coh-Metrix variables (syntactic similarity, lexical frequency, and content word overlap) were predictive of readability of texts for emergent bilinguals, I hypothesized that there would be both negative and positive relationships between linguistic characteristics of test items and students’ item-level scores, and these relationships would be stronger in the case of emergent bilingual students for some item characteristic variables with the strongest support in previous literature (e.g. item length, syntactic similarity, lexical frequency, and content word overlap).

The following section 3.3.3.1 outlines the findings for tests for interaction between the linguistic characteristic variables and ESOL group status, and then Section 3.3.3.2 provides comparisons of the ORs for the main effects of these variables for each group.

3.3.3.1. Tests for Interactions between Each of the Predictor Variables and ESOL Group Status

Tables 3.11 through Table 3.14 summarize the findings from the multivariate/interaction MLM analyses for RQ3 and are organized according to the
linguistic category of the predictor variables (Format, Syntax, Lexis, and Discourse). Although the main effects for each of the variables are reported, they are not interpreted here; rather, the important information is the significance ($p$ value) of the interaction between the two variables. Thus, I have bolded the rows in the tables representing the interactions so that the $p$ value information may become more salient.

Table 3.11. Overview of Findings for Format Variables Research Question Three

<table>
<thead>
<tr>
<th>Linguistic Characteristic (Predictor Variable)</th>
<th>Main Effects and Interactions with ESOL Group</th>
<th>B (SE)</th>
<th>$P$</th>
<th>Lower Bound</th>
<th>Odds Ratio</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>-Target Idea</td>
<td>-Target Idea Main Effect</td>
<td>-0.164 (0.019)</td>
<td>.000</td>
<td>0.849</td>
<td>0.807</td>
<td>0.894</td>
</tr>
<tr>
<td></td>
<td>ESOL Group Main Effect</td>
<td>-0.700 (0.093)</td>
<td>.000</td>
<td>0.414</td>
<td>0.497</td>
<td>0.596</td>
</tr>
<tr>
<td></td>
<td>-Target Idea *ESOL Group Interaction</td>
<td>0.042 (0.038)</td>
<td>(.268)</td>
<td>0.968</td>
<td>1.043</td>
<td>1.124</td>
</tr>
<tr>
<td>-Selected or Constructed</td>
<td>-Selected or Constructed Main Effect</td>
<td>-0.930 (0.072)</td>
<td>.000</td>
<td>0.343</td>
<td>0.395</td>
<td>0.454</td>
</tr>
<tr>
<td></td>
<td>ESOL Group Main Effect</td>
<td>-0.449 (0.087)</td>
<td>.000</td>
<td>0.538</td>
<td>0.638</td>
<td>0.757</td>
</tr>
<tr>
<td></td>
<td>-Selected or Constructed *ESOL Group Interaction</td>
<td>-0.302 (0.101)</td>
<td>.003</td>
<td>0.607</td>
<td>0.739</td>
<td>0.900</td>
</tr>
<tr>
<td>-Item Length (# words)</td>
<td>-Item Length (# words) Main Effect</td>
<td>0.038 (0.003)</td>
<td>.000</td>
<td>1.033</td>
<td>1.038</td>
<td>1.044</td>
</tr>
<tr>
<td></td>
<td>ESOL Group Main Effect</td>
<td>0.366 (0.250)</td>
<td>(.143)</td>
<td>0.884</td>
<td>1.442</td>
<td>2.353</td>
</tr>
</tbody>
</table>

Although these findings are presented in tables that are clustered according to linguistic category, they represent the results of a series of separate, multivariate analyses to test for interactions between the variables.
<table>
<thead>
<tr>
<th></th>
<th>(# words) *ESOL Group Interaction</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Item Length (# sentences)</strong></td>
<td>Item Length (# sentences) Main Effect</td>
<td>0.020</td>
<td>(0.036)</td>
<td>0.439</td>
<td>0.970</td>
</tr>
<tr>
<td></td>
<td>ESOL Group Main Effect</td>
<td>-0.564</td>
<td>(0.281)</td>
<td>0.044</td>
<td>0.328</td>
</tr>
<tr>
<td></td>
<td>*<em>Item Length (# sentences) <em>ESOL Group Interaction</em></em></td>
<td>-0.006</td>
<td>(0.037)</td>
<td>(.871)</td>
<td>0.924</td>
</tr>
<tr>
<td><strong>Distance Text Stem to Prompt</strong></td>
<td><strong>Distance Text Stem to Prompt Main Effect</strong></td>
<td>-0.416</td>
<td>(-0.146)</td>
<td>.000</td>
<td>0.597</td>
</tr>
<tr>
<td></td>
<td>ESOL Group Main Effect</td>
<td>-0.733</td>
<td>(0.099)</td>
<td>.000</td>
<td>0.395</td>
</tr>
<tr>
<td></td>
<td>*<em>Distance Text Stem to Prompt <em>ESOL Group Interaction</em></em></td>
<td>0.101</td>
<td>(0.074)</td>
<td>(.168)</td>
<td>0.958</td>
</tr>
<tr>
<td><strong>Distance Picture to Question</strong></td>
<td><strong>Distance Picture to Question Main Effect</strong></td>
<td>-0.092</td>
<td>(0.031)</td>
<td>.003</td>
<td>1.032</td>
</tr>
<tr>
<td></td>
<td>ESOL Group Main Effect</td>
<td>-0.385</td>
<td>(0.063)</td>
<td>.000</td>
<td>0.601</td>
</tr>
<tr>
<td></td>
<td>*<em>Distance Picture to Question <em>ESOL Group Interaction</em></em></td>
<td>-0.264</td>
<td>(0.045)</td>
<td>.000</td>
<td>.703</td>
</tr>
</tbody>
</table>

Again, in this table, I do not interpret the main effects of the variables; rather, I focus on whether there is a significant interaction between the variables. Table 3.11 shows that, of the format variables, there are significant interactions between the variables–item-level scores relationship and ESOL Group Status (i.e. significantly different ORs for each group, or relationships that vary across groups) for three variables:

- **Selected or Constructed*ESOL Group Interaction (p = .003)**
- **Item Length (# words)*ESOL Group Interaction (p = .000)**
- Distance Picture to Question*ESOL Group Interaction \((p = .000)\)

Thus, the ORs for the main effects of these variables for each group are compared in Section 3.3.3.2 below to show how the relationships vary across groups.

**Table 3.12. Overview of Findings for Syntax Variables Research Question Three**

<table>
<thead>
<tr>
<th>Linguistic Characteristic (Predictor Variable)</th>
<th>Main Effects and Interactions with ESOL group</th>
<th>B (SE)</th>
<th>(P)</th>
<th>Lower Bound</th>
<th>Odds Ratio</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>-Sentence Length (# words)</td>
<td>-Sentence Length (# words) Main Effect</td>
<td>0.263 (0.018)</td>
<td>.000</td>
<td>1.255</td>
<td>1.301</td>
<td>1.348</td>
</tr>
<tr>
<td></td>
<td>ESOL Group Main Effect</td>
<td>0.132 (0.236)</td>
<td>(.577)</td>
<td>0.718</td>
<td>1.141</td>
<td>1.813</td>
</tr>
<tr>
<td></td>
<td>-Sentence Length (# words) *ESOL Group Interaction</td>
<td>-0.086 (0.026)</td>
<td>.001</td>
<td>0.871</td>
<td>0.917</td>
<td>0.966</td>
</tr>
<tr>
<td>-OrdinalModifiers per Noun Phrase</td>
<td>-OrdinalModifiers per Noun Phrase Main Effect</td>
<td>-0.997 (0.048)</td>
<td>.000</td>
<td>0.336</td>
<td>0.369</td>
<td>0.405</td>
</tr>
<tr>
<td></td>
<td>ESOL Group Main Effect</td>
<td>-0.591 (0.100)</td>
<td>.000</td>
<td>0.455</td>
<td>0.554</td>
<td>0.674</td>
</tr>
<tr>
<td></td>
<td>-OrdinalModifiers per Noun Phrase *ESOL Group Interaction</td>
<td>-0.112 (0.068)</td>
<td>(.099)</td>
<td>0.783</td>
<td>0.894</td>
<td>1.021</td>
</tr>
<tr>
<td>-Words before Main Verb</td>
<td>-Words before Main Verb Main Effect</td>
<td>-0.030 (0.058)</td>
<td>(.603)</td>
<td>0.919</td>
<td>1.031</td>
<td>1.156</td>
</tr>
<tr>
<td></td>
<td>ESOL Group Main Effect</td>
<td>-0.034 (0.130)</td>
<td>(.796)</td>
<td>0.749</td>
<td>0.967</td>
<td>1.248</td>
</tr>
<tr>
<td></td>
<td>-Words before Main Verb *ESOL Group Interaction</td>
<td>-0.408 (0.085)</td>
<td>.000</td>
<td>0.563</td>
<td>0.665</td>
<td>0.786</td>
</tr>
<tr>
<td>-Ordinal Logical Operators</td>
<td>-Ordinal Logical Operators Main Effect</td>
<td>-0.954 (0.049)</td>
<td>.000</td>
<td>0.350</td>
<td>0.385</td>
<td>0.424</td>
</tr>
</tbody>
</table>
Table 3.12 shows that, of the syntax variables, there are significant interactions with ESOL Group Status (i.e. significantly different ORs for each group; relationships vary across groups) for two variables:

- **Sentence Length (# words)***ESOL Group Interaction (p = .001)
- **Words before Main Verb***ESOL Group Interaction (p = .000)

Thus, the ORs for the main effects of these variables for each group are compared in Section 3.3.3.2 below to show how the relationships vary across groups.

**Table 3.13. Overview of Findings for Lexis Variables Research Question Three**

<table>
<thead>
<tr>
<th>Linguistic Characteristic (Predictor Variable)</th>
<th>Main Effects and Interactions with ESOL Group</th>
<th>B (SE)</th>
<th>P</th>
<th>Lower Bound</th>
<th>Odds Ratio</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>+LogFrequency of Lexical Items</td>
<td>+LogFrequency of Lexical Items Main Effect</td>
<td>-0.213 (0.074)</td>
<td>.004</td>
<td>0.699</td>
<td>0.808</td>
<td>0.934</td>
</tr>
<tr>
<td>ESOL Group Main Effect</td>
<td></td>
<td>-0.386 (0.296)</td>
<td>(.193)</td>
<td>0.380</td>
<td>0.680</td>
<td>1.215</td>
</tr>
<tr>
<td>+LogFrequency of Lexical Items*ESOL Group Interaction</td>
<td></td>
<td>-0.081 (0.105)</td>
<td>(.437)</td>
<td>0.750</td>
<td>0.922</td>
<td>1.132</td>
</tr>
<tr>
<td>-Verb Hypernymy</td>
<td>-Verb Hypernymy Main Effect</td>
<td>-1.116 (0.124)</td>
<td>.000</td>
<td>0.257</td>
<td>0.327</td>
<td>0.417</td>
</tr>
<tr>
<td>ESOL Group Main Effect</td>
<td></td>
<td>-1.487 (0.306)</td>
<td>.000</td>
<td>0.124</td>
<td>0.226</td>
<td>0.412</td>
</tr>
<tr>
<td>-Verb Hypernymy*ESOL</td>
<td></td>
<td>0.520 (0.179)</td>
<td>.004</td>
<td>1.183</td>
<td>1.681</td>
<td>2.388</td>
</tr>
</tbody>
</table>
Table 3.13 shows that, of the syntax variables, there are significant interactions with ESOL Group Status (i.e. significantly different ORs for each group; relationships vary across groups) for two variables:

- Verb Hypernymy*ESOL Group Interaction ($p = .004$)
- +Concreteness*ESOL Group Interaction ($p = .000$)

Thus, the ORs for the main effects of these variables for each group are compared in Section 3.3.3.2 below to show how the relationships vary across groups.

Table 3.14. Overview of Findings for Discourse Variables Research Question Three

<table>
<thead>
<tr>
<th>Linguistic Characteristic (Predictor Variable)</th>
<th>Main Effects and Interactions with ESOL Group</th>
<th>B (SE)</th>
<th>P</th>
<th>Lower Bound</th>
<th>Odds Ratio</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>+Connectives</td>
<td>+Connectives Main Effect</td>
<td>-0.047</td>
<td>0.003</td>
<td>0.949</td>
<td>0.954</td>
<td>0.959</td>
</tr>
<tr>
<td></td>
<td>ESOL Group Main Effect</td>
<td>-0.185</td>
<td>(0.231)</td>
<td>0.528</td>
<td>0.831</td>
<td>1.308</td>
</tr>
<tr>
<td></td>
<td>+Connectives*ESOL Group Interaction</td>
<td>-0.009</td>
<td>(0.004)</td>
<td>0.984</td>
<td>0.991</td>
<td>0.999</td>
</tr>
<tr>
<td>+OrdinalSyntactic Repetition</td>
<td>+OrdinalSyntactic Repetition Main Effect</td>
<td>-0.041</td>
<td>(0.046)</td>
<td>.368</td>
<td>0.878</td>
<td>0.960</td>
</tr>
<tr>
<td></td>
<td>ESOL Group Main Effect</td>
<td>-0.763</td>
<td>(0.067)</td>
<td>.000</td>
<td>0.409</td>
<td>0.466</td>
</tr>
<tr>
<td></td>
<td>+OrdinalSyntactic Repetition*ESOL Group Interaction</td>
<td>0.222</td>
<td>(0.065)</td>
<td>.001</td>
<td>1.100</td>
<td>1.249</td>
</tr>
<tr>
<td>+OrdinalContent</td>
<td>+OrdinalContent</td>
<td>1.006</td>
<td>.000</td>
<td>2.493</td>
<td>2.733</td>
<td>2.997</td>
</tr>
</tbody>
</table>
Table 3.14 shows that, of the format variables, there are significant interactions with ESOL Group Status (i.e. significantly different ORs for each group; relationships vary across groups) for three variables:

- **+Connectives*ESOL Group Interaction** *(p = .027)*

- **+OrdinalSyntactic Repetition*ESOL Group Interaction** *(p = .001)*

- **+Intentional Cohesion*ESOL Group Interaction** *(p = .000)*

Thus, the ORs for the main effects of these variables for each group are compared in Section 3.3.3.2 below to show how the relationships vary across groups.

In sum, the following 10 variables have significant interactions with ESOL Group Status:

*Format*
- Selected or Constructed*ESOL Group Interaction \( (p = .003) \)
- Item Length (# words)*ESOL Group Interaction \( (p = .000) \)
- Distance Picture to Question*ESOL Group Interaction \( (p = .000) \)

**Syntax**
- Sentence Length (# words)*ESOL Group Interaction \( (p = .001) \)
- Words before Main Verb*ESOL Group Interaction \( (p = .000) \)

**Lexis**
- Verb Hypernymy*ESOL Group Interaction \( (p = .004) \)
- Concreteness*ESOL Group Interaction \( (p = .000) \)

**Discourse**
- Connectives*ESOL Group Interaction \( (p = .027) \)
- Ordinal Syntactic Repetition*ESOL Group Interaction \( (p = .001) \)
- Intentional Cohesion*ESOL Group Interaction \( (p = .001) \)

Thus, I compare the ORs for each of these 10 variables across ESOL Groups to see how the relationships vary across groups in the next Section 3.3.3.2.

**Section 3.3.3.2. Differences between Odds Ratios for Main Effects of Linguistic Complexity Variables Across ESOL Groups**

Tables 3.15 through 3.18 show the differences between ORs across the ESOL groups and are organized according to the linguistic category of the predictor.
variables (Format, Syntax, Lexis, and Discourse). In these tables, it is important to compare the OR across the two groups to look for differences that would index a differential effect of the predictor variable within each group. Thus, I have bolded the columns in the tables representing the ORs so that they may become more visually salient.

Table 3.15. Differences between Odds Ratios Across ESOL Groups for Significant Format Predictor Variables

<table>
<thead>
<tr>
<th>Linguistic Characteristic (Predictor Variable) for Each ESOL Group</th>
<th>B (SE)</th>
<th>P</th>
<th>Lower Bound</th>
<th>Odds Ratio</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selected or Constructed Response</td>
<td>Emergent Bilinguals</td>
<td>-1.244 (0.071)</td>
<td>.000</td>
<td>0.251</td>
<td><strong>0.288</strong></td>
</tr>
<tr>
<td>Native English Speaking Students</td>
<td>-0.926 (0.70)</td>
<td>.000</td>
<td>0.345</td>
<td><strong>0.396</strong></td>
<td>0.454</td>
</tr>
<tr>
<td>Item Length (# words)</td>
<td>Emergent Bilinguals</td>
<td>0.022 (0.003)</td>
<td>.000</td>
<td>1.017</td>
<td><strong>1.022</strong></td>
</tr>
<tr>
<td>Native English Speaking Students</td>
<td>0.037 (0.003)</td>
<td>.000</td>
<td>1.033</td>
<td><strong>1.038</strong></td>
<td>1.043</td>
</tr>
<tr>
<td>Distance Picture to Question</td>
<td>Emergent Bilinguals</td>
<td>-0.172 (0.033)</td>
<td>.000</td>
<td>0.789</td>
<td><strong>0.842</strong></td>
</tr>
<tr>
<td>Native English Speaking Students</td>
<td>0.091 (0.030)</td>
<td>.003</td>
<td>1.031</td>
<td><strong>1.095</strong></td>
<td>1.162</td>
</tr>
</tbody>
</table>

Again, although these findings are presented in tables that are clustered according to linguistic category, they represent the results of a series of separate, univariate analyses to test for the main effects of these predictor variables within each group of students.
Here we see that selected response items are significantly more difficult for both emergent bilingual and native English speaking students; however, this relationship is more severe in the case of emergent bilingual students because the OR of the relationship is larger (0.288 for emergent bilinguals vs. 0.396 for non-emergent bilingual students).

Table 3.15 also demonstrates that -Item Length has a slightly positive relationship with students’ item-level scores for both groups, although the ORs for both of these main effects are quite small and do not vary much across groups.

An interesting finding from the comparisons of ORs here, is that the -Distance Picture to Question variable actually has a very slight positive relationship with students’ item-level scores of native English speaking students; whereas, -Distance Picture to Question has a negative relationship with emergent bilingual students, and this relationship has a moderately large effect size (OR = 0.842). This has interesting implications given the literature on the importance of visuals and possibilities of computerized tests for emergent bilingual students (e.g. Martiniello, 2008, Logan-Terry & Wright, 2010; Wright & Logan-Terry, in preparation), which is further discussed in Section 3.4 below.

<table>
<thead>
<tr>
<th>Linguistic Characteristic (Predictor Variable) for Each ESOL Group</th>
<th>B (SE)</th>
<th>P</th>
<th>Lower Bound</th>
<th>Odds Ratio</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>-Sentence Length Emergent Bilinguals</td>
<td>0.178</td>
<td>.000</td>
<td>1.152</td>
<td>1.195</td>
<td>1.240</td>
</tr>
<tr>
<td>Native</td>
<td>0.262</td>
<td>.000</td>
<td>1.254</td>
<td>1.299</td>
<td>1.345</td>
</tr>
</tbody>
</table>

Table 3.16. Differences between Odds Ratios Across ESOL Groups for Significant Syntax Predictor Variables
Table 3.16 demonstrates that *Sentence Length* actually has a slightly positive relationship with students’ item-level scores for both emergent bilingual and non-emergent bilingual students. This is somewhat surprising given past literature, but is discussed further in Section 3.4 below.

