

University of Massachusetts Boston

## ScholarWorks at UMass Boston

---

Graduate Doctoral Dissertations

Doctoral Dissertations and Masters Theses

---

6-1-2015

# Analysis of English Language Learner Performance on the Biology Massachusetts Comprehensive Assessment System: The Impact of English Proficiency, First Language Characteristics, and Late-entry ELL Status

Mary A. Mitchell

*University of Massachusetts Boston*

Follow this and additional works at: [https://scholarworks.umb.edu/doctoral\\_dissertations](https://scholarworks.umb.edu/doctoral_dissertations)



Part of the [Bilingual, Multilingual, and Multicultural Education Commons](#), [Educational Assessment, Evaluation, and Research Commons](#), [First and Second Language Acquisition Commons](#), and the [Science and Mathematics Education Commons](#)

---

### Recommended Citation

Mitchell, Mary A., "Analysis of English Language Learner Performance on the Biology Massachusetts Comprehensive Assessment System: The Impact of English Proficiency, First Language Characteristics, and Late-entry ELL Status" (2015). *Graduate Doctoral Dissertations*. 211.  
[https://scholarworks.umb.edu/doctoral\\_dissertations/211](https://scholarworks.umb.edu/doctoral_dissertations/211)

This Open Access Dissertation is brought to you for free and open access by the Doctoral Dissertations and Masters Theses at ScholarWorks at UMass Boston. It has been accepted for inclusion in Graduate Doctoral Dissertations by an authorized administrator of ScholarWorks at UMass Boston. For more information, please contact [scholarworks@umb.edu](mailto:scholarworks@umb.edu).

ANALYSIS OF ENGLISH LANGUAGE LEARNER PERFORMANCE ON THE  
BIOLOGY MASSACHUSETTS COMPREHENSIVE ASSESSMENT SYSTEM: THE  
IMPACT OF ENGLISH PROFICIENCY, FIRST LANGUAGE CHARACTERISTICS,  
AND LATE-ENTRY ELL STATUS

A Dissertation Presented

by

MARY A. MITCHELL

Submitted to the Office of Graduate Studies,  
University of Massachusetts Boston,  
in partial fulfillment of the requirements for the degree of

DOCTOR OF EDUCATION

June 2015

Leadership in Urban Schools Program

© 2015 Mary A. Mitchell  
All rights reserved

ANALYSIS OF ENGLISH LANGUAGE LEARNER PERFORMANCE ON THE  
BIOLOGY MASSACHUSETTS COMPREHENSIVE ASSESSMENT SYSTEM: THE  
IMPACT OF ENGLISH PROFICIENCY, FIRST LANGUAGE CHARACTERISTICS,  
AND LATE-ENTRY ELL STATUS

A Dissertation Presented

by

MARY A. MITCHELL

Approved as to style and content by:

---

Wenfan Yan, Professor  
Chairperson of Committee

---

Zeena Zakharia, Assistant Professor  
Member

---

Jack Levy, Professor  
Member

---

Dr. Tricia Kress, Graduate Program Director  
Leadership in Urban Schools Program

---

Dr. Wenfan Yan, Chairperson  
Leadership in Urban Schools Program

## ABSTRACT

# ANALYSIS OF ENGLISH LANGUAGE LEARNER PERFORMANCE ON THE BIOLOGY MASSACHUSETTS COMPREHENSIVE ASSESSMENT SYSTEM: THE IMPACT OF ENGLISH PROFICIENCY, FIRST LANGUAGE CHARACTERISTICS, AND LATE-ENTRY ELL STATUS

June 2015

Mary A. Mitchell, A.B., Wellesley College  
M.B.A., Boston University  
M.A.T., Salem State University  
Ed.D., University of Massachusetts Boston

Directed by Professor Wenfan Yan

This study analyzed English language learner (ELL) performance on the June 2012 Biology MCAS, namely on item attributes of domain, cognitive skill, and linguistic complexity. It examined the impact of English proficiency, Latinate first language, first language orthography, and late-entry ELL status. The results indicated that English proficiency was a strong predictor of performance and that ELLs at higher levels of English proficiency overwhelmingly passed. The results further indicated that English proficiency introduced a construct-irrelevant variance on the Biology MCAS and raised validity issues for using this assessment at lower levels of English proficiency. This study

also found that ELLs with a Latinate first language consistently had statistically significant lower performance. Late-entry ELL status did not predict Biology MCAS performance.