The more interesting finding in Table 3.16 is that *Words before Main Verb* does not have a significant relationship with native English speaking students’ item level scores; however, it has a relatively strong (OR = 0.609) negative relationship with emergent bilingual students’ scores. In other words, this syntactic complexity measure seems to have a more negative relationship with emergent bilingual students’ scores and may thus be causing items to be biased against this group of students. This is a significant finding for my overarching research question about the sources of achievement gaps in the SCALE-uP corpus and is discussed again in the implications of Section 3.4.
Table 3.17. Differences between Odds Ratios Across ESOL Groups for Significant Lexis Predictor Variables

<table>
<thead>
<tr>
<th>Linguistic Characteristic (Predictor Variable) for Each ESOL Group</th>
<th>B (SE)</th>
<th>P</th>
<th>Lower Bound</th>
<th>Odds Ratio</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>-Verb Hypernymy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emergent Bilinguals</td>
<td>-0.607 (0.130)</td>
<td>.000</td>
<td>0.422</td>
<td><strong>0.545</strong></td>
<td>0.703</td>
</tr>
<tr>
<td>Native English Speaking Students</td>
<td>-1.110 (0.121)</td>
<td>.000</td>
<td>0.260</td>
<td><strong>0.329</strong></td>
<td>0.418</td>
</tr>
<tr>
<td>+Concreteness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emergent Bilinguals</td>
<td>0.001 (.462)</td>
<td>.998</td>
<td>1.001</td>
<td><strong>1.009</strong></td>
<td>1.004</td>
</tr>
<tr>
<td>Native English Speaking Students</td>
<td>0.009 (0.030)</td>
<td>.000</td>
<td>1.006</td>
<td><strong>1.009</strong></td>
<td>1.011</td>
</tr>
</tbody>
</table>

In terms of the findings for differences between ORs for the Lexis variables, there are some surprising findings here. -Verb Hypernymy has the predicted negative relationship with both ESOL groups; however, the relationship is more extreme in the case of native English speaking students, which is the opposite of what I predicted based on the readability literature. This means that the items with a large number of verbs with multiple meanings are more difficult for both groups, but the monolingual students are more negatively affected by this aspect. As for the +Concreteness (mean concreteness value of all content words in the item prompt that match a word in the MRC database) variable, the main effect is not significant for emergent bilingual students (p = .462) and, though it is significant for native English Speaking Students students, the relationship has a negligible effect size and thus does not predict difficulty or item bias (OR = 1.009).
Table 3.18. Differences between Odds Ratios Across ESOL Groups for Significant Discourse Predictor Variables

<table>
<thead>
<tr>
<th>Linguistic Characteristic (Predictor Variable) for Each ESOL Group</th>
<th>B (SE)</th>
<th>P</th>
<th>Lower Bound</th>
<th>Odds Ratio</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>+Connectives Emergent Bilinguals</td>
<td>-0.056 (0.003)</td>
<td>.000</td>
<td>0.940</td>
<td>0.946</td>
<td>0.951</td>
</tr>
<tr>
<td>Native English Speaking Students</td>
<td>-0.046 (0.003)</td>
<td>.000</td>
<td>0.950</td>
<td>0.955</td>
<td>0.960</td>
</tr>
<tr>
<td>+OrdinalSyntactic Repetition Emergent Bilinguals</td>
<td>0.179 (0.046)</td>
<td>.000</td>
<td>1.093</td>
<td>1.196</td>
<td>1.309</td>
</tr>
<tr>
<td>Native English Speaking Students</td>
<td>-0.041 (0.045)</td>
<td>(.360)</td>
<td>0.879</td>
<td>0.960</td>
<td>1.048</td>
</tr>
<tr>
<td>+Intentional Cohesion Emergent Bilinguals</td>
<td>0.011 (0.004)</td>
<td>.018</td>
<td>1.002</td>
<td>1.011</td>
<td>1.020</td>
</tr>
<tr>
<td>Native English Speaking Students</td>
<td>0.035 (0.004)</td>
<td>.000</td>
<td>1.027</td>
<td>1.035</td>
<td>1.044</td>
</tr>
</tbody>
</table>

Table 3.18 demonstrates a very interesting pattern in terms of differences between significance and ORs across groups for the +OrdinalSyntactic Repetition variable. The variable is not significant as a main effect for native English speaking students ($p = .360$); however, the relationship is significant and positive for the emergent bilingual group. Although the effect size is small (OR = 1.196), this provides some limited evidence that syntactic repetition in item prompts could be useful in reducing achievement gaps for emergent bilingual students. Further implications of this finding are discussed in Section 3.4 below.

The other findings for the discourse variables are less interesting; the +Connectives variable demonstrates significant negative relationships with students’
item-level scores for both ESOL groups, although the effect sizes are negligible.
Similarly, +Intentional Cohesion demonstrates the predicted significant positive relationship with both ESOL groups’ item-level scores; however, the ORs for these relationships are relatively small.

Section 3.3.3.3. Summary of Findings for Research Question Three: Does the Relationship between Linguistic Characteristics of Test Items and Students’ Item-Level Scores Vary across ESOL Groups?

In sum, my findings for this research question both confirm and refute some of my hypotheses. There are significant interactions between 10 linguistic characteristics variables and the ESOL Group Status variable: -Selected or Constructed*ESOL Group Interaction (p = .003); -Item Length (# words)*ESOL Group Interaction (p = .000); -Distance Picture to Question*ESOL Group Interaction (p = .000); -Sentence Length (# words)*ESOL Group Interaction (p = .001); -Words before Main Verb*ESOL Group Interaction (p = .000); -Verb Hypernymy*ESOL Group Interaction (p = .004); +Concreteness*ESOL Group Interaction (p = .000); +Connectives*ESOL Group Interaction (p = .027); +OrdinalSyntactic Repetition*ESOL Group Interaction (p = .001); +Intentional Cohesion*ESOL Group Interaction (p = .000).

However, upon comparison of the difference in ORs for the main effects of each of the linguistic characteristic variables across ESOL groups, it is clear that only
some of these variables actually have interesting differences between terms of the differences in the slopes of the relationships with students’ item-level scores.

For example, -Selected or Constructed Response has a significantly more negative relationship with the item-level scores of emergent bilingual students than the native English speaking students. Similarly, -Words before Main Verb is predictive of a significant negative relationship with emergent bilingual students’ scores which is not true of the native English speaking students. In other words, this syntactic complexity measure seems to have a more negative relationship with emergent bilingual students’ scores and may thus be causing items to be biased against this group of students.

In contrast, the discourse-level variable +OrdinalSyntactic Repetition also does not have a significant main effect for native English speaking students, but does have a significant positive relationship with the item-level scores of the emergent bilingual students. Although the effect size is relatively small (OR = 1.196), this provides some limited evidence that syntactic repetition in item prompts could be useful in reducing achievement gaps for emergent bilingual students. Surprisingly, some of the linguistic characteristics of item prompts that have been argued to be most salient in past literature (e.g. lexical frequency and content word overlap, Crossley et al., 2008) do not demonstrate significant differences across groups.

Thus, I would argue that these findings for Research Question Three generally lend some additional evidence to the argument that the linguistic complexity of test items penalizes emergent bilingual students (e.g. Abedi, 2006; Martiniello, 2008; Wolf & Leon, 2009); however, some types of discourse-level
"complexity" may be important for emergent bilingual students to process item prompts. In addition, much of the variation in students' scores is not accounted for by the linguistic complexity of test items (e.g. -Target Idea accounted for some of the variation) and the implications of this are further discussed in the next Section 3.4.

Finally, Table 3.19 provides an overview and summary of the findings for research question three (Does the relationship between linguistic characteristics of test items and students' item-level scores vary across ESOL groups?).

Table 3.19. Summary of Findings for Research Question Three

<table>
<thead>
<tr>
<th>Linguistic Area of Predictor Variable</th>
<th>Linguistic Characteristic (Predictor Variable)</th>
<th>Definition/Gloss</th>
<th>Interaction, Effect Size, and Directionality Comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
<td>Selected or Constructed</td>
<td>Whether the item is selected among three multiple choice items or requires a constructed response from a student</td>
<td>(p = .003) Emergent Bilingual OR: 0.288 Native English OR: 0.396</td>
</tr>
<tr>
<td></td>
<td>Item Length (# words)</td>
<td>Number of words in the item prompt</td>
<td>(p = .000) Emergent Bilingual OR: 1.022 Native English OR: 1.038</td>
</tr>
<tr>
<td></td>
<td>Distance Picture to Question</td>
<td>Amount of distance (measured in terms of other questions) between the picture association with an item and the eventual item question</td>
<td>(p = .000) Emergent Bilingual OR: 0.842 Native English OR: 1.095</td>
</tr>
<tr>
<td>Syntax</td>
<td>Sentence Length</td>
<td>Mean number of words per sentence in the item prompt</td>
<td>(p = .001) Emergent Bilingual OR: 1.195 Native English OR: 1.299</td>
</tr>
<tr>
<td></td>
<td>Words before Main Verb</td>
<td>Mean number of words before the main verb of the main clause in sentences in the item prompt—argued to be taxing on working memory</td>
<td>(p = .000) Emergent Bilingual OR: 0.687 Native English OR: 1.030</td>
</tr>
<tr>
<td>Lexis</td>
<td>-Verb Hypernymy</td>
<td>Mean hypernym (multiplicity of meaning) value of main verbs in the item prompt—argued to contribute to text difficulty</td>
<td>(p = .004) Emergent Bilingual OR: 0.545 Native English OR: 0.329</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>+Concreteness</td>
<td>Mean concreteness value of all content words in the item prompt that match a word in the MRC database</td>
<td>(p = .000) Emergent Bilingual OR: 1.001 Native English OR: 1.009</td>
</tr>
<tr>
<td>Discourse</td>
<td>+Connectives</td>
<td>Incidence of all connectives in the item prompt (e.g. positive: and, after, because; negative: but, until, although; additives: also, moreover, however, but; causal: because, so, consequently, although, nevertheless; logical: or, actually, if; temporal: after, before, when, until)</td>
<td>(p = .027) Emergent Bilingual OR: 0.946 Native English OR: 0.955</td>
</tr>
<tr>
<td></td>
<td>+Syntactic Repetition</td>
<td>Proportion of intersection tree nodes between all adjacent sentences (algorithms build an intersection tree between two syntactic trees, one for each of the two sentences being compared)—a measure of cohesion</td>
<td>(p = .001) Emergent Bilingual OR: 1.196 Native English OR: 0.960</td>
</tr>
<tr>
<td></td>
<td>+Intentional Cohesion</td>
<td>Ratio of intentional particles (e.g., in order to, so that, for the purpose of, by means of, by, wanted to) to the incidence of causal content (intentional actions/events)</td>
<td>(p = .000) Emergent Bilingual OR: 1.011 Native English OR: 1.035</td>
</tr>
</tbody>
</table>
3.4 CONCLUSION

3.4.1. Summary

In sum, I find that some linguistic characteristics of test items (format, syntax, lexis, and discourse) are important predictors of all students' item-level scores, and some of these relationships vary significantly across ESOL groups (e.g. the format variable of -Selected or Constructed Response, the syntax variable of -Words before Main Verb, and the discourse variable of +OrdinalSyntactic Repetition). In other words, linguistic characteristics of test items can in some cases differentially affect test outcomes for native English speaking and emergent bilingual students. This finding provides some limited evidence that native English speaking content tests do conflate the assessment of English language proficiency and content knowledge, thereby exacerbating achievement gaps for emergent bilingual students.

However, this study also indicates that variation in students’ scores is also due to factors other than the linguistic characteristics of test item prompts (as evidenced by the non-significant results found for many of the linguistic variables tested in these analyses – only 10 of the 20 variables tested had significantly different relationships with students’ item-level scores across the two ESOL groups.)
3.4.2. Discussion and Implications

3.4.2.1. Research Question One: Are Linguistic Characteristics of Test Items Related to Students' Item-Level Scores?

In this study, linguistic characteristics of test items related to format (\(-\text{Selected or Constructed Response Type} \text{ and } \text{-Distance Text Stem to Prompt}\)), syntax (\(-\text{OrdinalLogical Operators} \text{ and } \text{-OrdinalModifiers per Noun Phrase}\)), lexis (\(-\text{Verb Hypernymy}\)), and discourse (\(+\text{OrdinalContent Word Repetition}\)) are found to be more predictive of students' outcomes on tests than were the underlying target ideas of each item. This finding is significant because it highlights the degree to which these "content" tests are actually tests of both content and language for all students (reading comprehension tests, arguably). Therefore, test developers should attempt to use the most accessible linguistic characteristics of test items, even when constructing tests for native English speaking students.

While the majority of the linguistic complexity measures are found to have a negative relationship with students' item-level scores, \(+\text{OrdinalContent Word Repetition}\) lap—a measure of cohesion—demonstrates a positive relationship with students' correct answers. This finding highlights the limitations of assuming that linguistic complexity leads to difficulty for all students and has implications for the operationalization of linguistic characteristics of test items in future studies. This study shows that the Coh-Metrix tool is particularly promising for addressing this...
issue because it offers a range of measurements that take into account insights from readability and processing literature.

However, some of the variables that emerge as significant with the biggest effect sizes in the univariate and multivariate analyses are not those with the strongest support in previous studies and some variables with strong support from past literature are not predictive in the expected direction (e.g. +LogLexical Frequency and +SqRtTemporal Cohesion); thus, further work on improving operationalization of linguistic characteristics of test items is needed.

3.4.2.2. Research Question Two: Are There Differences between Students’ Item-Level Scores across ESOL Groups?

That there are achievement gaps between emergent bilingual students and their native English speaking peers is not really an interesting finding, given the national backdrop and the findings from the backdrop Chapter 2 of this dissertation; however, combined with the findings from Research Question One above, the fact that students’ item-level scores do differ across ESOL groups, with emergent bilingual students having significantly fewer correct responses than their native English speaking peers helps to motivate Research Question Three below, which investigates the interactions and comparisons of ORs across groups—speaking to the sources of achievement gaps that is the primary issue of concern for this study and the larger dissertation effort.
3.4.2.3. Research Question Three: Does the Relationship between Linguistic Characteristics of Test Items and Students’ Item-Level Scores Vary across ESOL Groups?

The most important finding of this study is that certain linguistic characteristics of test items disadvantage emergent bilingual students while other linguistic characteristics of test items reduce achievement gaps for emergent bilingual students. For example, constructed response item format and syntactic complexity of item prompts are related to increased differences between achievement between emergent bilingual student and native English speaking students. In other words, asking emergent bilingual students to produce written language causes a significantly increased difference in their outcomes as compared with native English speaking students. I further explore this key finding in the subsequent Chapter 4 of this dissertation, a qualitative analysis of students’ constructed responses.

In addition, increasing amounts of syntactic complexity, as measured by the number of words before the main verb in the sentence, in item prompts are found to be related to increases in achievement gaps between emergent bilingual and native English speaking students. While the words-before-main-verb syntactic complexity variable emerges as significant, it is important to note that the other three syntactic complexity variables that I investigate in this study are not found to be significant. Thus, I would argue that we need much more research into which aspects of syntactic complexity are important for emergent bilingual students in order to
improve upon existing simplification-based test accommodation approaches, many of which are not grounded in actual research as to the effects of complexity—or lack thereof—on students’ scores. These findings of non-significant relationships for certain syntactic variables notwithstanding, the evidence I find related to increasing achievement gaps due to increasingly complex linguistic characteristics of test items (e.g. number of words before the main verb in the sentence) does provide some limited evidence that these tests conflate the assessment of content knowledge and language proficiency.

In addition, the -Distance Picture to Question proves interesting in that it has a negative relationship with the scores of emergent bilingual students but not for native English speaking students. This provides additional evidence of the importance of visuals for emergent bilingual students and the promise of computerized/multimodal testing options for emergent bilingual students (e.g. Martiniello, 2008, Logan-terry & Wright, 2010; Wright & Logan-Terry, in preparation).

In contrast, the discourse-level variable +OrdinalSyntactic Repetition has a significant positive relationship with the item-level scores of the emergent bilingual students. Although the effect size is small (OR = 1.196), this provides some limited evidence that syntactic repetition in item prompts could be useful in reducing achievement gaps for emergent bilingual students. More research is needed to see whether this pattern bears out in additional data sets, as it could have important implications for how to accommodate tests in order to make them more accessible for emergent bilingual students.
While the significant results for the four variables discussed above (constructed vs. selected response type, syntactic complexity in terms of words before main verb, distance from visual accompaniment, and syntactic repetition) are important, it is also important and interesting to note that the other 16 variables tested in this analysis yield non-significant results with regard to differential scores between ESOL groups. Surprisingly, some of the linguistic characteristics of item prompts that have been argued to be most salient in past literature (e.g. lexical frequency and content word overlap, Crossley et al., 2008) do not demonstrate significant differences across groups.

In this way, much of the variation in students’ scores does not seem to be accounted for by the linguistic complexity measures of test items employed in this study, and this issue needs further exploration. This is particularly important in light of arguments other researchers have made about variation in test scores that can be accounted for by other factors such as differences in underlying content knowledge, socioeconomic status, and the construct relevance of item prompt language (Abedi & Gándara, 2006; Bachman & Koenig, 2004; Elliot, 2008; Farnsworth, 2008, in preparation; Ockey, 2007).

In sum, I argue that the findings from the analytic probe in this chapter lend some limited additional evidence to the argument that the linguistic complexity of test items penalizes emergent bilingual students (e.g. Abedi, 2006; Martiniello, 2008; Wolf & Leon, 2009).
3.4.2.4. Further Significance of Study Findings

These findings about linguistic characteristics of test items that are predictive of increased or decreased achievement gaps have potential implications for the growing field of research related to test accommodations for emergent bilingual students (e.g. Hudicourt-Barnes, Noble, Roseberry, Suarez, & Warren, 2008; Kopriva, 2008; Rivera & Collum, 2006; Wolff, Herman; see also Kieffer, Lesaux, Rivera, & Francis, 2009 for an extensive meta-analysis of research in this area). For example, linguistic simplification—one of the most popular accommodations for emergent bilingual students—may benefit from decreased syntactic complexity as measured in this study (words before the main verb). However, simplification-based accommodations must also take into account the importance of some elements of text for potentially leveling the playing field (e.g. syntactic repetition) by creating cohesion and comprehension for all students. Also, the finding that increased distance from the test item visual to the test item question disadvantages emergent bilingual students is important as it speaks to the importance of visuals for emergent bilingual students and the possibility of computerized, multimodal testing as an accommodation for emergent bilingual students (e.g. Martiniello, 2008, Logan-Terry & Wright, 2010; Wright & Logan-Terry, in preparation).

On a methodological level, this study also has great significance as it highlights the promise of incorporating readability literature and especially computerized readability tools such as Coh-Metrix, as mentioned above. In addition,
the innovative use of a longitudinal MLM statistical approach taken here has many advantages over those taken in other item characteristics studies and could be fruitfully applied to large-scale testing data. Using MLM instead of DIF or Rasch or CFA would allow researchers to take into account both fixed and random effects for multiple levels of variables (e.g. both item-level as well as student-level variables), thereby avoiding the assumption of independence of observations via calculations of the random effect blocks (Fields, 2009; see also Hayes, 2006; Heck, Thomas, & Tabata, 2010; Singer & Willlet, 2003 for additional information on the advantages of MLM).

Finally, the implications of these findings for my overarching dissertation research questions (Are there test-related sources of achievement gaps for emergent bilingual students?) are significant. I find that some linguistic characteristics of test items do contribute to achievement gaps for emergent bilinguals students, although not always in ways that are predicted by previous research and not necessarily in ways that are easy to turn into practical recommendations. I do argue that native English speaking content tests conflate students’ language proficiency with their content knowledge to a certain extent, and serve as gatekeeping encounters for emergent bilingual students (see also the discussion of gatekeeping encounters in Chapter 1, Chapter 4, and conclusion Chapter 5). However, much additional research is needed to uncover other factors that account for variation in students’ item-level scores across ESOL groups.
3.4.3. Limitations and Suggestions for Future Research

Some of the surprising results in this study may be due to limitations in variable operationalization and analysis. For instance, the lack of the predicted relationship between lexical frequency and student outcomes may be due to the fact that Coh-Metrix calculates this variable based on the Celex corpus, which is antiquated (texts from 1991) and based on adult texts. As discussed above, it may be more appropriate to measure lexical frequency in a way that is more context-specific, such as using a middle-school science corpus or aggregated terminology from textbooks and/or classroom discussion. Furthermore, the Coh-Metrix tool is designed to measure the readability of longer passages than the 60-70 word item prompts in the present study, so future studies may want to explore other readability measures in operationalizing linguistic characteristics of test items. Thus, some of the measures I use as variables in the present study may not be as reliable as if they were measured using a computational readability tool specifically designed for shorter texts.

The use of an existing test data corpus also proves limiting in several ways. For example, the QAM test only features ten total items, which very much reduces the generalizability of my findings to other contexts. Furthermore, I do not include visual accompaniments to items in the MLM analyses because all items in the QAM tests have similar visual elements. In addition, the existing data set does not allow me to explore the influence of other important predictor variables such as students’ English language proficiency (the only measure available was whether or not they
were receiving ESOL services), length of time living in the United States, and family language policy.

It is also important to note that some of the “native English speaking” students in my study may in fact speak an additional language at home; I categorize them as native English speaking based on the fact that they have never received ESOL services as this is the only information available to me in the SCALE-uP corpus. Some literature shows that bilingual “generation 1.5” students do perform like monolingual native speaker students in literacy assessments (e.g. di Gennaro, 2009); however, it would be good to investigate generation 1.5 students as well as those students previously receiving ESOL services in future studies of linguistic characteristics of test items and student outcomes. It would also be very useful to test the generalizability of these item characteristics findings to large-scale, standardized, high-stakes content tests.

The study also has several limitations due to the depth of my analyses. Additional qualitative analysis of test item prompts would be beneficial in shedding further light on the interesting quantitative results discussed above. In addition, analysis of the actual constructed responses given by emergent bilingual and native English speaking students is an important next step, as selected versus constructed response type is a key item characteristic variable in the present study. Thus, the issue of why and how students’ written answers are scored as “incorrect” is dealt with in subsequent qualitative analytic Chapter 4 of the dissertation.

Finally, this study does not address what I believe to be an underlying issue of monolingual content assessments for emergent bilingual students: not allowing
students to utilize their entire linguistic repertoire in demonstrating content knowledge. Even if emergent bilingual students have underlying content knowledge, I argue that they will be met with issues in demonstrating that knowledge on monolingual tests. Recent studies investigating multilingual, multimodal assessments and employing introspective methodologies (e.g. Kopriva, 2008; Logan-Terry & Wright, 2010; Schissel, 2011; Solano-Flores & Li, 2009; Rea-Dickins, 2011; Wright & Logan-Terry, in preparation) are promising in addressing this underlying issue. I see the present analysis of monolingual content tests as complementary to the approach of studying multilingual, multimodal assessments for emergent bilingual students.
CHAPTER 4

STUDENTS' CONSTRUCTED RESPONSES:
An Interactional Approach to the Study of Achievement Gaps

The present qualitative analysis further investigates an issue that arose in
the previous quantitative analysis found in Chapter 3: increased achievement gaps
between emergent bilingual students and native English speaking students on
constructed response items. As mentioned in the previous chapter, I found that both
emergent bilingual students and native English speaking students had significantly
lower scores on constructed response items than on selected response items.
However, the slope of the constructed response variable–item score relationship
was significantly more negative for emergent bilingual students. In other words,
asking emergent bilingual students to produce written language increased the
achievement gaps between them and their native English speaking peers. Thus, I
begin the present analysis with the general research question, “Why are constructed
response items so problematic for emergent bilingual students as compared with
native English speaking students?”

Employing analytic tools from Interactional Sociolinguistics (Gumperz, 1982,
1992; 1999; Schiffrin, 1994; Tannen, 2005) and schema theory (Anderson & Pichert,
1978; Pearson, 1985; Rowe, 1986; Rowe & Rayford, 1987; Ruddell & Unrau, 2004), I
show how the knowledge schemas (Tannen & Wallat, 1993) of each of the
interlocutors involved in the test interactions—test developers, students, and
raters—play a crucial role in the construction of meaning during the test taking
process. My findings also illustrate patterns of schema activation miscues related to
the sequence of items, specific lexical aspects of item prompts, grammatical
constructions in item prompts, and visual accompaniments. I also show how these
schema activation miscues were more of a problem for emergent bilingual students
than for native English speaking students.

In the following introduction to the chapter (Section 4.1), I identify and
discuss a key pattern that emerges in the data: students’ constructed responses
often engaged with ideas other than the ones targeted by the QAM items, especially
in the case of emergent bilingual students. In this introductory Section 4.1 I also
present a qualitative analytic framework that I use to illuminate how this pattern is
discursively constructed in the data. The next section of this chapter (Section 4.2)
characterizes the data—the QAM test items and the classroom of students whose
tests I analyze. In the analysis Section 4.3 I present a series of examples that
demonstrate the types of schema activation miscues that occur as well as ways in
which these schema activation miscues are overcome. Finally, the conclusion
(Section 4.4) provides summary along with discussion, implications, limitations, and
suggestions for future research.
4.1. INTRODUCTION

In approaching this analysis, I apply a variety of qualitative frameworks to examine students’ data from the QAM test. I also consider some ethnographic data (classroom video of the implementation of the curriculum unit, students' workbooks, interviews with the students and the teacher) from one particular classroom of students, a classroom with the pseudonym of “Eastmond.” These types of ethnographic data are collected from six sample classrooms as part of the original SCALE-uP study.

I choose the Eastmond classroom to analyze for this chapter because it consists of the highest proportion of ESOL students—12 currently receiving services, 15 who have never received services, and one student who has previously received ESOL services. This high proportion of students currently receiving ESOL services (12 out of 28 students, or 43 percent of the classroom) does not exist in any of the other classrooms sampled by the original SCALE-uP researchers. In this way, Eastmond stands out to me as uniquely well suited for comparison of emergent bilingual students’ and native English speaking students’ constructed responses. As with the previous analyses, I focus this analysis on the students who are currently receiving ESOL services (“emergent bilingual students”) and those who have never received services (“native English speaking students”) in order to have the clearest representation of groups that demonstrate achievement gaps. Thus, the test data from Yolanda, the one student who has exited ESOL services, are not considered.
As an initial starting point in investigating why constructed response items are so problematic for emergent bilingual students as compared with native English speaking students, I consider whether the emergent bilingual students’ constructed responses might simply be rated as wrong due to second language (L2) writing issues. However, I find that this is not the case. While there is ample evidence of emergent bilinguals’ interlanguage in their written responses, higher numbers of these language issues do not seem to be related to more frequent ratings of “incorrect” by the raters.\textsuperscript{32}

As I examine and compare the constructed responses of the Eastmond emergent bilingual and native English speaking students, a different, more salient pattern emerges: students’ constructed responses often address ideas other than the ones targeted by the specific test item.\textsuperscript{33} For example, consider the QAM Item #9 ("Rolling Ball Part I") found in Figure 4.1:

\textsuperscript{32} Please see Section 4.3.4 \textit{Example 5} and \textit{Example 6}, which show how emergent bilingual students’ \textit{correct} answers feature L2 writing problems.

\textsuperscript{33} I use the QAM Assessment Manual found in Appendices B and C for this information.
The correct answer to the constructed portion of this item is “no” and students are meant to provide a constructed response that engages with the rule of inertia—an object at rest stays that way unless acted on by a force and an object in motion will continue to move unabated unless acted on by a force. However, students often select “yes” and then wrote in their constructed responses about ideas other than inertia, such as gravity, wind, or other ideas related to the foundational QAM target idea that motion is caused by forces.

This pattern of the ideas provided in students’ constructed responses differing from the intended target of the item seems particularly strong in emergent bilingual students’ tests. I consider whether this pattern might account for the statistically significant increased achievement gaps between emergent bilingual
students and native English speaking students identified in Chapter 3, and this serves as the initial motivation for this qualitative analysis.

In order to further investigate whether this pattern might shed light on the sources of achievement gaps across the ESOL groups, I count each time a student's constructed response addresses a different idea than the one targeted by the QAM item. The following Table 4.1 and Table 4.2 show an overview of the students' constructed responses from each group (emergent bilingual and native English speaking, respectively) that relate to a different target idea than the one intended by the test developers. In each case that is flagged (marked with a black dot in the table), the students receive no credit for the item.
Table 4.1. Emergent Bilingual Students’ Constructed Responses that Addressed a Different Idea than the Target Idea of the Item

<table>
<thead>
<tr>
<th>Emergent Bilingual Students</th>
<th>QAM #2 Pushing a Couch</th>
<th>QAM #3 A Full Truck</th>
<th>QAM #5 Gravity &amp; Tennis Ball Part I.b.</th>
<th>QAM #6 Gravity &amp; Tennis Ball Part II.a.</th>
<th>QAM #7 Gravity &amp; Tennis Ball Part II.b.</th>
<th>QAM #9 Rolling Ball Part I.</th>
<th>QAM #10 Rolling Ball Part II.</th>
<th>Student Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ralph</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Paul</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Dylan</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Greg</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Ethan</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Valerie</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Hank</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Michelle</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Henry</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Faith</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Walter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Zoe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Item Totals</td>
<td>1/</td>
<td>2/</td>
<td>2/</td>
<td>7/</td>
<td>4/</td>
<td>10/</td>
<td>3/</td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2. Native English Speaking Students’ Constructed Responses that Addressed a Different Idea than the Target Idea of the Item

<table>
<thead>
<tr>
<th>Native English Speaking Students</th>
<th>QAM #2 Pushing a Couch</th>
<th>QAM #3 A Full Truck</th>
<th>QAM #5 Gravity &amp; Tennis Ball Part I.b.</th>
<th>QAM #6 Gravity &amp; Tennis Ball Part II.a.</th>
<th>QAM #7 Gravity &amp; Tennis Ball Part II.b.</th>
<th>QAM #9 Rolling Ball Part I.</th>
<th>QAM #10 Rolling Ball Part II.</th>
<th>Student Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Andrew</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Carrie</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Ella</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Stacy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Beth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Nathan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Brad</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Ian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Quincy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Luke</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Derek</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Owen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Kyle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Joy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Item Totals</td>
<td>0/</td>
<td>1/</td>
<td>2/</td>
<td>5/</td>
<td>3/</td>
<td>7/</td>
<td>1/</td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>12</td>
<td>12</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

Student Avg. 2.4

Student Avg. 1.2
As shown in these overview tables, the different-than-target-idea pattern is stronger for emergent bilinguals than for native English speaking students. On average, emergent bilingual students’ answers are flagged twice as often as those of native English speaking students—emergent bilingual students average 2.4 mismatches on the seven constructed response items while native English speaking students average 1.2 mismatches across the seven constructed response items. Another way to emphasize the differences in this pattern across the ESOL groups is to focus on the respective total items flagged. Across all seven of the constructed response items, emergent bilingual students’ responses are flagged 29 out of 84 times (35% of their answers relate to a different target idea than the one intended). In contrast, native English speaking students’ responses are flagged only 18 out of 105 times (17% of their answers relate to a different target idea).

There are also differences across the ESOL groups related to the range of individual students’ mismatches across items: while emergent bilingual students’ mismatches range from zero out of seven items to five out of seven items, native English speaking students’ mismatches only range from zero out of seven items to three out of seven items. Thus, while the general pattern features more mismatches for emergent bilingual students than for native English speaking students, there are some emergent bilingual students whose constructed response items feature fewer mismatches than some of the native English speaking students.

It is also important to note that there are other incorrect answers that are incorrect for the right reasons (students address the targeted idea but do not show evidence of understanding). The incorrect answers flagged in the above table are a
subset of the total incorrect answers: the ones that are incorrect and in which the student addresses an idea other than the target idea of the item.

In fact, out of the 150 item interactions of the native English speaking students, 65 of them received a score of zero and a subset of 19 of those incorrect answers were flagged for mismatches in target idea. In other words, approximately 29% of the native English speaking students’ incorrect answers featured a mismatch in ideas targeted. This pattern was more extreme for emergent bilingual students: 53 of their 120 item interactions were incorrect, and 29 of those were flagged for mismatches (approximately 55%). This compelling as it indicates that over half of the emergent bilinguals incorrect answers fit within this pattern of mismatches of target ideas.

It is also evident in the above tables that the “Rolling Ball Part I” item is the most problematic for both groups (flagged 10 out of 12 times for the emergent bilingual students and seven out of 15 times for the native English speaking students). There are plenty of instances of students giving off-target responses across the other items in the test as well (e.g. “Gravity & Tennis Ball Part II.a” is flagged seven out of 12 times for emergent bilingual students and five out of 15 times for native English speaking students).

Indeed, the “Gravity and a Tennis Ball” and the “Rolling Ball” series of items seem to be particularly problematic for both emergent bilingual students and native English speaking students. For the five constructed responses in those item series, emergent bilingual students’ answers are flagged 26 out of 60 times (43% of their answers discuss an idea other than the one targeted by the item). The native English
speaking students’ responses are flagged 17 out of 75 times (23% of their answers are not about the target idea). Thus, I choose to focus qualitative analysis on those two series of items, as they demonstrate the clearest pattern.

Having identified the above broad pattern in the data (mismatches in intended target ideas and the ideas addressed in students’ constructed responses), my next analytic step is to conduct qualitative analysis of both the emergent bilingual and native English speaking students constructed responses to try to figure out how and why the pattern occurs. For these analyses, I draw on frameworks that have been shown to be particularly well suited to the study of miscommunication and mismatches in expectations. These analytic frameworks are outlined in the following Section 4.1.1.

4.1.1. Analytic Framework

Interactional Sociolinguistics (Gumperz, 1982, 1992, 1999; Schiffrin, 1994; Tannen, 2005) is a qualitative analytic framework that has been fruitfully applied to uncover how patterns are constructed interactionally across a range of contexts (e.g. Davies, 2003; Hamilton, 1994; Schiffrin, 1987). Anthropologist John Gumperz, one of the founders of Interactional Sociolinguistics, developed this approach as a means of addressing social justice issues surrounding minority populations’ interactions in crucial institutional contexts. In a similar spirit, the present analysis endeavors to better understand discursive construction of institutional social justice issues.
(educational achievement gaps) for the minority populations of emergent bilingual students.

Gumperz (1982, 1992) emphasizes that interlocutors may share grammatical knowledge yet have different ways of contextualizing language that are socially and culturally relative. In particular, Gumperz points to contextualization cues that interlocutors use to signal how talk is meant to be interpreted. These studies focus on face-to-face communication; thus, the contextualization cues described in this early work relates to aspects of language such as intonation, lexical choice, and speech rhythm. In the context of this chapter, I consider written contextualization cues that sometimes overlap with those found in face-to-face interaction (e.g. lexical choice, grammatical constructions) and sometimes differ (e.g. underlining written words for emphasis).

The work of Erving Goffman has also heavily influenced Interactional Sociolinguistics. His research highlights how contextualization cues can function to frame interaction (e.g. serious, joking, formal, informal, etc.). Framing is a crucial aspect of the testing process, as I will discuss in my analysis. Goffman (1971) also focuses on both the information given (which is intentional) and the information given off (unintentionally) during interaction. Much like in face-to-face interaction, written interaction also features information intentionally given as well as information that is given off unintentionally, and this is a factor in the constructed response interactions that I analyze for this chapter.

While the foundational work in Interactional Sociolinguistics and much of the research employing its methodologies are applied to face-to-face interactions (e.g.
Hamilton, 1994; Schiffrin, 1987; Tannen, 1989/2007; Tyler, 1994; Wright, 2008), I apply Interactional Sociolinguistics to the written interactions that students have during the test taking process. I argue that interactions with the Questions about Motion test items constitute a sequential, triadic participant structure (Philips, 1972), as seen in the following Figure 4.2:

*Figure 4.2. Sequential, Triadic Participant Structure of Test Interactions*

In each phase of the interaction, the background expectations and assumptions of each of the three interactional parties (test developers, students, and raters)—their knowledge schemas—play a crucial role in the construction of meaning. Knowledge schemas have been discussed and defined in various ways across different fields of literature (e.g. Bartlett, 1932/1967; Brown & Yule, 1983; Minsky, 1975; Rumelhart, 1975; Tyler, 1995). Here I use the term in line with Tannen & Wallat (1993)'s definition: “participants’ expectations about objects, people, settings, ways to interact, and anything else in the world [that are] continually checked against experience and revised” (ibid: 61). It is also important to note that the three-part test interaction (test developers, students, test raters) is
shaped not only by the interlocutors’ background knowledge, expectations, and assumptions related to physics content but also their respective structured expectations about testing as a communicative event.

Knowledge schemas have also been discussed in schema-theoretic literature in terms of the “transactional view” of reader and text (Rosenblatt, 1985). Goodman & Goodman (2004) highlight the similarities and parallels between the interactional nature of reading and that of face-to-face communication. Researchers working in the area of reading have pointed out that “much of the meaning understood from a text is really not actually in the text, per se, but in the reader, in the background knowledge or schematic knowledge of the reader” (Carrell & Eisterhold, 1983: 559). Indeed, studies have shown that readers can read a text and come away with very different interpretations, depending on their respective background knowledge and expectations (Anderson & Pichert, 1978; Pichert & Anderson, 1977; Ruddell & Unrau, 2004).

This is particularly salient in the testing context in which students are presented with a series of brief texts that may or may not be related to one another thematically. The ordering of the content becomes crucial as students work through items and experience waves of schema activation. In work related specifically to the assessment of reading comprehension, Rowe (1986) and Rowe and Rayford (1987) have shown that meaning is constructed as a result of the interaction between knowledge schemas and cues available in the text—I refer to these in the present chapter as written contextualization cues. In order to comprehend the series of texts presented in tests, students must make inferences about the content of the text in
the process of schema activation. Once a schema has been activated, the content of the text is filtered through the expectations and assumptions affiliated with that particular schema. Importantly, schemas are also shaped during this process (Tannen & Wallat, 1993).

For the present analysis of the QAM tests, I consider this schema activation along with the additional layers of interaction inherent in the test taking communicative event. These interactions do not only hinge on students’ reading and interpretation of item prompt text but also their construction of written responses that are then read and interpreted by raters. In both the test item text and the students’ constructed response text there are subtle contextualization cues that signal how the discourse is meant to be interpreted. These written contextualization cues—whether consciously or unconsciously employed—are an important aspect of how the text creators (both test developers and students) can frame their text as relating to certain key target ideas. I highlight in my analysis of the QAM data how each level of the interaction is shaped by variation in use of these contextualization cues, framing, and the interlocutors’ knowledge schemas.

4.2. DATA

The data under consideration for this analysis consist of the Eastmond classroom students’ QAM posttests. In the following Section 4.2.1, I first remind the reader of the nature of the QAM test items that were initially introduced and discussed in Chapter 1. I also discuss the target ideas of the “Gravity and a Tennis
Ball” and the “Rolling Ball” series of items, as they are the focus of the qualitative analysis.

Then, in Section 4.2.2 I give more detail about the Eastmond classroom and the students within the classroom. This serves to provide some context that will help the reader orient to the students’ test data that are discussed in the qualitative analyses in Section 4.3.

4.2.1. The QAM Test

In the QAM test, the “Gravity and a Tennis Ball” and the “Rolling Ball” series of items target ideas relate to motion and forces in a variety of ways. The foundational target idea (Idea #1, as referenced by the test developers) is: changes in speed or direction are caused by forces (“Gravity and Tennis Ball Part I.a,” “Gravity and Tennis Ball Part II.a,” and “Rolling Ball Part II”). A slightly more complex idea (Idea #2) is: the greater the force is, the greater the change in motion will be (“Gravity and a Tennis Ball Part II.b”). These series of items also target the highest level Idea #4 that an object at rest stays that way unless acted on by a force and an object in motion will continue to move unabated unless acted on by a force (the rule of inertia), through the “Rolling Ball Part I” item that we saw earlier in the introduction. In this way, the target ideas are not ordered sequentially on the test in terms of difficulty, nor are they necessarily clustered within certain item series. The below Figure 4.3 overviews the Target Ideas for each of the QAM items analyzed in this chapter.
As shown in the above figure, as students move through the test items, they are asked to engage with Target Idea #1 (changes in speed or direction are caused by forces, specifically as related to gravity). The next item (“Gravity and a Tennis Vall Part I.a”) also targets Idea #1 as related to gravity. “Gravity and a Tennis Ball Part II.b” then asks a question that is meant to target a different idea, Idea #2 (the greater the force is, the greater the change in motion will be). After this series about gravity and motion, students encounter items that are meant to target different ideas—“Rolling Ball Part I” targets Idea #4 (an object at rest stays that way unless acted on by a force and an object in motion will continue to move unabated unless acted on by a force). The same Target Idea #1 (changes in speed or direction are caused by forces) is addressed in the final “Rolling Ball Part II” item, except in this case the idea is meant to cue students’ understanding of the specific force of friction (rather than gravity).

I primarily focus my analyses on students’ posttest data related to the above QAM constructed response items as those were the data analyzed in the previous
dissertation chapter that helped motivate the current study (Chapter 3); however, I do also consider the students’ pretests in certain analyses to help identify patterns and to see whether they differed before and after the curriculum implementation.\(^{34}\)

4.2.2. Eastmond Classroom Students

As mentioned above, data from the Eastmond 6\(^{th}\) grade physics classroom are selected for the case study presented in this chapter. The Eastmond classroom consists of 28 students—12 emergent bilinguals (currently receiving ESOL services), 15 native English speaking students (never having received ESOL services), and one student who has previously received ESOL services (Yolanda).\(^{35}\) As mentioned above, I focus this analysis on the students who are currently receiving ESOL services and those who have never received services in order to have the clearest representation groups that demonstrate achievement gaps in larger studies (e.g. Chapter 2). Thus, Yolanda’s test data are not considered in this analysis.

As mentioned in the backdrop Chapter 2 of this dissertation, native English speaking students make significantly larger gains using the EMF physics curriculum than the emergent bilingual students (average gains for native English speaking students are approximately 17\% and the average gains for emergent bilingual

\(^{34}\) As a general overview, the pretest patterns are similar to those discussed below for the posttest. The schema activation miscues are somewhat ameliorated on the posttests, although they are still abundant, as I show in the findings Section 4.3.

\(^{35}\) As mentioned in previous chapters, there are many limitations with this classification of students; however, it is the only classification available to me within the existing SCALE-uP corpus.
students are only approximately 11%). The below Table 4.3 has an overview of the achievement gaps across the corpus as compared with the Eastmond classroom.

Table 4.3 Overview of Achievement Gaps Across the Corpus as Compared with the Eastmond Classroom

<table>
<thead>
<tr>
<th></th>
<th>Emergent Bilinguals</th>
<th>Native English Speaking Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus Pretest</td>
<td>31.4</td>
<td>43.5</td>
</tr>
<tr>
<td>Corpus Posttest</td>
<td>40.0</td>
<td>58.4</td>
</tr>
<tr>
<td><strong>Avg. Gains</strong></td>
<td><strong>8.6</strong></td>
<td><strong>14.9</strong></td>
</tr>
<tr>
<td>Eastmond Pretest</td>
<td>39.7</td>
<td>34.7</td>
</tr>
<tr>
<td>Eastmond Posttest</td>
<td>43.8</td>
<td>52.1</td>
</tr>
<tr>
<td><strong>Avg. Gains</strong></td>
<td><strong>4.1</strong></td>
<td><strong>17.4</strong></td>
</tr>
</tbody>
</table>

As shown in above table, the achievement gap trend is even more exaggerated in the Eastmond classroom than in the rest of the SCALE-uP corpus. Emergent bilingual students have relatively high pretest scores (approximately 40%), but their average gains are only approximately four percent. In contrast, the native English speaking students in the Eastmond classroom have relatively low scores on the pretest (an average of approximately 35%), with much bigger gains on the posttest (average gains for native English speaking students in Eastmond are approximately 17%). This makes Eastmond a particularly interesting classroom to study, as the achievement gap trend found across the corpus is particularly

---

36 These percentage scores are based on the weighting outlined in the QAM Assessment Manual in Appendix B that is also discussed in the introductory chapter of the dissertation. Here I have combined the comparison and treatment data in the “corpus” comparison numbers. In Chapter 2 these categories are parsed out and discussed in more detail, as those issues are more salient for the analysis in Chapter 2.
problematic in Eastmond, and the Eastmond classroom has a particularly large proportion of emergent bilingual students.