## DEDICATION

I would like to thank my father and my aunt for their encouragement, support, and belief in me as I embarked on a new chapter in my life.

.

## PREFACE

Mid-career, I left the corporate world to teach English as a second language, and I have spent the last decade working in urban high schools in gateway communities in the Boston metropolitan area. My undergraduate degree in molecular biology landed me in sheltered English instruction (SEI) science content classrooms teaching English language learners (ELLs). I have observed first-hand the vast differences between ELLs who enter our schools at the secondary level. Some enter with grade-level academic language and literacy in their first language, and others enter with minimal or interrupted schooling. Minimal schooling includes never having been in a formal learning environment, while interrupted schooling includes students who finished up to Grade 6 in their country but who have been out of school for several years before arriving in the United States. Each ELL has strengths and weaknesses unique to the intersection of innate capabilities, background experiences, background knowledge, culture, and experiences in the United States. ELLs who enter U.S. schools at age 12 or older are the poster children for the heterogeneity of the ELL subgroup. But they have one thing in common—they all struggle to reach grade-level English proficiency in the short time high school allots them.

Realizing the disconnect between research that suggests it takes four to seven years to acquire academic language and ELLs who enter in high school and are expected to pass high-stakes content tests in English, I decided to become part of the process to see how late-entry ELLs might succeed when the odds were stacked against them. Thus began my involvement with the Massachusetts Department of Elementary and Secondary



Education (MA DESE) and its statewide assessments. For three years, I was a member of the assessment development committee (ADC) for the Massachusetts English Proficiency Assessment (MEPA), a statewide assessment given yearly to all ELLs in Massachusetts public schools. I also served as a member of the 2005 MEPA Standards Setting Panel. I have also been involved for the past eight years with the Biology Massachusetts Comprehensive Assessment System (MCAS), serving on that assessment's ADC. My work on the Biology MCAS ADC led to my appointment to the MCAS Performance Appeals Panel, which decides if other evidence of student academic performance justifies granting an exemption from the graduation requirement of passing the MCAS in that content area. I have also been involved with Massachusetts assessments for licensure, having served on the 2008-2009 Objective Review and Qualifying Score Committees for the Massachusetts Test for Educator Licensure (MTEL) in Earth Science. Most recently, I was the science content item reviewer from Massachusetts and a member of the Bias and Sensitivity Review panel for ACCESS 2.0, the English proficiency assessment given to ELLs in the more than 30 states and territories that are part of the WIDA consortium.

As an SEI biology teacher, I strove not only for my students to learn biology content but also to be prepared to demonstrate their knowledge on the Biology MCAS. My work with the Biology MCAS development process informed my instruction. Specifically, I recognized the need for parallel academic language development in my SEI biology classroom, including the development of academic literacy and discourse expected of native speakers in the content area of biology. I also recognized that my students needed foundational and definitional knowledge, but, moreover, they needed to recognize science concepts when presented in unfamiliar contexts. My ultimate goal was

to enable students to synthesize knowledge from different domains of biology and situate it in overarching themes so that they could constructively problem-solve.

In efforts to discover an effective pedagogy to help ELLs succeed on the Biology MCAS, I have analyzed my students' MCAS scores and the linguistic demands of the Biology MCAS, and implemented the explicit teaching of academic language and the discourse of the Biology MCAS. I have seen many students successfully pass the MCAS, often before they pass the English Language Arts MCAS or the Mathematics MCAS and often without having completed my school's two-year biology curriculum. My classroom experiences teaching biology to ELLs and my experiences in the development process of the Biology MCAS have culminated in my doctoral research.

## TABLE OF CONTENTS

DEDICATION .....	vi
PREFACE .....	vii
LIST OF TABLES .....	xvi
LIST OF FIGURES .....	xx
CHAPTER	Page
1. INTRODUCTION .....	1
Problem Statement .....	2
Background and Context.....	4
English language learners .....	4
Early-entry ELLs .....	4
Late-entry ELLs .....	5
ELL designation in Massachusetts.....	5
The MCAS .....	7
History.....	7
Who takes the MCAS?.....	8
The Biology MCAS .....	9
Pedagogy.....	10
ELL achievement gap .....	11
Research Purpose and Rationale .....	12
Research Questions .....	14
Significance of the Study .....	19
Definition of Terms.....	21
Limitations and Delimitations.....	24
Assumptions.....	25
Summary .....	27
2. LITERATURE REVIEW .....	29
Review of Second Language Acquisition Theories.....	31
Second language acquisition theoretical schools.....	31
Behaviorism .....	31
Innatism.....	31
Interactionism .....	32
Krashen's monitor model and five hypotheses .....	32
Critical period hypothesis .....	35
Late-entry ELLs .....	38
Review of Academic Language Literature .....	40
Cummins' model of BICS and CALP .....	41
Characteristics of academic language.....	44