The Eastmond classroom curriculum implementation and Questions about Motion testing occurred in the spring of 2004 over the course of a two-month period. The teacher in this classroom, a teacher with the pseudonym “Mr. Bob,” is an experienced science teacher and has taught in the district for approximately 20 years. This is his first time using the EMF curriculum and it is the first time the curriculum has been used in the school district. According to the teacher, the seating arrangement of the classroom is organized into heterogeneous lab groups of approximately four students. The 28 students in the Eastmond science classroom are of different ethnic, socioeconomic, and linguistic backgrounds.

The Figure 4.4 below provides a seating chart of the entire Eastmond classroom, with the emergent bilingual students indicated with bolding and underlining (please see legend on bottom right of figure).
Figure 4.4. Eastmond Classroom Seating Chart

Teacher

Table 1
Andrew  Beth
Carrie   Derek

Table 2
Nathan   Owen
Paul     Ralph

Table 3
Ian      Quincy
Brad

Table 4
Michelle  Luke
Kyle     Joy

Table 5
Yolanda *
Hank     Zoe

Table 6
Ella     Greg
Henry    Faith

Table 7
Stacy    Ethan
Valerie  Walter

Table 8
Adam     Dylan

Emergent Bilingual
* Exiled ESOL Services
Native English-Speaking
In order to provide additional information about the demographics for each of the Eastmond students, I provide summary Table 4.4 and Table 4.5 of the ethnic and socioeconomic details that were available to me from the SCALE-uP corpus.

Table 4.4. Emergent Bilingual Students’ Reported Demographic Information

<table>
<thead>
<tr>
<th>Demographic Information</th>
<th>Counts</th>
<th>Student Pseudonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>4/12</td>
<td>Faith, Michelle, Valerie, Zoe</td>
</tr>
<tr>
<td>Male</td>
<td>8/12</td>
<td>Walter, Dylan, Greg, Ralph, Paul, Henry, Ethan, Hank</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>5/12</td>
<td>Paul, Faith, Michelle, Valerie, Zoe</td>
</tr>
<tr>
<td>African American</td>
<td>0/12</td>
<td>N/A</td>
</tr>
<tr>
<td>Hispanic</td>
<td>2/12</td>
<td>Greg, Ralph</td>
</tr>
<tr>
<td>Asian</td>
<td>5/12</td>
<td>Dylan, Ethan, Henry, Walter, Hank</td>
</tr>
<tr>
<td>Free-and-Reduced Meals Services (a proxy for socioeconomic status)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never received</td>
<td>8/12</td>
<td>Dylan, Ethan, Henry, Paul, Faith, Michelle, Valerie, Zoe</td>
</tr>
<tr>
<td>Prior received</td>
<td>0/12</td>
<td>N/A</td>
</tr>
<tr>
<td>Currently receiving</td>
<td>4/12</td>
<td>Walter, Hank, Greg, Ralph</td>
</tr>
</tbody>
</table>

These demographic data differ from the national trends for emergent bilingual students. Whereas on a national level, emergent bilingual students tend to include a large proportion of Hispanic students from lower socioeconomic backgrounds, the Eastmond students do not fit this trend. The Eastmond classroom includes a number of Asian students, and the emergent bilingual students do not have a high proportion of students receiving Free-and-Reduced Meals Services (FARMS, the SCALE-uP proxy for socioeconomic measures). Those students who do
receive FARMS are the two Hispanic students, and two South Asian students (other non-FARMS Asian students tend to be from East Asia). The Caucasian emergent bilingual students include immigrants and adoptees from Eastern European countries, as well as some students of Hispanic heritage who self-identify as “white” in their interviews.

### 4.5. Native English Speaking Students’ Reported Demographic Information

<table>
<thead>
<tr>
<th>Demographic Information</th>
<th>Counts</th>
<th>Student Pseudonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>5/15</td>
<td>Carrie, Ella, Beth, Stacy, Joy</td>
</tr>
<tr>
<td>Male</td>
<td>10/15</td>
<td>Nathan, Owen, Brad, Luke, Derek, Andrew, Quincy, Ian, Adam, Kyle</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>8/15</td>
<td>Nathan, Owen, Luke, Derek, Beth, Joy, Stacy, Kyle</td>
</tr>
<tr>
<td>African American</td>
<td>6/15</td>
<td>Andrew, Brad, Quincy, Ian, Ella, Adam</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1/15</td>
<td>Carrie</td>
</tr>
<tr>
<td>Asian</td>
<td>0/15</td>
<td>N/A</td>
</tr>
<tr>
<td>Free-and-Reduced Meals Services</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a proxy for socioeconomic status)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never received</td>
<td>10/15</td>
<td>Andrew, Quincy, Ella, Nathan, Owen, Luke, Derek, Beth, Joy, Kyle</td>
</tr>
<tr>
<td>Prior received</td>
<td>0/15</td>
<td>N/A</td>
</tr>
<tr>
<td>Currently receiving</td>
<td>5/15</td>
<td>Brad, Ian, Stacy, Carrie, Adam</td>
</tr>
</tbody>
</table>

As with the emergent bilingual students in the Eastmond classroom, the native English speaking students include a large proportion of male students. In contrast to the emergent bilingual group, the native English-speakers include
students who self-identify as African American. While the ratio of students receiving FARMS is the same for native English speaking students as for emergent bilinguals, a large proportion of the FARMS students are those who self-identify as African American. This does fit with the national trends, although much of the Eastmond demographic data are quite different than in other contexts.

All of the above discussion echoes the issues of representativeness of these data from Montgomery County, Maryland. As I discuss in other chapters of this dissertation, none of my findings are meant to be generalizable to other contexts. These studies serve as analytic probes (testing out innovative methodologies that could then be applied to a larger sample) and simply are intended to reach a better understanding of the specific test-related issues that contribute to the achievement gap trend in the SCALE-uP corpus in particular. In the following analytic Section 4.3 I illustrate discursive patterns in students’ constructed responses that seem to contribute to the achievement gaps between emergent bilingual and native English speaking students in the Eastmond classroom.

4.3. ANALYSIS

In the subsequent analytic Sections 4.3.1-4.3.4 I present in depth discussions of a series of example interactions that illustrate key patterns in the Eastmond students’ constructed response data. As a means of orienting the reader, I provide here a brief overview of the patterns before discussing each of them in detail through a series of illustrative examples.
As mentioned in the introduction, a central pattern that emerges from the data relates to mismatches in the ideas addressed in the students’ answers and the ideas intended by the test developers. Emergent bilingual students’ constructed responses exhibit more evidence of these mismatches. I refer to this phenomenon in terms of schema activation “miscues,” in line with schema-theoretic literature showing how similar types of miscues can occur during the reading process (e.g. Rosenblatt, 1985; Goodman & Goodman, 2004). Within these schema activation miscues, there are two distinct types.

The first type of miscue involves the students’ schema related to motion and forces. With these miscues in the students’ organized background knowledge specific to physics, students produce constructed responses that relate to a different target idea about motion and forces than the one intended by the test developers. In many cases, this type of miscue can be traced to the sequence in which the test items are presented (e.g. a series of items related to gravity immediately followed by items targeting inertia and friction).

The second, less common type of schema activation miscue exhibited in the emergent bilingual students’ answers is an activation of another, off topic schema. Constructed responses that display this second type of schema activation miscue relate to students’ outside (non-physics) knowledge, often prompted by some specific lexical or grammatical aspect of the test item prompt language or the accompanying visual.

While native English speaking students do display some of the same evidence of schema miscues as the emergent bilingual students—especially on the “Rolling
Ball” items—there are relatively more native English speaking students who are able to navigate the test item prompts in order to address the intended target idea, unlike the emergent bilingual students. Native English speaking students seem to notice subtle yet crucial contextualization cues in the test prompts (as evidenced, for example, by underlining of key word in the item prompts), employ more repetition of key words from the item prompts in their answers, and utilize techniques such as underlining to signal meaning in their responses to test items. I interpret this as relatively more successful communication between the native English speaking students, the test developers, and the test raters.

In what follows, I discuss a series of examples that illustrate how these patterns manifest discursively. I first discuss schema activation as related to the “Rolling Ball” series of items in Sections 4.3.1 and 4.3.2, as this series of items is most problematic for emergent bilingual students in terms of schema activation and results in the largest achievement gaps. Then, in Sections 4.3.3 and 4.3.4, I examine several more examples of schema activation miscues that arise in relation to the “Gravity and a Tennis Ball” series of items. Section 4.3.5 provides a summary of these findings, and the concluding Section 4.4 provides additional discussion and contextualization of the findings.

4.3.1. Schema Activation Miscues in the “Rolling Ball” Items

The first series of examples that I use to demonstrate schema activation miscues relate to the “Rolling Ball” items. “Rolling Ball Part I” was discussed in the
introduction to this chapter, and I present it again here to remind the reader of its basic elements.

Rolling Ball Part I.

As mentioned earlier, students can receive full credit on the selected response portion of this item by selecting either yes or no, depending on how they explained their answers in part B. The answer that the test developers intend as the correct answer for this item was “no” for the selected response with an explanation that one force starts the push but does not constantly push on it as it moves; however, students can choose “yes” and explain that the force of friction is constantly acting on the ball as an alternative way of achieving a correct item-level
score. In this way, the item is meant to target QAM Idea #4—an object at rest stays that way unless acted on by a force and an object in motion will continue to move unabated unless acted on by a force (the rule of inertia).

However, many of the students in the Eastmond classroom (17 out of 27, or 63%) do not engage with Target Idea #4 when responding to this item and thus receive no credit in their ratings for the item. In this section, I discuss three examples that demonstrate the types of schema miscues that occur in response to the “Rolling Ball” questions, causing students to engage with different target ideas than the ones intended by the test developers. First, I show how one emergent bilingual student, Walter, seems to continue to use his gravity schema as he works through the “Rolling Ball” item. Then, I show an example from a native English speaking student, Nathan, who also struggles with the item in a similar way.

After showing this first type of schema activation that is shared across the emergent bilingual and English speaking groups—although it is more prevalent in the case of emergent bilingual students—I show a different type of schema miscue that occurs only for the emergent bilingual students. In this third example, Ralph discusses other target ideas in his constructed response. The fourth example shows a schema miscue from the test of another emergent bilingual student, Dylan, which may be related to the visual accompaniment for the item.
Example 1: Walter’s Gravity Schema Miscue

In this case, Walter answers “Yes” for the selected response and in his constructed response mentions that “The force is pulling the ball down.” I interpret this as a reference to gravity, although he does not use that key word here. The sequence of items may be a factor in contributing to this schema miscue, along with other elements of Walter’s interaction with the item. The previous series of items, the “Gravity and a Tennis Ball” items, cue Walter’s schema related to gravity. Thus, it is unsurprising that Walter interprets the “Rolling Ball” item as potentially also related to gravity.

Indeed, many students—both native English speaking (three out of seven miscues) and emergent bilingual (two out of 10 miscues)—mention gravity in their responses to this item. Much like the emergent bilingual student above, Nathan, a native English speaking student also exhibits a gravity schema miscue in Example 2 below.

---

37 On Walter’s pretest and in his assessment interview, Walter mentions gravity when prompted with this “Rolling Ball” question. This additional information bolsters my interpretation that “the force” he is referring to here is gravity.
Example 2: Nathan (Native English speaking Student) Gravity Schema Miscue

While gravity serves as a common red herring for students on the “Rolling Ball” items, there are other miscues in their motion and forces schemas, as evidenced in the next two examples from the emergent bilingual students Ralph and Dylan.

Example 3: Ralph’s Target Idea #1 and #2 Schema Miscue

Here Ralph discusses two target ideas of the QAM assessment: 1) that motion is caused by forces (“...the push give the ball rolls”), and 2) the greater the force, the greater the motion (“if the boy gives a bigger push the ball will roll fast and far away...”). Instead of engaging with the intended Target Idea #4, Ralph demonstrates
accurate knowledge of two other target ideas; however, because he does not engage
with Target Idea #4, the raters give Ralph a score of zero for this item.

   It is interesting to note that Ralph also struggles with this item on his pretest
in which he notes: “he is only stand and lisent music he are not doing force” [sic]
which indicates a possible miscue related to the visual graphic accompanying the
item which features a boy wearing headphones:

   Dylan’s posttest constructed response for the “Rolling Ball Part I” item also
shows a schema miscue that may be related to the visual accompaniment to the
item, as he describes, “Yes, because the wind might pushing the ball and so the ball
is forced to move forward.”
Example 4: Dylan Discussing Wind and Target Idea #1

b. Explain your answer.

Yes, because the wind might push the ball and so the ball is forced to move forward.

This answer accurately engages with Target Idea #1 (changes in speed or direction are caused by forces, with wind being the force in this case). However, like many of the other emergent bilingual students, Dylan fails to address the intended Target Idea #4 and receives no credit for this answer. Here, too, it seems that subtle aspects of the visual accompaniment to the item may play a role in schema activation miscues. This reiterates findings from Martiniello (2008), Logan-Terry & Wright (2010), and Wright & Logan-Terry (in preparation) as to the particular importance of item visuals for emergent bilingual students. In the visual accompaniment to the item, there are lines drawn next to the ball, possibly to indicate the movement of air as a result of the ball’s motion:
This subtle aspect of the test item construction may give off information that causes Dylan’s motion and forces schema to cue “wind” instead of the intended target idea.

In sum, these examples from Walter, Nathan, Ralph, and Dylan's posttest constructed responses show the types of schema miscues that occur for emergent bilingual and native-English speaking students in response to the “Rolling Ball” questions. Some students seem to continue to use their background knowledge and expectations related to gravity when engaging with the “Rolling Ball Part I” item. This type of miscue may be related to sequential aspects of the QAM test (with inertia/friction questions immediately following a series of four questions related to gravity), as other schema activation studies have indicated (e.g. Pearson, 1985; Rowe, 1986; Rowe & Rayford, 1987).

Other students accurately describe additional QAM target ideas (most commonly Target Idea #1 and Target Idea #2) in response to the “Rolling Ball” prompts, yet do not seem to pick up on the key prompt word “constantly” that is meant to serve as a contextualization cue to signal that the item deals with Target Idea #4. In addition, some of the schema miscues with relation to this item may be linked to students’ interpretation of subtle aspects of the visual accompaniments to the item. This latter type of schema miscue in particular seems to be related to “information given off” (Goffman, 1981) in aspects of visual accompaniments, rather than information intentionally given by the test item developers.
4.3.2. Overcoming Schema Activation Miscues in the “Rolling Ball” Items

As mentioned in the previous section, schema activation miscues do arise in the native English speaking students' interactions with the “Rolling Ball” items as well. However, at least some of the native English speaking students are able to notice the contextualization cues (e.g., the bolded “constantly” in the item prompt) that the test developers use to signal that this item related to Target Idea #4. I present two examples to demonstrate this point. First, one student, Ella, underlines the key word in the item prompt, and in the second example Kyle also repeats the key word “constant” in his constructed response.

Example 1 - Ella Underlining Key Words in Item Prompt

![Image of an example response]

38 It is impossible to know exactly how many students notice this particular contextualization cue. The ratio of native English speaking students with schema miscues for this item (seven out of 15) as compared with that of the emergent bilinguals (10 out of 12) may indicate an increased likelihood of noticing these sorts of contextualization cues for native English speakers.
Ella’s explanation: "No because if it was constantly pushing or pulling the ball it would keep moving" features a repetition of the key word from the item prompt. This is in line with findings from Reynolds (1995) that showed native speakers are more adept at repeating key aspects of texts than emergent bilinguals. In addition, I interpret Ella’s underlining of “constantly” in the test prompt as evidence that she has noticed the importance of this word in showing which target idea is being addressed here. This also relates to the work of Gumperz and Goffman, who argue that meaning is constructed through the interpretation of contextualization cues and framing. I view the bolding of “constantly” by the test makers as explicit, conscious information given as a way to cue Target Idea #4 in this item.39

Similar to Ella’s example, Kyle repeats this key word in his own constructed response to the “Rolling Ball Part I” item.

Example 2: Kyle Repeats the Word “Constant”

39 This view is bolstered by an informal interview that I conducted with one of the test developers, who explicitly mentioned the importance of the bolding of the word “constantly” in this item prompt. He described how the test developers intend for students to notice the bolding as a sign that this item relates to the rule of inertia rather than one of the other key ideas from the QAM test.
Again, the bolding of the word "**constantly**" in the QAM #9 item prompt can be seen as a contextualization cue on the part of the test developers, serving to frame (Goffman, 1981) the item as an item concerning the higher level concept of inertia (Target Idea #4) that the students have recently discussed in class. This framing seems to be recognized by Ella and Kyle and thus the item activates their physics knowledge schemas in a more successful way than in the case of the emergent bilingual students. This is in line with evidence from other studies employing introspective methodologies (e.g. Logan-Terry & Wright, 2010; Rea-Dickins & Khamis, 2010; Wright & Logan-Terry, in preparation) that also indicates these sorts of subtle contextualization cues in test item prompts are more likely to be noticed by native English speakers than by emergent bilingual students.

In addition, the use of repetition of key item prompt language here not only relates to the findings of Reynolds (1995) about differences in emergent bilingual and native-English speaking students’ usages; it also relates to Deborah Tannen’s Interactional Sociolinguistics work (1989/2007) showing the importance of this type of allo-repetition in creating coherence and demonstrating understanding in both spoken and written contexts.

I now present a third example that actually comes from one of the rare (two out of 12) interactions with Rolling Ball Part I in which an emergent bilingual student, Michelle, is able to overcome schema activation miscues and achieve a correct score on the item. Her example of overcoming schema activation miscues shares some similarities with that of the native-English speaking students; however, there are also key differences.
Michelle’s knowledge schema seems to have been cued in a different way than the other students; she mentions friction instead of inertia or gravity. In terms of structure, her answer is similar to some of the native English speaking students’ answers that overcome schema activation miscues: it uses key words and it features a relatively simple sentence structure (the same type of “pithy” response that I discuss for native English speakers). It is interesting to note that Michelle is the only student in the Eastmond classroom to mention friction for this item and achieve a correct answer in this way. Also interesting is the fact that Michelle does not receive credit for the follow-on question in the “Rolling Ball” series that actually intends to target the notion of friction, even though her organized background knowledge related to this concept seems to have been activated in the previous item.

Indeed, the follow-on question in the “Rolling Ball” series—Part II—further highlights the differences between schema activation for emergent bilinguals and native English speaking students. While not a single one of the emergent bilingual students mentions the target idea of friction for this item, a relatively large proportion of the native English speaking students are able to read the cues that this item is about friction (six out of fifteen native English speaking students explicitly
mention “friction;” other students write about the notion without using the keyword), an instantiation of Target Idea #1 (that changes in speed or direction are caused by forces).

I now present examples four through seven in order to showcase how native English speaking students use a variety of resources in their constructed responses to the follow-on “Rolling Ball Part II” item that helps signal to the raters that they understand the target idea of that item (friction).

**Rolling Ball Part II.**

![Part II. The ball slows down and stops before reaching you.]

What caused the ball to stop?

The native English speaking students who succeed on this item use various techniques in the construction of their responses to signal their understanding of what the test developers were targeting with the item. Many students simply provide the one key word answer “Friction,” as evidenced by the below examples from Quincy and Brad. Others, like Nathan, employ more elaborate sentences that
discuss how gravity contributes to friction. One student, Owen, makes use of a technique similar to Ella’s underlining of the key word in the test prompt: Owen underlines the key words in his own constructed text. I interpret this as his way of highlighting and framing his answer as relating to the idea targeted by the test developers.

*Example 4: Quincy Using Simple, Key Word Answer*

![Example 4: Quincy Using Simple, Key Word Answer](image)

*Example 5: Brad Using Misspelled Key Word Answer*

![Example 5: Brad Using Misspelled Key Word Answer](image)

In this case, Brad still receives credit for this answer, even though he has misspelled “Friction” as “Fritshun.”

*Example 6: Nathan’s Description of Gravity and Friction*

![Example 6: Nathan’s Description of Gravity and Friction](image)
Here we see Nathan also mentioning “friction” in addition to his discussion of how gravity contributes to friction. This demonstrates mastery of this target idea, and Nathan receives full credit for this answer.

Example 7: Owen 10 Underlining Key Word in His Constructed Response

Nathan writes: “The ball stops because of Friction. If you kick the ball hard then the ball will go far. The ball might stop but at least it doesn’t go forever.” In addition to mentioning the key word “Friction” and underlining it to highlight its salience in his answer, Nathan goes on to deal with another target idea (that increased force causes increased motion, Target Idea #2). The third sentence in Nathan’s answer is difficult to interpret. I would argue that language construction issues like those found in Nathan’s third sentence do not interfere with students’ ability to communicate with the raters (indeed, Nathan received full credit for this answer). In this way, it seems that relative success of item schema activation during students’ interactions with items may trump writing issues for native English
speaking students just as it does for emergent bilinguals (as mentioned in the introduction).

The examples from the posttests of Ella, Kyle, Michelle, Quincy, Brad, Nathan, and Owen in this section show how native English speaking students’ interactions with test items are generally more successful than those of the emergent bilingual students. While Michelle’s example shows how emergent bilingual students sometimes overcome schema activation miscues by using pithy, key word answers (e.g. repetition of the key word “friction” in a simple sentence), native English speaking students overcome schema activation miscues more often (as evidenced by the much lower number of schema miscues in their tests) and in more diverse ways (as evidence by the range of types of successful constructed responses).

These examples show that Native English speaking students seem to be able to notice the contextualization cues that the test developers use to signal the target idea of items. The native English speaking students also utilize techniques such as underlining key words in both the test prompt and their own constructed responses as contextualization cues and framing devices in order to signal to the test raters which target idea they interpret the item to be about. The native English speaking students also repeat key words from the item prompts in their own constructed responses. I argue that all of these factors contributed to their higher scores on constructed response items—thereby also helping to explain the increased

---

40 As mentioned in the introduction to this chapter, emergent bilingual students’ responses are flagged for schema miscues 29 out of 84 times (35% of their answers relate to a different target idea than the one intended) across the total seven constructed responses. In contrast, native English speaking students’ responses are flagged for schema miscues only 18 out of 105 times (17% of their answers relate to a different target idea) across the total seven constructed responses.
achievement gaps between emergent bilingual and native English speaking students on the constructed response items.

4.3.3. Schema Activation Miscues in the “Gravity and a Tennis Ball” Items

The next series of examples that I use to demonstrate schema activation miscues relate to the “Gravity and a Tennis Ball” items.
Gravity and a Tennis Ball Part I.

Both emergent bilingual and native-English speaking students perform relatively well on the initial constructed response item in the “Gravity and a Tennis Ball Part I” series (seen above). The answer that garners credit here is that the
speed of the ball increases due to gravity pulling it down. Most students are able to repeat the word “Gravity” from the item title for this question.

In contrast, I find evidence of schema activation miscues in the emergent bilingual students interactions with the follow-on “Gravity and a Tennis Ball Part II” items:

\textit{Gravity and a Tennis Ball Part II.a.}

\begin{itemize}
  \item Part II. Maria throws the tennis ball into the air. Before it gets to Katie, it falls back to Maria.
  \item a. What caused the tennis ball to change its direction from going up to going down?
\end{itemize}

For this item, the correct answer is, again, gravity (Target Idea #1—changes in speed or direction are caused by forces). While the first two constructed responses in this series both relate to gravity, the next question in this series shifts to a different topic idea:

\textit{Gravity and a Tennis Ball Part II.b.}

\begin{itemize}
  \item b. What does Maria have to do to the tennis ball to get it up to Katie?
\end{itemize}
In this final question in the series, the target idea is Idea #2—the greater the force is, the greater the change in motion will be. Students receive credit for this item from the raters if they write things like “use more force” or “throw it harder.”

I now present two examples of the types of schema activation miscues that emergent bilingual students encounter when interacting with “Gravity and a Tennis Ball Part II.a.”

In the first example, from the posttest of Ralph, we see the student’s answer relates to the speed of the ball rather than the force that causes it to change direction. I argue that this type of miscue may be due to residual schema activation from the first two items (“Gravity and a Tennis Ball Part I”) in this series that cue students to think about the speed of the ball as it is in the air.

**Example 1: Ralph Discusses Speed of the Ball Instead of Gravity**

```
a. What caused the tennis ball to change its direction from going up to going down?

the ball will be increases because she threw up to go more fast.
```

Here Ralph mentions “the ball will be increases” which relates directly to the prompt in the previous questions (“Will the speed of the ball increase, decrease, or stay the same?”) as opposed to the force that causes the change in motion. I argue
that this schema miscue, then, may be related to the sequence of the items that is more problematic for emergent bilinguals than for native English speaking students. It is also interesting to note that Ralph received no credit for the previous constructed response (“Gravity and a Tennis Ball Part I.a.”), as well, as he choose “decreases” and writes, “decreases because Katie is standing up of the windows and Maria is outside waiting the ball is falling down,” [sic] a description of what he is seeing in the visual accompaniment. It seems that his expectations and knowledge schemas related to the test as a communicative event may lead him to believe that, of the choices given in the initial selected response item in the series (“increases,” “decreases,” or “stay the same”), there are probably items in the continuation of the series that deal with the other options.

In the second example, we see that Michelle writes about Target Idea #2 (the greater the force is, the greater the change in motion will be) but does not address the intended Target Idea #1 of the item. Her answer would have garnered her full credit on the “Gravity and a Tennis Ball Part II.b” that targets Idea #2; here her schema miscue yields an answer that is rated for zero credit on the “Gravity and a Tennis Ball Part II.a.”
Michelle writes, “Maria did not throw the ball hard enough.” Unlike Ralph, she does seem to engage with the change in direction aspect of the prompt; however, the question cues a different aspect of her knowledge schema related to motion and forces than the test developers intend. Michelle miscues to Target Idea #2 (the greater the force is, the greater the change in motion will be), which is actually the target of the next item, but she does not address the intended Target Idea #1 of this item (motion is caused by forces) and thus receives no credit for this item.

Indeed, interactions with the final item in the “Gravity and a Tennis Ball” series are very interesting. They produce the most examples of the second type of schema activation miscue (evidence of engagement outside the students' physics schemas). In this section, I present an example of this type of schema activation miscue, one in which Valerie discusses the details of playing tennis instead of answering the question. In addition, I will present a second example of an interaction from this item, in which Greg does not seem to understand at all what the test developers are asking. Finally, I present an example from a native English
speaking student, Carrie, who also has schema activation miscues with this item. In her case, the type of schema activation is more in line with the types we saw for previous items (discussing a different QAM target idea than the one intended for that item).

Example 1: Valerie Schema Miscue Outside of Physics Schemas

Valerie struggles with this particular series of items. She answers, “Maria needs to push the ball with the rakets, so it could get on Katie’s side.” This does feature the physics key word “push;” however, Valerie has also provided evidence that she is thinking of a tennis game in the hypothetical scenario created in this item—as seen in her use of “rakets” and “Katie’s side.” This sort of schema miscue related to outside, non-physics knowledge is more rare than miscues related to other target ideas related to motion and forces, but its occasional occurrence in the data highlights the importance of constructing items that might avoid this type of miscue. For example, the story for this item could simply have involved the dropping/throwing up “a ball” instead of “a tennis ball” as some students may think that the tennis aspect of the item is salient, as it seems Valerie does here. This also
points to the importance of specific lexical items in test prompts that may cause schema miscues for certain students, depending on their knowledge schemas.

*Example 2: Greg Struggles to Understand “Gravity and a Tennis Ball Part II”*

One emergent bilingual student, Greg, struggles considerably with these “Gravity and a Tennis Ball Part II” follow-on items in the series. He does well on the initial questions in the series on both the pretest and the posttests. In addition, he demonstrates ample understanding of Target Idea #1 and Target Idea #2 (the target ideas of these follow-on items) in other contexts. However, he leaves “Gravity and a Tennis Ball Part II.a” blank on both the pretest and the posttest. In Greg’s response to the “Gravity and a Tennis Ball Part II.b” item seen above, it is unclear to me what he is trying to communicate. As expected, the raters give him zero credit for this item.

It is possible that certain grammatical aspects of the item may cause confusion for Greg and other students. The test developers reported consciously writing the test prompts in “colloquial” language in hopes that it would resonate

---

41 Preliminary analysis of students’ demonstration of knowledge in test contexts as compared with their classroom discourse, workbooks, and assessment interviews indicate that emergent bilingual students are able to demonstrate greater evidence of understanding of key target ideas outside of the testing context. A more detailed analysis is forthcoming (please see Logan-Terry, in preparation).
more with students. In this case, it seems that the covert modality of the item prompt construction ("What does Maria have to do ..." vs. “What would Maria have to do ...”), coupled with the idiomatic purpose clause expression “have to do with the tennis ball to get it up to Katie” (featuring multiple prepositional clauses with null subjects), may cause schema miscues for certain students, especially emergent bilingual students.\textsuperscript{42}

\textit{Example 3: Carrie (Native English Speaking Student) Addresses Target Idea 1 Instead of Target Idea 2}

\begin{quote}
\textbf{b. What does Maria have to do to the tennis ball to get it up to Katie?}
\underline{Throw the ball back to her}
\end{quote}

This example shows how a native English speaking student also struggles with schema activation for this item. However, in contrast with the schema miscues or her emergent bilingual peers, Carrie seems to engage with a different target idea (that motion is caused be forces) rather than the one intended by the test developers for this item (that more force is needed to cause more motion). In this way, her schema miscue is more line with examples from the “Rolling Ball” series in which a different physics target idea is addressed, as opposed to the type of miscue that this item produced from some of the emergent bilingual students: cue of outside (non-physics) knowledge schemas.

\textsuperscript{42} I owe thanks to Paul Portner, Justin Kelly, and Gergana Boft for their insights into why this prompt may have been particularly problematic for emergent bilingual students.
As this section has shown, these interactions with “Gravity and a Tennis Ball Part II” from the posttests of Ralph, Michelle, Valerie, Greg, and Carrie show various schema activation miscues that occur often for emergent bilingual students and occasionally for native English speaking students. We have seen that the sequence of the items may influence students’ constructed responses (e.g. discussing the speed of the ball in the follow-on questions that come after the initial questions about the speed of the ball). There are also miscues with other QAM target ideas that are not the intended target of the particular items in this series. We also saw in this section the more rare type of schema miscue that involves students’ outside, non-physics knowledge (e.g. Valerie’s discussion of tennis). Finally, I argue that the grammatical structure of the latter item in the “Gravity and a Tennis Ball” series may cause problems for emergent bilingual students as well as native English speaking students.

4.3.4. Overcoming Schema Activation Miscues in the “Gravity and a Tennis Ball” Items

Much like the findings from the “Rolling Ball” items, the “Gravity and a Tennis Ball” items prove less problematic in terms of schema activation for native English speaking students than for emergent bilingual students. I present here four examples of native English speaking students’ interactions with the “Gravity and a Tennis Ball” items. The examples from the constructed responses of Joy, Nathan, and
Stacy shows a variety of ways that native English speaking students are able to succeed on this item, despite possible schema activation miscues.

Finally, I present two examples from an emergent bilingual student, Hank, who uses slightly different techniques than the native English speaking students to achieve full credit on these “Gravity and a Tennis Ball” items.

**Example 1: Joy’s Verbose Answer Overcomes Schema Activation Miscues**

![Example Image]

Joy first discusses the “force that Maria had on the ball” instead of the target idea of gravity. In this way, Joy exhibits some similar schema activation miscues as found in the answers of the emergent bilingual students (engagement with ideas other than the one targeted); however, she is able to overcome this by constructing additional sentences that deal with the key target idea (“gravity”). Joy is able to use multiple sentences to overcome any possible confusion about which target idea the item is oriented toward by addressing multiple target ideas.
Example 2: Nathan’s Pithy Use of Key Word “Gravity”

a. What caused the tennis ball to change its direction from going up to going down?

gravity causes the tennis ball to come back down.

This example from Nathan’s post-test represents a pithy, keyword (a single, simple sentence that begins with the keyword “gravity”) native English speaking student response that is given full credit by the raters. This sort of response was not uncommon among the correct answers of the native English speaking students.

Example 3: Joy Discusses Force

b. What does Maria have to do to the tennis ball to get it up to Katie?

Maria has to put a lot of force in to throw the tennis ball back up to a high level building.

In this example, Joy mentions key words such as “force” and shows understanding of the hypothetical scenario (“high level building”), garnering her full credit on this item. While these previous examples involve the use of key words,
Stacy’s example below shows a different way of succeeding on this item, despite possible schema activation miscues.

**Example 4: Stacy Discusses Manner of Throwing**

b. What does Maria have to do to the tennis ball to get it up to Katie?

She has to throw it really hard.

Stacy’s answer does not use key physics words as Joy’s answer does, but the activation of her knowledge schema related the manner of throwing is also effective in succeeding in communicating to the raters that she understands the target idea (more force causes more motion), even though she uses a more colloquial construction.

The following two examples from the posttest of an emergent bilingual student, Hank, show how—despite L2 writing issues related to article usage and prepositional phrases, for example—some emergent bilingual students are able to overcome schema activation miscues and produce responses that feature key words to get them full credit on the “Gravity and a Tennis Ball” items.
Example 5: Hank (Emergent Bilingual Student) Using Key Words to Get Credit

Hank’s answer, “The gravity pulled it down when the Force applied into it to throw the ball to Katie,” features two key words, “gravity” and “Force,” that help him achieve full credit for this item. Hank is also careful to self-repair in certain instances, including the prepositional phrase “… in into it to throw …” and the spelling of the name of one of the characters in the hypothetical scene created by the test developers for this item: “Kate Katie.” I interpret Hank’s use of self-repair as an indication of the L2 writing trouble he has as a relatively low English proficiency emergent bilingual. Despite this trouble, he is able to achieve a score of “correct” because he does not seem to experience a schema activation miscue as to the target idea of the item and is able to communicate to the raters that he understands the target concept.
Example 6: Hank (Emergent Bilingual Student) Again Using Key Words to Receive Credit

b. What does Maria have to do to the tennis ball to get it up to Katie?

Use a more force to throw it higher

Again, Hank’s constructed response for this item mentions the key words “Use a more Force to throw it higher” [sic], and he receives full credit for this item.

In sum, these examples from the interactions of native English speaking students, Joy, Nathan, and Stacy, with the “Gravity and a Tennis Ball Part II” items show discursive evidence of how native English speaking students are able to achieve full credit on items that cause relatively more schema activation miscues for a portion (13 out of 36 item interactions, or 36%) of emergent bilingual students. Students who might experience similar schema miscues as some of the emergent bilingual students (e.g. Joy mentioning Target Idea #2 on “Gravity and a Tennis Ball Part II.a”) use multiple sentences to include additional target ideas; thus, the raters are able to give them credit for the item. In addition, native English speaking students repeat key words from the item prompts in their answers, as well as more common sense phrases (“throw it really hard”) to communicate with the raters. Echoing the findings from the “Rolling Ball” items, it seems that native English speaking students are able to avoid/overcome schema activation miscues, thereby contributing to increased gaps between their achievement on these items and the achievement of the emergent bilingual students. Hank’s examples show how
emergent bilingual students are also able to achieve full credit on these items by using key word answers in some cases.

4.3.5. Summary

In sum, upon examination of students’ constructed responses, I find that emergent bilingual students struggle more with two types of schema activation: 1) miscues in the students’ schema about motion and forces, resulting in constructed responses that related to a different target idea about motion and forces than the one intended by the test developers and 2) activation of another, off topic (non-physics) schema. Each of these schema activation miscues are linked to specific issues related to: 1) the sequence of test items (e.g. inertia and friction items following a series of items that targeted gravity), 2) the particular lexical items provided in the text prompts (e.g. “tennis ball” as opposed to just “ball” as the students’ tennis schema could be activated, 3) grammatical aspects of test prompts (e.g. the covert modality of the item prompt construction “What does Maria have to do to the tennis ball to get it up to Katie?”) and 4) the visual accompaniment of items (e.g. inclusion of headphones in the drawing related to the inertia item that may cause schema miscues for certain students).

While emergent bilingual students struggled with these schema miscues more often, many of their native English speaking peers are able to achieve full credit on these items. Indeed, as mentioned in the introduction to this chapter, emergent bilingual students’ answers are flagged 26 out of 60 times (43% schema
miscue rate) on the five constructed responses in the “Gravity and a Tennis Ball” and the “Rolling Ball” series of items. In contrast, the native English speaking students’ responses are only flagged 17 out of 75 times (23% schema miscue rate). Through the use of 1) multiple sentences that targeted multiple target ideas, 2) pithy (single sentence) keyword answers, 3) repetition of key words from prompt language (e.g. “gravity” or “friction”), and 4) contextualization cues such as underlining both prompt and answer language, it seems native English speaking students are able to more successfully interact with the QAM test items.

4.4. CONCLUSION

In this conclusion, I briefly overview and reiterate the main contributions of this qualitative analysis. I then discuss my findings from this analysis as they relate to existing literature. After this contextualization of the findings, I discuss implications of this chapter’s findings for test development and the overarching research questions of my dissertation effort. Finally, I highlight caveats and shortcomings of the analysis that help motivate areas for future research.

A key contribution of this chapter is the demonstration that knowledge schemas (Tannen & Wallat, 1993) of each of the three interactional parties involved in test interactions—test developers, students, and raters—play a crucial role in the construction of meaning during the test taking process. I discuss issues in the activation of these knowledge schemas and argue that emergent bilingual students
struggle more with these schema activation miscues than native English speaking students.

Furthermore, at each level of student–test interaction (the test construction and the students’ reading of the item and subsequent construction of a constructed response that is then read by the test raters), I find important contextualization cues (Gumperz, 1982, 1999) that sometimes overlap with those found in face-to-face interaction (e.g. lexical choice, grammatical constructions) and sometimes differ from those used in face-to-face interaction (e.g. underlining written words for emphasis, aspects of drawings and other visual accompaniments to items). Much like in the spoken context, these contextualization cues function to frame (Goffman, 1974, 1981) the test interaction as relating to a certain target idea. I find mismatches in the framing of these interactions that seem to contribute to miscommunication between test developers, students, and raters.

I discuss this phenomenon here in terms of schema “miscues,” in line with research that has shown readers can interpret a text very differently, depending on their background knowledge and expectations (Anderson & Pichert, 1978; Pichert & Anderson, 1977; Ruddell & Unrau, 2004) as well as research showing how students make inferences about the content of text as a result of cues in the text (Pearson, 1985; Rowe, 1986; Rowe & Rayford, 1987). The contextualization cues identified in this study include choice of lexical items, grammatical structures, and bolding and underlining of key words in both the test item prompts as well as the students’ constructed responses. I argue that all of these contextualization cues are pivotal in the process of schema activation and communication during the test taking
interactions and should thus be taken into consideration in the test development process.

The types of schema miscues I identify here relate to various aspects of the test interactions. The sequence of test items is shown to be a crucial factor, which highlights the importance of considering sequence and framing in test development. The particular lexical items provided in the text prompts, grammatical aspects of test prompts, and visual accompaniment of items are also shown to be important. This confirms Interactional Sociolinguistic notions of the relevance of information given explicitly (e.g. key words in item prompts) as well as information given off unconsciously (e.g. headphones in the visual accompaniment as an indicator of the activity in progress) that can result in framing mismatches and miscommunication. These findings also reiterate results from work that has shown the particular importance of visual accompaniments for activating emergent bilingual students’ knowledge schemas, perhaps as a result of language proficiency issues (Logan-Terry & Wright, 2010; Martiniello, 2008; Wright & Logan-Terry, in preparation).

All of these findings as to the four main sources of miscommunication between test developers and emergent bilingual students (sequence, lexical items, grammatical constructions, visual accompaniment) have important implications for test development. As I mentioned in the implications of the quantitative study in Chapter 3, researchers can utilize these findings about specific problematic aspects of test items to help inform test accommodations. The preliminary implications for test development of these patterns and patterns from previous chapters are discussed in more detail in the Conclusion Chapter 5.
An additional interesting outcome of these analyses is the identification of ways in which students, especially native English speaking students, were able to achieve correct answers on constructed response items, including the use of multiple sentences that targeted multiple ideas, pithy, keyword answers, repetition of key words from prompt language, and underlining of key words. These findings echo past research findings about the importance of contextualization cues (Gumperz, 1982, 1999; Tyler 1995), repetition in discourse (Tannen, 1989, 2007) as well as native speakers’ increased likelihood to make use of these strategies (Reynolds, 1995).

By extending the Interactional Sociolinguistic analytic framework to this testing context, my analysis also adds to literature in the area of Interactional Sociolinguistics. I argue that the study presented in this chapter is innovative in that it pilots the analysis of written test data from an interactional perspective and helps to show that testing is, in a sense, a gatekeeping (Erickson, 1975; Erickson & Schutlz, 1982; see also discussions of this concept in the Introduction Chapter 1 and Conclusion Chapter 5 of this dissertation) conversation between students and representatives of the educational institution, one of the main contributions of my dissertation effort.

In terms of the overarching validation study of the QAM test that was discussed in the Introduction Chapter 1 of this dissertation, I see the findings from this analysis as bolstering an argument against the validity of the QAM tests for emergent bilingual students, since I provide evidence that emergent bilingual
students struggle more with knowledge schema activation miscues during the test taking process than native English speaking students.

Finally, I should address one of the biggest limitations of this study that points to an important area for future research: after conducting this analysis, I still have a lingering question about the degree to which the nature of students’ physics knowledge schemas contributed to the patterns in schema activation identified here. Schema activation miscues for emergent bilingual students are certainly shaped by the discursive patterns I have illustrated here; however, I am still uncertain about the extent to which emergent bilingual students experience these issues due to differences in the amount and quality of their physics knowledge schemas at the time of test taking.

As Anderson, Spiro, and Anderson (1978) have pointed out, the state of a reader’s relevant knowledge schema at the time of reading influences how likely that reader will be to understand texts. For example, in the case of the QAM test interactions, those students whose physics knowledge schemas have been constructed in ways that are congruent with the ways that the physics concepts are cued on the test will be more likely to succeed in understanding the item (and, thus, potentially address the intended topic idea and achieve a correct answer). In order to begin to address this lingering issue, my next analytic step is to conduct follow-on analyses of evidence of students’ knowledge in alternative contexts (in their classroom discourse, workbooks, and assessment interviews) in order to shed further light on this issue. This and other suggestions for future research are also discussed in the following Conclusion Chapter 5.
CHAPTER 5

CONCLUSION

This concluding chapter provides a brief summary of research findings from each of the analytic chapters in Section 5.1. I then discuss the implications of these findings for the overarching research questions of this dissertation in Section 5.2. The practical, theoretical, and methodological contributions of the dissertation are outlined in Section 5.3. The final Section 5.4 discusses the limitations of my dissertation research and the potential areas for future study.

5.1. SUMMARY

After the introductory Chapter 1 outlines the achievement gap problem and general approach of this study, Chapter 2 serves as an analytic backdrop, wherein I find ESOL group status to be a significant predictor of differences in achievement between emergent bilingual and native English speaking students at the national, state, and classroom level. These analyses in Chapter 2 help to motivate the subsequent, more targeted analyses of sources of achievement gaps in the SCALE-uP corpus.

Chapter 3 presents an investigation of linguistic characteristics of assessment items as sources of achievement gaps through the use of computerized readability tools and multilevel modeling. Key findings from the analyses in Chapter 3 include the variation in relationships between linguistic characteristics of test items and
students’ item-level scores across emergent bilingual and native English speaking
groups of students for certain format, syntax, and discourse variables. This provides
some evidence that the QAM tests conflate the assessment of English language
proficiency and content knowledge, thereby exacerbating achievement gaps for
emergent bilingual students in this context. However, the lack of significant results
for many of the variables tested in this analysis also points to the fact that much of
the variation in students’ item-level scores is not explained by linguistic aspects of
the test item prompts.

As a means of further exploring one of the key findings from the quantitative
analysis in Chapter 3 (that constructed response items were particularly difficult for
emergent bilingual students as compared with native English speaking students),
Chapter 4 examines students’ constructed responses using a qualitative,
interactional framework. In this analysis, I show how students’ interactions with
test items are shaped by the knowledge schemas of each of the interlocutors
involved in the test interactions—test developers, students, and raters. I discuss
examples as a means of showing how students’ schemas can be miscued during their
interactions with test items. These miscues are related to missed contextualization
cues that the test developers employ to frame items as relating to certain target
ideas. They are also linked to issues of sequence of items, lexical and grammatical
aspects of items, as well as the visual accompaniment to items. I argue that the
prevalence of these types of schema activation miscues for emergent bilingual
students seem to be contributing to achievement gaps.
5.2. IMPLICATIONS

In the Introduction Chapter 1 of this dissertation, I propose to investigate achievement gaps in the SCALE-uP corpus. I also frame my study as an evidence-based validity study (Kane, 2010), with potential implications as to whether the QAM tests in the SCALE-uP corpus function as gatekeeping (Erickson, 1975; Erickson & Schultz, 1982) encounters for emergent bilingual students.

While this dissertation is oriented toward investigating achievement gaps, I argue here that the findings from my analyses actually problematize the entire notion of “achievement gaps” by showing how they are—at least to a certain extent—constructed by the very testing practices used to measure achievement. In this way, my analyses shed new light on the how and why of achievement gaps.

The Figure 5.1 below provides a simplified, graphical representation of how I view the findings from the two main analytic chapters of this dissertation study (with Chapter 2 serving as an analytic background chapter) as related to the construction of achievement gaps, with the key findings from Chapter 3 (the left-most two boxes, highlighted in green) and the key findings from Chapter 4 (the left-most two boxes, highlighted in orange. Please note that the issues highlighted in Figure 5.1 are not meant to represent all the sources of achievement gaps for emergent bilingual students; the figure is simply a summary of key findings from my dissertation study.
Furthermore, I argue that my dissertation serves as an argument against the validity of the QAM test because language-related issues serve as sources of construct-irrelevant variance in the tests (e.g. differential effects of linguistic complexity of test prompt language across ESOL groups, test item schema miscues for emergent bilingual students).

Finally, I discuss here the implications of my research as they relate to the notion of gatekeeping mentioned in the introduction to this dissertation study. I argue that my findings demonstrate how the test taking process is, in a sense, a gatekeeping encounter between emergent bilingual students and representatives of the educational institution. Chapter 3 investigates the role of linguistic complexity of item prompts in contributing to achievement gaps, and Chapter 4 examines the
discursive processes in students’ constructed responses that contribute to achievement gaps. In these ways, the analyses provide evidence that the QAM tests serve as linguistic events in which emergent bilingual students (members of a minority community) require approval of the educational institution (the majority community) in order gain access to resources and opportunities for success.

5.3. CONTRIBUTIONS

Throughout each of the analytic chapters, specific contributions of the work are discussed in depth. Here I provide a synopsis and discuss themes that emerged across the analytic chapters. This section also provides additional discussion of the practical, theoretical, and methodological contributions of the overarching dissertation effort.

As mentioned above, the analyses found in this dissertation contribute to the debate surrounding language as a source of construct-irrelevant variance (e.g. Abedi, 2006; Bailey, 2005; Farnsworth, 2008; Martiniello, 2008; Wolf & Leon, 2009) by providing some evidence that linguistic features of test items do differentially affect emergent bilingual and native English speaking students. The analyses in Chapter 3 and Chapter 4 also offer promising directions for future research that can continue to address this debate. For example, Chapter 3 shows that the use of computerized readability tools and multilevel statistical modeling in the investigation of large-scale, standardized tests may uncover additional language-related barriers. Chapter 4 points to the promise of employing targeted qualitative
to shed light on *how* language-related issues may be contributing to achievement gaps.

The findings from each of the analytic chapters also contribute to the ongoing discussion and understanding of the possibilities and limitations of test accommodations for emergent bilingual students (e.g. Kieffer, Lesaux, Rivera, & Francis, 2009; Wolf, Herman, Bachman, Bailey, & Griffin, 2008). Chapter 3’s findings show that certain types of test accommodations (e.g. linguistic simplification) should take into account the aspects of complexity that seem to cause problems for emergent bilinguals (e.g. syntactic complexity, as measured by number of words before the main verb of a sentence) as well as those aspects of complexity (e.g. repetition of syntactic constructions) that may serve to aid in comprehension of test items.

The findings from Chapter 4 related to aspects of test items that contribute to schema miscues for emergent bilingual students (the sequence of test items, the particular lexical items provided in the text prompts, grammatical aspects of test prompts, and visual accompaniment of items) also have significance for the field of test accommodations and test development. I argue that test developers should be aware of the ways in which students’ knowledge schemas are cued and design tests that, for example, do not present a series of items related to one key concept followed by a series of items related to a different key concept without explicitly marking the shift in the focus of the items. Test developers should also be careful in their use of lexical items that may cause students to miscue to schemas outside the
realm of the content testing (e.g. the use of “tennis ball” may cue students’ schemas related to playing tennis rather than the intended physics target idea of gravity).

In addition, evidence from both Chapter 3 and Chapter 4 speak to the importance of considering the types of grammatical constructions that may cause problems for emergent bilingual students. My analyses in Chapter 3 provide evidence that sentences with many words before the main verb seem to be problematic for emergent bilingual students. Chapter 4 shows that sentences with covert modality (e.g. “What does Maria have to do …” vs. “What would Maria have to do …”) and multiple prepositional clauses with null subjects may also be particularly problematic for emergent bilingual students. More research is needed to determine whether these patterns appear in other contexts; however, my work here serves to point out areas that researchers and test developers may want to target in future studies.

In the case of the QAM assessment, the test developers explicitly state that they attempt to use relatively more “colloquial” language to make items more accessible for all students. However, this may actually serve to make items more difficult for emergent bilinguals, as evidenced in some of my analyses (e.g. emergent bilingual students struggling with the idiomatic expressions seen in Chapter 4). Additional research is needed to aid test developers who seek to accommodate tests through the use of linguistic simplification and other test item modifications, as use of these strategies that is not grounded in research may—at best—be ineffective, and—at worst—ironically serve to disadvantage emergent bilingual students.

---
43 This was stated during an informal interview that I conducted with one of the QAM test developers.
Findings from both the quantitative analyses in Chapter 3 and the qualitative analyses in Chapter 4 highlight the importance of visual accompaniments to test items for emergent bilingual students. In Chapter 3, I show statistically significant differences in achievement based on the distance of an item prompt from the item's visual accompaniment. The examples of schema miscues related to visual accompaniments of test items discussed in Chapter 4 shed further light on how and why visuals are important for emergent bilingual students. These findings relate to other studies that have pointed to the salience of visuals for emergent bilingual students (e.g. Martiniello, 2008) as well as research that addresses visual accompaniment through the investigation of computerized and multimodal testing options for emergent bilingual students (e.g. Logan-Terry & Wright, 2010; Wright & Logan-Terry, in preparation).

Another contribution of this dissertation effort is the provision of examples of additional types of test validity studies that are available to researchers. In particular, the chapters demonstrate the usefulness in test validation research of both innovative quantitative methodologies (i.e. the computerized readability tools and multilevel statistical modeling in Chapter 3) as well as rigorous qualitative methodologies (i.e. the interactional and schema-theoretic approach in Chapter 4). The types of analyses found in Chapter 4 are particularly novel in the testing context as they demonstrate the value of a systematic qualitative approach to studying test validity, an area of research that has predominantly been approached via quantitative, psychometric analyses.
As an illustration of the usefulness of qualitative research in test validity studies, I highlight a key contribution of the findings from Chapter 4: the demonstration that knowledge schemas of each of the three interactional parties involved in test interactions—test developers, students, and raters—play a crucial role in the construction of meaning during the test taking process. I argue that test development would benefit from increased consideration of the ways in which individual students’ background knowledge and expectations shape the way they interpret test items. In this way, tests could be created that more clearly address the specific needs of minority populations of students, including emergent bilingual students. Much more research is needed to investigate the ways in which this might be accomplished; I hope my work in this dissertation provokes additional exploration into this important vein of research.

As discussed above, the dissertation makes theoretical and practical contributions to the area of testing research. I also view the work as an important example of applied discourse analysis. In the study described in Chapter 4, I use tools from Interactional Sociolinguistics to address students’ interactions with test items. By extending this interactional framework to a testing context, I show how Interactional Sociolinguistics can be fruitfully applied to different types of interactions than the ones that have been most often studied in the past research employing this framework (i.e. face to face communication). In addition, I see my work as building on the longstanding tradition of utilizing Interactional Sociolinguistics as a means to help address pressing social justice issues (e.g. Gumperz, 1982). In this case, I show how larger, high-stakes gatekeeping
“conversations” between emergent bilingual students and educational institutions are instantiated in students’ individual interactions with test items.

Finally, I view one of the biggest contributions of my dissertation work to be its innovative, mixed methods methodology. As mentioned in the introduction Chapter 1, I employ a sequential, mixed quantitative/qualitative model in order to conduct my analyses (Creswell, 2008). The analysis in Chapter 2 first quantitatively characterizes a broad problem across contexts: achievement gaps for emergent bilingual students. As a next step in investigating the sources of these achievement gaps, the analyses in Chapter 3 employ computerized readability tools and multilevel modeling in order to identify linguistic characteristics of test items that contribute to statistically significant differences in achievement between emergent bilingual students. Motivated by findings from the quantitative analyses in Chapter 3 (i.e. that constructed response items are particularly problematic for emergent bilingual students), the qualitative analyses in Chapter 4 uncover the specific discursive processes that contribute to the patterns identified in the previous quantitative analyses. In this way, the mixed methods design situates my dissertation findings in such a way as to show the larger patterns of achievement gaps, as well as a close analysis of particular features that give rise to these patterns.

5.4. LIMITATIONS AND FUTURE RESEARCH

Having discussed above the contributions of this dissertation, I now turn to discussion of the limitations of my work and the areas of future research that are born out of these limitations.
As discussed in more detail in each of the analytic chapters, there are inherent limitations in working with existing data. For example, my ESOL categorization-based operationalization of “emergent bilingual” and “native English speaking” students is problematic. Some of the “native English speaking” students in my study may in fact speak an additional language at home, even though they have never received ESOL services. In addition, this categorization does not provide any other pertinent information such as the emergent bilingual students’ English language proficiency, the length of time that immigrant emergent bilinguals have lived in the United States, and the family language policies that have been shown to shape emergent bilinguals’ language acquisition and usage across contexts (King, Fogle, & Logan-Terry, 2008).

Another important limitation related to the use of existing data is the very small number of items on the QAM test. I employ analytic tools (e.g. multilevel modeling and qualitative analyses) to address this limitation to the degree possible, but it still constrains the generalizability of my work considerably. As such, I frame the studies in these analytic chapters as pilots or methodological probes that can be applied in future research to large-scale, standardized test data and across larger samples of classroom students.

One issue that has proven particularly challenging to address is the question of students’ underlying content knowledge and its relationship to my findings here. Recent studies have argued that variation in test scores can be accounted for by factors other than language, including students’ underlying knowledge and classroom opportunities to learn (Abedi & Herman, 2010; Koenig & Bachman, 2004;
Gee, 2008). This lingering question motivates an in-progress subsequent analysis in which I compare evidence of students’ knowledge in testing contexts to alternative contexts (i.e. their classroom discourse, workbooks, and assessment interviews) in order to shed further light on this issue (Logan-Terry, 2011, in preparation).

As mentioned in the conclusion section of Chapter 3, I view any study of monolingual content testing for emergent bilingual students—including this dissertation work—as problematic because it does not address the underlying issue that emergent bilingual students cannot utilize their entire linguistic repertoire in demonstrating content knowledge on monolingual content tests. Recent studies investigating multilingual, multimodal assessments and employing introspective methodologies (e.g. Kopriva, 2008; Logan-Terry & Wright, 2010; Schissel, 2011; Solano-Flores & Li, 2009; Rea-Dickins, 2011; Wright & Logan-Terry, in preparation) are promising in addressing this underlying issue.

Finally, I would like to mention some of the process-related limitations of this dissertation study. It attempts to bridge quantitative and qualitative paradigms in order to provide a richer understanding of the achievement gap problem. The work also represents my own first attempt at conducting large-scale, mixed methods research, and I have found this to be a very challenging and humbling process. Concurrent mixed qualitative/quantitative methodologies (as opposed to the sequential model employed in this dissertation), in which there is additional iteration and integration of approaches, have been argued to generate much more robust understanding (e.g. Creswell, 2008). Thus, in future research, I plan to employ concurrent mixed methods designs.
In sum, despite the limitations and issues discussed here, I hope this dissertation work is provocative to other researchers working to build a better understanding of the complexities surrounding achievement gaps for emergent bilingual students.
Appendix A: Questions About Motion Test (QAM)

Date ________________

Questions About Motion

When you answer these questions, think about how motion is caused by pushes and pulls.

Thank you for your time and effort in response to the following questions. Please answer with as much detail as possible.

© 2013 SCALE-CP
Questions About Motion Item 1 (labeled ‘a’ below) and Item 2 (labeled ‘b’ below)

Pushing a Couch

Lucy is pushing a couch across the floor. It is moving slowly. Lucy pushes the couch with the same force and her sister begins to help push it in the same direction.

a. What will happen to the motion of the couch when Lucy’s sister helps push?
Fill in the oval above your answer:

It increases.  It decreases.  It stays the same.

b. Why does the motion of the couch increase, decrease, or stay the same when Lucy’s sister helps her push?

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________
Questions About Motion Item 3

A Full Truck

Bob drives a truck. Sometimes the truck is full; sometimes it is empty. He always applies the same force to the brake pedal to make the truck stop. It takes his truck longer to stop when it is full.

Empty Truck

Full Truck

Explain why it takes longer for the truck to stop when it is full?

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________

Turn the page→
Questions About Motion Item 4 (labeled ‘a’ below) and Item 5 (labeled ‘b’ below)

Gravity and a Tennis Ball

Part I. Katie and her friend Maria are playing with a tennis ball. Katie lets go of the ball and it falls down to Maria who is standing outside.

a. What happens to the speed of the tennis ball as it falls? Fill in the oval above your answer:
   - Speed increases
   - Speed decreases
   - Speed stays the same

b. Explain what caused the speed of the tennis ball to increase, decrease, or stay the same.

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
Part II. Maria throws the tennis ball into the air. Before it gets to Katie, it falls back to Maria.

a. What caused the tennis ball to change its direction from going up to going down?

b. What does Maria have to do to the tennis ball to get it up to Katie?

Turn the page→
Questions About Motion Item 8 (labeled ‘a’ below), Item 9 (labeled ‘b’ below), and Item 10 (labeled ‘Part II’)

A Rolling Ball

Part I. Your friend rolls a ball on the floor toward you. The ball continues to roll.

a. Is there a force constantly pushing or pulling the ball as it rolls toward you? Fill in the oval above your answer: Yes      No

b. Explain your answer.

Part II. The ball slows down and stops before reaching you.

What caused the ball to stop?

Measurement

The Motion and Forces Assessment (MFA)

A panel of experts used an assessment development procedure that was collaboratively created by staff from SCALE-uP and AAAS Project 2061 (AAAS, 2003b) to develop the Motion and Forces Assessment (MFA). The MFA is a curriculum independent assessment aligned with student understanding of force and motion as articulated in Benchmarks for Science Literacy (AAAS, 1993). The assessment map in Figure 1 shows the organization of target ideas and related ideas for the concept at different grade levels.

The assessment is composed of selected and constructed response items designed for ease of use in a diverse classroom setting, using language and illustrations that allow it to be read and understood by a maximum number of 6th grade students (Pyke & Ochsendorf, 2004). MFA consists of 10 items (6 constructed response and 4 selected response) that require understanding of forces and motion. Items are situated in four contexts: pushing a couch, a full truck, gravity and a tennis ball, and a rolling ball. As part of the development procedure, the assessment was subjected to a rigorous pilot testing and revision process using Project 2061’s assessment analysis criteria (AAAS, 2003a; Stern & Ahlgren, 2002). Student responses to the constructed response items were judged by trained raters. All inter-rater reliability Kappa statistics for pairs of raters for each item are between .52 and .93. Average Kappa coefficients for each item across all raters are between .67 and .88. Table 4 shows the average Kappa coefficients by rater pairs for each constructed response task considered in the calculation of scale scores.

Table 4
Cohen’s Kappa measurements of inter-rater reliability on constructed response items on MFA

<table>
<thead>
<tr>
<th>Pushing a Couch: Constructed Response</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
<th>Rater 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td>0.78</td>
<td>0.80</td>
<td>0.88</td>
<td>0.87</td>
<td>0.88</td>
</tr>
<tr>
<td>Rater 2</td>
<td>-</td>
<td>0.75</td>
<td>0.81</td>
<td>0.77</td>
<td>0.76</td>
</tr>
<tr>
<td>Rater 3</td>
<td>-</td>
<td>-</td>
<td>0.79</td>
<td>0.75</td>
<td>0.72</td>
</tr>
<tr>
<td>Rater 4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.86</td>
<td>0.85</td>
</tr>
<tr>
<td>Rater 5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.88</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A Full Truck: Constructed Response</th>
<th>Rater 2</th>
<th>Rater 3</th>
<th>Rater 4</th>
<th>Rater 5</th>
<th>Rater 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td>0.67</td>
<td>0.64</td>
<td>0.75</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>Rater 2</td>
<td>-</td>
<td>0.52</td>
<td>0.72</td>
<td>0.67</td>
<td>0.58</td>
</tr>
<tr>
<td>Rater 3</td>
<td>-</td>
<td>-</td>
<td>0.58</td>
<td>0.66</td>
<td>0.54</td>
</tr>
<tr>
<td>Rater 4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.75</td>
<td>0.68</td>
</tr>
<tr>
<td>Rater 5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.74</td>
</tr>
</tbody>
</table>
Gravity and a Tennis Ball, Part II-a. Constructed Response.

<table>
<thead>
<tr>
<th>Rater</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Average Kappa=0.74</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td>0.68</td>
<td>0.65</td>
<td>0.81</td>
<td>0.77</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>Rater 2</td>
<td>-</td>
<td>0.71</td>
<td>0.74</td>
<td>0.68</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>-</td>
<td>-</td>
<td>0.64</td>
<td>0.64</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.85</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>Rater 5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.80</td>
<td></td>
</tr>
</tbody>
</table>

Gravity and a Tennis Ball, Part II-b. Constructed Response.

<table>
<thead>
<tr>
<th>Rater</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Average Kappa=0.83</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td>0.78</td>
<td>0.78</td>
<td>0.89</td>
<td>0.93</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>Rater 2</td>
<td>-</td>
<td>0.75</td>
<td>0.76</td>
<td>0.79</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>-</td>
<td>-</td>
<td>0.75</td>
<td>0.77</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.92</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>Rater 5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.93</td>
<td></td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Rater</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Average Kappa=0.88</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td>0.93</td>
<td>0.87</td>
<td>0.89</td>
<td>0.90</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>Rater 2</td>
<td>-</td>
<td>0.84</td>
<td>0.88</td>
<td>0.89</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>-</td>
<td>-</td>
<td>0.77</td>
<td>0.83</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.86</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>Rater 5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.93</td>
<td></td>
</tr>
</tbody>
</table>

A Rolling Ball, Part II. Constructed Response.

<table>
<thead>
<tr>
<th>Rater</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Average Kappa=0.72</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 1</td>
<td>0.76</td>
<td>0.62</td>
<td>0.76</td>
<td>0.71</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>Rater 2</td>
<td>-</td>
<td>0.66</td>
<td>0.72</td>
<td>0.72</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>Rater 3</td>
<td>-</td>
<td>-</td>
<td>0.62</td>
<td>0.68</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>Rater 4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.67</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>Rater 5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.80</td>
<td></td>
</tr>
</tbody>
</table>

Note. Cohen’s Kappa coefficient provides an index of agreement between ratings of two variables or ratings of the same variable by two raters when the same scale is used for both. A value of 1 indicates perfect agreement, and a value of 0 indicates that agreement is no better than chance, while a value of −1 indicates perfect disagreement (Howell, 1997). Generally, a Kappa greater than .7 is considered satisfactory.
Once raw scores (sum of correct responses) had been determined, a panel of experts and George Washington University researchers engaged in an analytic judgment process for standard setting (Plake & Hambleton, 2001; Pyke & Hansen, 2005). The standard setting process was undertaken to determine cut scores that distinguish among four levels of understanding: flexible understanding, fluency with ideas, context-limited understanding, and no understanding. The holistic judgments from the standard setting process and analyses of item statistics suggested that three items be excluded from the weighted score calculation. Using this information in conjunction with the concept map (Figure 1), individual item analyses, and input from the panel of experts, a weighting scheme was determined for the MFA. Cronbach’s Alpha for the ten items of the MFA used in calculating the weighted score indicated that the assessment has acceptable internal consistency (α = 0.52).

The weighting scheme distributes the contribution of students’ ideas about:

a. Changes in speed or direction are caused by forces (60% of the total score)
b. The more massive an object is, the less effect a given force will have (15% of the total score)
c. The greater the force, the greater the change in motion will be (15% of the total score)
d. An object at rest stays that way unless acted on by a force/an object in motion will continue to move unabated unless acted on by a force (10% of the total score)

In addition, the scheme balances the contribution of selected response and constructed response items and the difficulty and discrimination of each item in the total score according to the formula:

\[
\text{Total score} = f_{pos} = [0.24(f\text{po}4\text{iicr}) + 0.06(f\text{po}3\text{isr}) + 0.18(f\text{po}3\text{icr}) + 0.12(f\text{po}3\text{iicr1})] + [0.15(f\text{po}2\text{cr})] + [0.03(f\text{po}1\text{asr}) + 0.045(f\text{po}1\text{bcr}) + 0.075(f\text{po}3\text{iicr2})] + [0.05(f\text{po}4\text{isr}) + 0.05(f\text{po}4\text{iicr})]
\]

Each term in the formula corresponds to an assessment item and its associated weight in the final score (e.g. item id fpo4iicr is multiplied by .24 to indicate that it accounts for 24% of the overall scaled score). This formula transforms the 10 items into a scale score that ranges from 0-100. The scaled scores distribute assessments into the four categories previously discussed:

• Scores in the 0-20 range show no consistent evidence of understanding the benchmark ideas.
• Scores in the 21-50 range show some evidence of understanding in specific contexts.
• Scores of 51-70 reveal some fluency with the ideas, but also misconceptions in certain contexts
• Scores of 71-100 were found to demonstrate a flexible understanding of, and commitment to, the benchmark ideas, with few errors or misconceptions.
Appendix C: Rating Guide for QAM Assessment

Rater__________

Motion and Forces Assessment Framework, Rating Information and Scoring Guides

Changes in speed or direction are caused by forces. The greater the force, the greater the change in motion will be. The more massive an object is, the less effect a given force will have. An object at rest stays that way unless acted on by a force, an object in motion will continue to move unabated unless acted on by a force.

Benchmarks for Science Literacy
(Project 2061, 1993)
Assessment Framework

Assessment Purpose

Goal: To determine a student’s scientific understanding of the forces and motion benchmark ideas. These ideas can be broken into four smaller target ideas; (a) changes in speed or direction are caused by forces (b) the greater the force is, the greater the change in motion will be (c) the more massive an object is, the less effect a given force will have and (d) an object at rest stays that way unless acted upon by a force, an object in motion will continue to move unabated unless acted upon by a force.

Impact: Data from this assessment will be used to judge the effectiveness of a new curriculum unit when compared to other curricula and to investigate patterns in responses among groups of students characterized by labels associated with the phrase “diverse learners.”

Assessment Design Specifications

Student Understanding: The assessment ought to target latent student characteristics that determine responses to assessment activities related to the forces and motion benchmarks. Specifically,

- Understanding of, and potential to apply, ideas about changes in speed or direction being caused by forces to phenomena in familiar contexts.

- Understanding of, and potential to apply, ideas about the relationship of force magnitude and motion to phenomena in familiar contexts.

- Understanding of, and potential to apply, ideas about the relationship between mass and force to phenomena in familiar contexts.

- Understanding of, and potential to apply, ideas about objects maintaining constant motion unless a force acts to phenomena in familiar contexts.

Student Performance: What students are expected to do to demonstrate achievement of the benchmark(s) in 20-30 minutes using a paper and pencil instrument.

Students will be expected to read text statements describing phenomena in familiar contexts and respond to prompts for responses. The assessment contains four tasks that require either selecting a response or writing extended responses. Specific qualities of responses are described in the rating guide.
Rating Guide for Pushing a Couch
and Sample Student Responses

Pushing a Couch: Rating Guide

Rationale

The task contains two elements for raters to record information about. There is one selected response item and one constructed response item. The first item asks students to predict the motion of the couch when Lucy's sister helps her push the couch (increase, decrease or stay the same). The constructed response item asks students why the motion of the couch increases, decreases or stays the same. Each item is to be rated independently and a score will be constructed from the independent ratings in a manner consistent with the assessment purpose.

Ratings –

Use these criteria for rating the selected response for Pushing a Couch (part a).

1 = student fills in the circle above “It increases.”
2 = student fills in the circle above “It decreases.”
3 = student fills in the circle above “It stays the same.”
5 = student does not fill in any circle (blank).
6 = student fills in multiple circles

Use these criteria for rating the constructed response item for Pushing a Couch (part b). The focus for rating constructed response items is on identifying the characteristics of a student's reasoning.

1 = rationale or explanation provided for “Applied Force Increases”
2 = rationale or explanation provided for “Change in Energy”
3 = rationale or explanation provided for “Increase in Mass/Weight”
4 = rationale or explanation provided for “Motion Decreased/Stayed the Same”
5 = No evidence of reasoning about the task (blank, I don't know confession, joking response)
6 = Not Able to Judge/High Inference (can't understand response, inaccurate vocabulary use, not enough information, response repeats given information, too much inference required to judge)

Turn to next page for the rating guide chart for Pushing a Couch.
<table>
<thead>
<tr>
<th>Rating Code</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td><strong>Pushing a Couch</strong></td>
</tr>
<tr>
<td>1</td>
<td><strong>Inference</strong></td>
</tr>
<tr>
<td>2</td>
<td><strong>Change in Energy</strong></td>
</tr>
<tr>
<td>3</td>
<td><strong>Increase in Mass/Weight</strong></td>
</tr>
<tr>
<td>4</td>
<td><strong>Motion Decreased/Stayed the Same explanation</strong></td>
</tr>
<tr>
<td>5</td>
<td><strong>No evidence of reasoning about the task</strong></td>
</tr>
<tr>
<td>6</td>
<td><strong>Not able to Judge/High Inference</strong></td>
</tr>
</tbody>
</table>

**Inference**
- Student writes about how there would be a change in energy (get/give more energy) that would increase the motion of the couch. Characteristics of student responses:
  - Because two people are pushing they get more energy than one person
  - Because when her sister came it gave her more energy to push

*see research base on Energy Conceptualization* (Benchmarks, p. 338)

**Inference**
- Student writes about how there would be an increase in mass/weight that would change the motion of the couch. Characteristics of student responses:
  - Because there is a heavier weight to push the couch
  - Maybe an older or bigger person might help it move

**Inference**
- Student selects increase and the student writes about how there would be an increase in mass/weight that would change the motion of the couch. Characteristics of student responses:
  - Because they used the same force
  - It decreases because there is more force pushing on the floor.
  - It decreases because you have more weight to push the couch
  - It decreases because once her sister came to help it was easier to move

**Inference**
- Student selects decreases or stays the same and writes about the motion decreasing or staying the same. Characteristics of student responses:
  - Because they used the same force
  - It decreases because there is more force pushing on the floor.
  - It decreases because you have more weight to push the couch
  - It decreases because once her sister came to help it was easier to move

**Inference**
- Student writes about how there would be a change in energy (get/give more energy) that would increase the motion of the couch. Characteristics of student responses:
  - Because two people are pushing they get more energy than one person
  - Because when her sister came it gave her more energy to push

*see research base on Energy Conceptualization* (Benchmarks, p. 338)

**Inference**
- Student writes about how there would be a change in energy (get/give more energy) that would increase the motion of the couch. Characteristics of student responses:
  - Because two people are pushing they get more energy than one person
  - Because when her sister came it gave her more energy to push

*see research base on Energy Conceptualization* (Benchmarks, p. 338)
Rating Guide for A Full Truck
and Sample Student Responses

A Full Truck: Rating Guide

Rationale

The task contains one element for raters to record information about. There is one constructed response item. The item asks students to explain why it takes longer for the truck to stop when it is full. Each item is to be rated independently and a score will be constructed from the independent ratings in a manner consistent with the assessment purpose.

Use these criteria for rating the constructed response item for A Full Truck. The focus for rating constructed response items is on identifying the characteristics of a student’s reasoning.

1 = rationale or explanation provided for “More Mass/Weight”
2 = rationale or explanation provided for “Inertia or Momentum Only”
3 = rationale or explanation provided for “Objects in Truck Possess Force”
5 = No evidence of reasoning about the task (blank, I don’t know confession, joking response)
6 = Not Able to Judge/High Inference (can’t understand response, inaccurate vocabulary use, not enough information, response repeats given information, too much inference required to judge)

Turn to the next page for the rating guide chart for A Full Truck.
<table>
<thead>
<tr>
<th>Rating Code</th>
<th>Inference</th>
<th>A Full Truck</th>
</tr>
</thead>
</table>
| 1           | More Mass/Weight                        | Student writes that the truck has more mass or weight so it takes longer to stop when the same amount of force is used. Additional explanations (i.e., more force needed to stop the truck in same distance) are acceptable. Characteristics of student responses are:  
- The truck has more mass/more weight/is heavier  
- The amount of weight of the objects in the truck.  
- It takes longer for the truck to stop when it is full because it is heavy causing an increase in inertia. (Inertia alone is not sufficient, see Category 2))  
- The brakes use the same force but they are stopping a heavier load  
- The more things that are in the truck the longer it will take to stop  
- Because F=MA, so to make more mass stop you must give more force  
- The bigger the mass, the more force needs to be applied  
- It takes more friction to come to a complete stop since the mass increased.  
- It takes longer because of momentum. The heavier something is the longer it takes for friction to act on a moving object.  
*Stopping force is implied by using the words friction, pressure, power, energy, strength, push, pull.* |
| 2           | Inertia or Momentum only                | Student writes about Inertia or Momentum but makes no mention of mass. These responses could show evidence of understanding more advanced ideas but high inference is required to judge students’ reasoning about the relationship between mass and force. Characteristics of student responses are:  
- Inertia  
- Because inertia is happening. An object in motion wants to stay in motion.  
- Because there is more momentum the truck is harder to stop  
- Because there is more momentum added to the truck when it is full, making it harder to stop. |
| 3           | Objects in Truck Possess Force          | Student writes that the reason the truck takes longer to stop is because the objects/mass/weight in the truck move or push forward. Characteristics of student responses are:  
- The objects in the truck are still in motion when the truck stops  
- When the truck stops it takes time for the load to stop  
- The stuff in the truck keeps moving forward  
- More force in the truck.  
- The weight pushes the truck/the objects push the truck |
| 5           | No evidence of reasoning about the task | ➢ blank, nothing is written in response area  
➢ I don’t know confessed  
➢ response not meaningful, joking  
➢ no reasoning, ex. Just because |
| 6           | Not able to judge/High Inference        | ➢ can’t understand response (language, grammar)  
➢ Inaccurate vocabulary use  
➢ not enough information (one word/incomplete answers)  
➢ response repeats given information  
➢ undocumented/unique explanation/conception (gravity pushes truck)  
➢ too much inference required to judge (not enough friction to stop)  
➢ student uses multiple contradictory ideas/categories/responses to explain  
➢ Response provides irrelevant (possibly accurate) information |
Rating Guide for Gravity and a Tennis Ball
and Sample Student Responses

Gravity and a Tennis Ball: Rating Guide

Rationale

The task contains four elements for raters to record information about. There is one selected response item and three constructed response items. The selected response item asks students if the speed of the tennis ball will increase, decrease or stay the same as it falls. The first constructed response item asks students to explain why it will increase, decrease or stay the same. The second constructed response item asks students what caused the ball to change direction from going up to going down when Maria throws it up in the air. The third constructed response item asks students what Maria has to do the ball to get it up to Katie. Each item is to be rated independently and a score will be constructed from the independent ratings in a manner consistent with the assessment purpose.

Ratings –

Use these criteria for rating the selected response for Gravity and a Tennis Ball (Part 1a)

1 = student fills in the circle above “Speed increases.”
2 = student fills in the circle above “Speed decreases.”
3 = student fills in the circle above “Speed stays the same.”
5 = student does not fill in any circle (blank).
6 = student fills in multiple circles

Use these criteria for rating the first constructed response item for Gravity and a Tennis Ball (Part 1b). The focus for rating constructed response items is on identifying the characteristics of a student’s reasoning.

1 = rationale or explanation provided for “Force (Gravity)”
2 = rationale or explanation provided for “Alternative Conceptions of Force”
3 = rationale or explanation provided for “Initial Force”
4 = rationale of explanation provided for “Property of the Ball”
5 = No evidence of reasoning about the task (blank, I don’t know confession, joking response)
6 = Not Able to Judge/High Inference (can’t understand response, inaccurate vocabulary use, not enough information, response repeats given information, too much inference required to judge)
Use these criteria for rating the second constructed response item for Gravity and a Tennis Ball (Part IIa). The focus for rating constructed response items is on identifying the characteristics of a student’s reasoning.

1 = rationale or explanation provided for “Force (Gravity)”
2 = rationale or explanation provided for “Alternative Conceptions of Force”
3 = rationale or explanation provided for “Initial Force”
4 = rationale or explanation provided for “Property of the Ball”
5 = No evidence of reasoning about the task (blank, I don’t know confession, joking response)
6 = Not Able to Judge/High Inference (can’t understand response, inaccurate vocabulary use, not enough information, response repeats given information, too much inference required to judge)

Use these criteria for rating the third constructed response item for Gravity and a Tennis Ball (Part IIb). The focus for rating constructed response items is on identifying the characteristics of a student’s reasoning.

1 = rationale or explanation provided for “Applied Force Increases”
2 = rationale or explanation provided for “Force Given to the Ball”
3 = rationale or explanation provided for “Quality of the Path of the Ball”
5 = No evidence of reasoning about the task (blank, I don’t know confession, joking response)
6 = Not Able to Judge/High Inference (can’t understand response, inaccurate vocabulary use, not enough information, response repeats given information, too much inference required to judge)

**Turn to the next three pages for the rating guide charts for Gravity and a Tennis Ball.**
### Gravity and a Tennis Ball Part Ib
#### Rating Category Descriptions

<table>
<thead>
<tr>
<th>Rating Code</th>
<th>Inference</th>
<th>Description</th>
</tr>
</thead>
</table>
| 1 | **Force (Gravity)**<br>(9,11,13,24,31,32,40,41,43,45,53,54,55,61,64) | Student must fill in bubble above “Increase” and write about a force (gravity/power/pressure/momentum) pulling the ball down. Characteristics of student responses are:  
- Because gravity is pulling it down (with a lot of force)/gravity pulls things to the ground/Earth’s surface  
- Gravity (this response alone explaining the “Increase” is taken as evidence of a force changing the speed).  
- The speed increases because of (power of) gravity/momentum.  
- Gravity is making it go down faster.  
- Gravity’s force pulls the ball toward the (center of the) Earth.  
- Gravity pulls on it and the longer gravity pulls on it, every second it grows faster.  
- Since gravity pulls everything to the center of the Earth  
- The force of gravity pulled the ball down making it go down faster.  
- Gravity caused the speed of the ball to increase. |
| 2 | **Alternative Conceptions of Force**<br>(2,6,17,24,46,47,48,59,62,49) | Student fills in bubble above “Increase” and write about a) the force gravity changing (increasing, pulling harder/faster/stronger) as the ball falls toward the ground b) student writes about gravity pushing the ball down to the ground or c) an object possessing a force. Characteristics of student responses are:  
- Gravity is pulling down harder/faster.  
- Gravity is pulling the ball down harder and harder and faster faster  
- Gravity pushes against the ball/pushes the ball down.  
- Speed stays the same because gravity pulls on it making it the same.  
- Pull of gravity stays the same/is constant so the tennis ball doesn’t change.  
- Stays the same because gravity is not pulling harder/same amount of gravity.  
- Stays the same because there is no pressure/other force to make it slower or faster.  
- Gravity’s force pulls the ball down harder/faster.  
- Gravity is making it go down faster.  
- Since gravity pulls everything to the center of the Earth  
- The force of gravity pulled the ball down making it go down faster.  
- Gravity caused the speed of the ball to increase. |
| 3 | **Initial Force** | Student writes about some aspect of the initial force used to make the ball move. |
| 4 | **Property of the Ball** | Student writes about some property of the ball causing the speed of the ball to increase, decrease or stay the same as it falls. Characteristics of student responses are:  
- The tennis ball builds up velocity/picks up speed  
- The weight of the ball pulls/pushes it down  
- The ball gains more momentum/velocity as it falls.  
- Because the ball is light/Because of the air in the ball.  
- The ball is heavy so the speed gets faster.  
- The more it falls, the heavier it gets. |
| 5 | **No evidence of reasoning about the task** |  
- blank, nothing is written in response area  
- I don’t know confessed  
- response not meaningful, joking  
- no reasoning, ex. Just because |
| 6 | **Not Able to Judge/High Inference** |  
- can’t understand response (language, grammar)  
- Inaccurate vocabulary use  
- not enough information (one word/incomplete answers) |
<table>
<thead>
<tr>
<th>Rating Code</th>
<th>Inference Not Able to Judge/High Inference</th>
<th>Gravity and a Tennis Ball Part IIa</th>
<th>Rating Category Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gravity and a force (gravity) changing (increasing, pulling harder/faster/stronger) as the ball falls toward the ground, student writes about gravity ‘pushing’ the ball down to the ground or c) an object possessing a force. Characteristics of student responses are:</td>
<td>student writes about a) the force (gravity) changing (increasing, pulling harder/faster/stronger) as the ball falls toward the ground, b) student writes about gravity ‘pushing’ the ball down to the ground or c) an object possessing a force. Characteristics of student responses are:</td>
<td>Gravity is pushing down harder/faster.</td>
</tr>
<tr>
<td>2</td>
<td>Alternative Conceptions of Force (3,19,23,31,42,48,55,)</td>
<td>High Inference Force (Gravity) (2,4,6,8,9,10,11,12,14,16,17,18,21,22,24,25,26,30,33,34,37,38,40,41,43,44,45,46,47,49,51,52,57,59,62,64,66)</td>
<td>Gravity forces/pulls it down/gravity pulls all objects to center of Earth.</td>
</tr>
<tr>
<td>3</td>
<td>Initial Force (13,27,28,36,50,58,60,61)</td>
<td>Student writes only about there not being enough initial force (applied by throw) as the reason for why the ball changes direction. Characteristics of student responses are:</td>
<td>There wasn’t enough power to get it to Katie Tennis ball won’t go to Maria unless Katie throws harder/didn’t throw it hard enough</td>
</tr>
<tr>
<td>4</td>
<td>Property of the Ball (1, 20, 35,53,56,63)</td>
<td>Student writes about some property of the ball causing the direction of the ball to change from going up to going down. Characteristics of student responses are:</td>
<td>Because of the weight of the ball pushed it down/the ball is heavy Because the ball bounces When it falls, the faster it will go/ it goes back because of the speed The ball is rotating There is air/no air in the ball.</td>
</tr>
<tr>
<td>5</td>
<td>No evidence of reasoning about the task</td>
<td></td>
<td>blank, nothing is written in response area I don’t know/confessed response not meaningful, joking no reasoning, ex. just because</td>
</tr>
<tr>
<td>6</td>
<td>Not Able to Judge/High Inference (7,29,32,39,)</td>
<td></td>
<td>can’t understand response (language, grammar) Inaccurate vocabulary use not enough information (one word/incomplete answers) undocumented/unique explanation/conception (inertia, the wind/air) response repeats given information too much inference required to judge (the throw is straight up) student uses multiple contradictory ideas/categories/responses to explain Response provides irrelevant (possibly accurate) information</td>
</tr>
<tr>
<td>Rating Code</td>
<td>Inference</td>
<td>Gravity and a Tennis Ball IIb Rating Category Descriptions</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-----------</td>
<td>------------------------------------------------------------</td>
<td></td>
</tr>
</tbody>
</table>
| 1           | **Applied Force Increases** (1,10,13,16,17,18,19, 20,21,22,23,25,27, 28,29,30,32,38, 41,42,43,44,45,48,50,51, 52,58,59,60,61,64,66) | Student writes about having to throw/bounce the ball with more force/strength/pressure/power/energy or throw it harder in order to have the ball make it all the way up to Katie. Explanations such as, *go inside and climb the stairs*, are acceptable in addition to ‘applied force increases’ responses. Characteristics of student responses include:  
  ➢ Throw/bounce the ball hard/harder or with more force  
  ➢ Apply more/put more/put a harder/use more/give a lot of/increase the amount of force on/to the ball.  
  ➢ She has to apply enough/certain amount of force to overcome the force of gravity long enough for the ball to reach Katie  
  ➢ Needs to use **more force and strength** for the ball to bounce higher  
  ➢ Use a lot of power/use more strength  
  *not all “bounce” answers show evidence of understanding* |
| 2           | **Force Given to the Ball** (2,11,36,40,59) | Student writes about how there would be a force given to the ball. *Students tend to think of force as a property of an object.* Characteristics of student responses are:  
  ➢ More force given to the ball  
  ➢ Give a lot of/more force to the tennis ball  
  ➢ Put more force into the tennis ball  
  *see research base on the Concept of Force (Benchmarks, p. 339)* |
| 3           | **Quality of the Path of the Ball** (3,4,6,8,12,14,24,34,46,49,54,55,56,57,62) | Student writes about throwing/bouncing the ball higher/straighter/further/again to Katie (without referencing a kind of increase in force). Characteristics of student responses are:  
  ➢ Maria has to throw/bounce the ball up/at/straight to her  
  ➢ Maria has to throw/bounce the ball higher  
  ➢ She needs to aim at Katie then throw the ball so it will reach  
  ➢ Throw it in an arc  
  ➢ She needs to throw it so Katie can catch it |
| 5           | **No evidence of reasoning about the task** | ➢ blank, nothing is written in response area  
  ➢ I don’t know confessed  
  ➢ response not meaningful, joking  
  ➢ no reasoning, ex. Just because |
| 6           | **Not able to judge/High Inference** (7,15,26,31,33,37,42,65) | ➢ can’t understand response (language, grammar)  
  ➢ Inaccurate vocabulary use  
  ➢ not enough information (one word/incomplete answers response repeats given information  
  ➢ undocumented/unique explanation/conception (take it upstairs, go inside and give it to her)  
  ➢ too much inference required to judge (Katie [not Maria] throws/drops the ball)  
  ➢ student uses multiple **contradictory** ideas/categories/responses to explain  
  ➢ Response provides irrelevant (possibly accurate) information |
Rating Guide for A Rolling Ball and Sample Student Responses

A Rolling Ball: Rating Guide

Rationale

The task contains **three elements** for raters to record information about. There is one selected response item and two constructed response items. The selected response item asks students if there is a force constantly pushing or pulling the ball as it rolls toward you. The first constructed response item asks students to explain their selected response. The second constructed response item asks students what caused the ball to stop. Each item is to be rated independently and a score will be constructed from the independent ratings in a manner consistent with the assessment purpose.

Ratings –

Use these criteria for rating the selected response for A Rolling Ball (Part 1a)

1 = student fills in the circle above “Yes”  
2 = student fills in the circle above “No”  
5 = student does not fill either circle (Blank)  
6 = student fills in multiple circles

Use these criteria for rating the first constructed response item for A Rolling Ball (Part 1b). The focus for rating constructed response items is on identifying the characteristics of a student’s reasoning.

1 = rationale or explanation provided for “Initial Force/Inertia”  
2 = rationale or explanation provided for “Alternative Conception of Force”  
3 = rationale or explanation provided for “Initial Force”  
4 = rationale of explanation provided for “Property of the Ball”  
5 = No evidence of reasoning about the task (blank, I don’t know confession, joking response)  
6 = Not Able to Judge/High Inference (can’t understand response, inaccurate vocabulary use, not enough information, response repeats given information, too much inference required to judge)  
7 = rationale or explanation provided for “Friction”
Use these criteria for rating the first constructed response item for A Rolling Ball (Part II). The focus for rating constructed response items is on identifying the characteristics of a student’s reasoning.

1 = rationale or explanation provided for “Forces Change Motion”
2 = rationale or explanation provided for “Alternative Conception of Force”
3 = rationale or explanation provided for “Initial Force”
4 = rationale of explanation provided for “Property of the Ball”
5 = No evidence of reasoning about the task (blank, I don't know confession, joking response)
6 = Not Able to Judge/High Inference (can't understand response, inaccurate vocabulary use, not enough information, response repeats given information, too much inference required to judge)

Turn to the next two pages for the rating guide charts for A Rolling Ball.
<table>
<thead>
<tr>
<th>Rating Code</th>
<th>Inference</th>
<th>Rating Category Descriptions</th>
</tr>
</thead>
</table>
| 1           | Initial Force/Inertia (Constant Force=No) (9,11,34,41,55,57,28,19,52,40,29) | Student must fill in bubble above “No” and write about no force acting on the ball after the initial force (kick, push, roll). Student may describe the initial pushing force. Characteristics of student responses are:  
- No force is constantly pushing but one force started the ball toward you.  
- The ball has inertia.  
- There is no force after the push from your friend.  
- My friend only gave one push then let it roll/ friend kicked the ball once.  
- No one is holding/pushing it the whole way, they gave it one big push.  
- There is no force pushing the ball anymore.  
- Once the ball is moving, there is no longer force.  
- It does not constantly be pushed or pulled, it is just rolling to another person. |
| 2           | Alternative Conceptions of Force (46,51,61) | Student writes about forces being transferred into objects or gravity ‘pushing’ the ball toward you. Characteristics of student responses are:  
- Gravity pushes the ball to the other guy.  
- Gravity helps push it to make it go faster.  
- Force/momentum from kick/push/initial force is transferred into the ball.  
- Once you push the ball, the ball will have force in it from pushing it.  
- (Yes is filled in) and student writes about Newton’s First Law. |
| 3           | Initial Force (Constant Force=Yes) (1,3,4,5,8,10,35,38,42,45,48,53,58,50,65,66) | Student must fill in bubble above “Yes” and the student writes about a force (friends foot, kick, roll) staying with the ball or slowly running out as it rolls. Characteristics of student responses are:  
- His friend is pushing the ball to him.  
- Someone is pushing/rolling the ball.  
- The force that pushed it is still there.  
- The friend pushed it, that makes the ball move. |
| 4           | Property of the Ball | Student writes about some property of the ball to explain why there is or is not a force acting on the ball. Characteristics of student responses are:  
- It pushes or pulls itself.  
- The ball is round and has no edges.  
- The ball is going fast. |
| 5           | No evidence of reasoning about the task | blank, nothing is written in response area  
- I don’t know confessed  
- response not meaningful, joking  
- no reasoning, ex. Just because |
| 6           | Friction (12,13,14,15,27,39,4 Not Able to Judge/ High Inference (12,15) | Student must fill in bubble above “Yes” (and write in) and states that a force (friction) constantly acting on the ball (friends shoe, rolling surface of ground) is why the ball is slowing down. Characteristics of student responses are:  
- friction is causing the ball to slow down  
- friction is not mentioned but the student describes it as a force (too much friction, ball has to stop)  
- friction is pulling on the ball (tear/stop ball)  
- student uses multiple contradictory ideas/categories/responses to explain  
- Selected response inconsistent with constructed response (selects no, describes friction)  
- Response provides irrelevant (possibly accurate) information (gravity pulls down). |
| 7           | Friction (12,13,14,15,27,39,4 Not Able to Judge/ High Inference (12,15) | Student must fill in bubble above “Yes” (and write in) and states that a force (friction) constantly acting on the ball (friends shoe, rolling surface of ground) is why the ball is slowing down. Characteristics of student responses are:  
- friction is causing the ball to slow down  
- friction is not mentioned but the student describes it as a force (too much friction, ball has to stop)  
- friction is pulling on the ball (tear/stop ball)  
- student uses multiple contradictory ideas/categories/responses to explain  
- Selected response inconsistent with constructed response (selects no, describes friction)  
- Response provides irrelevant (possibly accurate) information (gravity pulls down). |
<table>
<thead>
<tr>
<th>Rating Code</th>
<th>Inference</th>
<th>A Rolling Ball Part II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Rating Category Descriptions</td>
</tr>
<tr>
<td>1</td>
<td><strong>Forces Change Motion</strong></td>
<td>Student writes that because of force/friction/air resistance, the ball stops or because something got in its way and caused it to stop. Characteristics of student responses are:</td>
</tr>
<tr>
<td></td>
<td>(2,3,5,12,13,15,21,25,</td>
<td>➢ Friction.</td>
</tr>
<tr>
<td></td>
<td>27,29,30,39,41,45,49,</td>
<td>➢ Friction caused the ball to (slow down and then) stop.</td>
</tr>
<tr>
<td></td>
<td>58,63,66)</td>
<td>➢ Friction from the ball dragging on the floor, rubbing against the ground.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ Friction was pulling the ball.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ The ball stopped because of air/wind hitting the ball and gradually making it slow down.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ The ball stops when something gets in its way.</td>
</tr>
<tr>
<td>2</td>
<td><strong>Alternative Conception of Force</strong></td>
<td>Student writes about the force/energy/momentum being within the ball and/or running out/being finished or because there is no more force/energy/momentum in the ball (ball needs force to move). Characteristics of student responses are:</td>
</tr>
<tr>
<td></td>
<td>(1,4,7,9,10,18,22,26,48,</td>
<td>➢ If you don’t keep pushing the ball, it starts to slow down.</td>
</tr>
<tr>
<td></td>
<td>61,57)</td>
<td>➢ The ball stopped because the force/energy of the kick ran out/depleting force.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ The ball wasn’t being pushed by a force anymore/ there is no more force/force stops/energy from the kick is now finished.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ The first force stopped having affect or</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Student writes about the idea that constant motion requires constant force (force=motion or no force=no motion).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ If there was force, the ball would keep rolling.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ The lack of force caused the ball to stop.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ If you don’t keep pushing, the ball will slow down/stop.</td>
</tr>
<tr>
<td>3</td>
<td><strong>Initial Force</strong></td>
<td>Student writes about the initial force of the ball not being (hard) enough to keep the ball moving. Characteristics of student responses are:</td>
</tr>
<tr>
<td></td>
<td>(14,28,31,32,34,35,42,4</td>
<td>➢ The amount of force you used is what caused the ball to stop. If you used more/less force, the ball would travel more/less distance.</td>
</tr>
<tr>
<td></td>
<td>4,56,55,53,50)</td>
<td>➢ He didn’t kick/roll it hard enough.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ There wasn’t enough force to push it all the way.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ The amount of force that the other person used.</td>
</tr>
<tr>
<td>4</td>
<td><strong>Property of Ball</strong></td>
<td>Student writes about some property of the ball as the reason it stops.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ Ball loses speed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ Weight of the ball</td>
</tr>
<tr>
<td>5</td>
<td><strong>No evidence of reasoning about the task</strong></td>
<td>Student writes about the lack of evidence for the task.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ blank, nothing is written in response area</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ I don’t know confessed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ response not meaningful, joking</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ no reasoning, ex. Just because</td>
</tr>
<tr>
<td>6</td>
<td><strong>Not Able to Judge/High Inference</strong></td>
<td>Student writes that because of force/friction/air resistance, the ball stops or because something got in its way and caused it to stop. Characteristics of student responses are:</td>
</tr>
<tr>
<td></td>
<td>(6,8,60,</td>
<td>➢ can’t understand response (language, grammar)</td>
</tr>
<tr>
<td></td>
<td>46,47,37,36,17,11,16,19</td>
<td>➢ Inaccurate vocabulary use (Ball ran out of inertia)</td>
</tr>
<tr>
<td></td>
<td>,20,23,24,33,38,40,43,5</td>
<td>➢ not enough information (one word/incomplete answers)</td>
</tr>
<tr>
<td></td>
<td>1,52,54,59,62,64,65)</td>
<td>➢ undocumented/unique explanation/conception</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ response repeats given information</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ too much inference required to judge (gravity caused ball to stop)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ student uses multiple contradictory ideas/categories/responses to explain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>➢ Response provides irrelevant (possibly accurate) information (gravity pulls ball to ground)</td>
</tr>
</tbody>
</table>
Appendix D: Initial Item Characteristics Scores for Questions About Motion Test

* Those scores highlighted in yellow are calculated outside of Coh-Metrix. Most of the linguistic complexity variables are hypothesized to have a negative relationship with students' item-level scores; however, those measures which are highlighted in green are hypothesized to have a positive relationship with students' item-level scores.

**Initial Item Format Variables Calculated in Coh-Metrix**

<table>
<thead>
<tr>
<th>ITEM</th>
<th>Target Idea</th>
<th>Selected/Cons</th>
<th># words</th>
<th># sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF1</td>
<td>2</td>
<td>0</td>
<td>66</td>
<td>9</td>
</tr>
<tr>
<td>MF2</td>
<td>2</td>
<td>1</td>
<td>87</td>
<td>10</td>
</tr>
<tr>
<td>MF3</td>
<td>3</td>
<td>1</td>
<td>60</td>
<td>7</td>
</tr>
<tr>
<td>MF4</td>
<td>1</td>
<td>0</td>
<td>62</td>
<td>6</td>
</tr>
<tr>
<td>MF5</td>
<td>1</td>
<td>1</td>
<td>79</td>
<td>7</td>
</tr>
<tr>
<td>MF6</td>
<td>1</td>
<td>1</td>
<td>70</td>
<td>7</td>
</tr>
<tr>
<td>MF7</td>
<td>2</td>
<td>1</td>
<td>71</td>
<td>7</td>
</tr>
<tr>
<td>MF8</td>
<td>4</td>
<td>0</td>
<td>56</td>
<td>8</td>
</tr>
<tr>
<td>MF9</td>
<td>4</td>
<td>1</td>
<td>60</td>
<td>9</td>
</tr>
<tr>
<td>MF10</td>
<td>1</td>
<td>1</td>
<td>37</td>
<td>6</td>
</tr>
</tbody>
</table>

**Initial Item Format Variables Calculated outside of Coh-Metrix**

<table>
<thead>
<tr>
<th>ITEM</th>
<th>layout/visuals</th>
<th>distance pict. Text?</th>
<th>text stem and prompt distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>MF2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>MF3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>MF4</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>MF5</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>MF6</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>MF7</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>MF8</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>MF9</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>MF10</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
**Initial Item Syntactic Complexity Variables Calculated in Coh-Metrix**

<table>
<thead>
<tr>
<th>ITEM</th>
<th>sentence length</th>
<th># words before main verb</th>
<th>logical operators</th>
<th># modifiers per no</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF1</td>
<td>7.33</td>
<td>0.67</td>
<td>15.15</td>
<td>0.80</td>
</tr>
<tr>
<td>MF2</td>
<td>8.70</td>
<td>0.70</td>
<td>22.99</td>
<td>0.85</td>
</tr>
<tr>
<td>MF3</td>
<td>8.57</td>
<td>0.86</td>
<td>0.00</td>
<td>0.88</td>
</tr>
<tr>
<td>MF4</td>
<td>10.33</td>
<td>1.50</td>
<td>48.39</td>
<td>0.79</td>
</tr>
<tr>
<td>MF5</td>
<td>11.29</td>
<td>1.43</td>
<td>50.63</td>
<td>0.83</td>
</tr>
<tr>
<td>MF6</td>
<td>10.00</td>
<td>2.14</td>
<td>42.86</td>
<td>0.85</td>
</tr>
<tr>
<td>MF7</td>
<td>10.14</td>
<td>2.43</td>
<td>42.25</td>
<td>0.90</td>
</tr>
<tr>
<td>MF8</td>
<td>7.00</td>
<td>1.13</td>
<td>53.57</td>
<td>0.94</td>
</tr>
<tr>
<td>MF9</td>
<td>6.67</td>
<td>1.11</td>
<td>50.00</td>
<td>0.90</td>
</tr>
<tr>
<td>MF10</td>
<td>6.17</td>
<td>1.50</td>
<td>27.03</td>
<td>1.20</td>
</tr>
</tbody>
</table>

**Initial Lexical Complexity Variables Calculated in Coh-Metrix**

<table>
<thead>
<tr>
<th>ITEM</th>
<th>syllables per word</th>
<th>log avg. frequency of words</th>
<th>noun hypernyms</th>
<th>verb hypernyms</th>
<th>concreteness</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF1</td>
<td>1.33</td>
<td>1.93</td>
<td>7.57</td>
<td>1.94</td>
<td>248</td>
</tr>
<tr>
<td>MF2</td>
<td>1.32</td>
<td>1.95</td>
<td>7.16</td>
<td>1.95</td>
<td>248</td>
</tr>
<tr>
<td>MF3</td>
<td>1.17</td>
<td>2.32</td>
<td>8.12</td>
<td>1.75</td>
<td>225</td>
</tr>
<tr>
<td>MF4</td>
<td>1.32</td>
<td>2.10</td>
<td>4.66</td>
<td>1.42</td>
<td>248</td>
</tr>
<tr>
<td>MF5</td>
<td>1.30</td>
<td>2.04</td>
<td>4.59</td>
<td>1.49</td>
<td>248</td>
</tr>
<tr>
<td>MF6</td>
<td>1.33</td>
<td>2.33</td>
<td>4.79</td>
<td>1.36</td>
<td>287</td>
</tr>
<tr>
<td>MF7</td>
<td>1.30</td>
<td>2.38</td>
<td>4.76</td>
<td>1.53</td>
<td>251</td>
</tr>
<tr>
<td>MF8</td>
<td>1.27</td>
<td>2.31</td>
<td>5.14</td>
<td>1.96</td>
<td>216</td>
</tr>
<tr>
<td>MF9</td>
<td>1.28</td>
<td>2.30</td>
<td>4.9</td>
<td>2.08</td>
<td>216</td>
</tr>
<tr>
<td>MF10</td>
<td>1.19</td>
<td>2.22</td>
<td>5.28</td>
<td>1.71</td>
<td>256</td>
</tr>
</tbody>
</table>

**Initial Discourse Complexity Variables Calculated in Coh-Metrix**

<table>
<thead>
<tr>
<th>ITEM</th>
<th>pos &amp; neg connectives</th>
<th>syntactic repetition</th>
<th>content word overlap</th>
<th>causal cohesion</th>
<th>intentional cohesion</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF1</td>
<td>30.3</td>
<td>0.31</td>
<td>0.26</td>
<td>0.00</td>
<td>30.3</td>
</tr>
<tr>
<td>MF2</td>
<td>45.98</td>
<td>0.29</td>
<td>0.24</td>
<td>0.00</td>
<td>34.48</td>
</tr>
<tr>
<td>MF3</td>
<td>66.67</td>
<td>0.08</td>
<td>0.28</td>
<td>0.67</td>
<td>33.33</td>
</tr>
<tr>
<td>MF4</td>
<td>48.39</td>
<td>0.04</td>
<td>0.2</td>
<td>0.00</td>
<td>32.26</td>
</tr>
<tr>
<td>MF5</td>
<td>63.29</td>
<td>0.06</td>
<td>0.22</td>
<td>0.2</td>
<td>37.98</td>
</tr>
<tr>
<td>MF6</td>
<td>71.43</td>
<td>0.06</td>
<td>0.14</td>
<td>0.2</td>
<td>42.86</td>
</tr>
<tr>
<td>MF7</td>
<td>56.34</td>
<td>0.06</td>
<td>0.21</td>
<td>0.00</td>
<td>42.25</td>
</tr>
<tr>
<td>MF8</td>
<td>53.57</td>
<td>0.1</td>
<td>0.09</td>
<td>0.00</td>
<td>17.86</td>
</tr>
<tr>
<td>MF9</td>
<td>50</td>
<td>0.1</td>
<td>0.08</td>
<td>0.00</td>
<td>33.33</td>
</tr>
<tr>
<td>MF10</td>
<td>81.08</td>
<td>0.08</td>
<td>0.12</td>
<td>0.2</td>
<td>27.03</td>
</tr>
</tbody>
</table>
References


Functions of language in the classroom (pp. 370–394). New York: Teachers College Press.


Wright, L. J., & Logan-Terry, A. (In preparation). The wood does not - it will stay: Identifying trouble sources in English learners’ interactions with traditional and multisemiotic science assessment items. Assessment in Education.