CHAPTER	Page
Academic language and science .....	48
Academic language and ELLs .....	50
Academic language and achievement.....	53
Review of Literature on Critical Issues for Cognitive Load and	
Wide-Scale Assessments .....	54
Item difficulty .....	54
Cognitive complexity.....	55
Cognitive load.....	56
Cognitive load and second language learning .....	62
Differential item functioning .....	63
Review of Critical Issues for ELLs and High-Stakes Testing .....	65
Assessment of ELLs .....	65
Impact of linguistic complexity .....	68
Validity and reliability .....	72
Accommodations .....	75
Conceptual Framework.....	83
Summary .....	84
3. METHODS .....	87
Research Design and Data Sources.....	88
Research questions.....	90
Data sources .....	92
The Biology MCAS instrument.....	93
Reliability.....	93
Validity .....	94
The MEPA instrument .....	95
MEPA-R/W.....	95
MELA-O .....	96
Classical test theory .....	96
Item response theory .....	96
Reliability.....	97
Sample.....	97
Variables .....	98
Performance variables.....	98
Biology MCAS performance .....	99
English language proficiency (ELP).....	100
Demographic variables .....	101
First language characteristics.....	101
Age-related variables .....	103
Item attribute variables .....	104
Linguistic complexity .....	105
Domain.....	106
Cognitive skill level .....	107

CHAPTER	Page
Analysis Strategy—Phase I.....	107
Content domain.....	107
Cognitive skill level.....	108
Operationalization of item linguistic complexity.....	108
Stem lexical density.....	109
Stem syntax (SS).....	109
Total answer lexical density.....	110
Reading complexity score.....	110
Data preparation.....	112
Lexile Analyzer®.....	113
Data transformations and calculated variables.....	113
Calculated variables.....	113
Total lexical density.....	113
Answer lexical density.....	113
Stem syntactic density.....	113
Composite linguistic complexity computation.....	114
Statistical analyses.....	114
Analysis Strategy—Phase II.....	115
Data management.....	115
Data file preparation.....	115
Refining the data.....	115
Missing variables.....	116
Data transformations and calculated variables.....	117
Missing scaled MCAS scores.....	117
Item correct.....	117
Calculated item attribute performance.....	118
Age of entry variable.....	118
Statistical analyses.....	119
Biology MCAS performance.....	119
Impact of English proficiency.....	119
Impact of L1 family.....	120
Impact of L1 orthography.....	120
Impact of late-entry ELL status.....	121
Content domain performance.....	121
Impact of English proficiency.....	121
Impact of L1 family.....	121
Impact of L1 orthography.....	122
Impact of late-entry ELL status.....	122
Cognitive skill level performance.....	123
Impact of English proficiency.....	123
Impact of L1 family.....	123
Impact of L1 orthography.....	123
Impact of late-entry ELL status.....	124

CHAPTER	Page
Linguistic complexity performance .....	124
Impact of English proficiency.....	124
Impact of L1 family .....	125
Impact of L1 orthography .....	125
Impact of late-entry ELL status .....	125
4. RESULTS .....	127
Sample Characteristics.....	127
English proficiency level .....	132
First language characteristics .....	134
First language family .....	136
First language orthography .....	137
Late-entry ELLs .....	139
Summary of sample characteristics .....	141
Analysis of the MCAS Instrument.....	142
Content domains .....	143
Cognitive skill level .....	144
Linguistic complexity .....	146
Content domain linguistic complexity .....	147
Cognitive skill linguistic complexity .....	148
Summary of linguistic elements.....	152
Composite linguistic complexity .....	152
Summary of June 2012 MCAS instrument.....	156
ELL Performance on the June 2012 Biology MCAS .....	157
English proficiency impact .....	159
First language family impact .....	168
First language orthography impact .....	178
Late-entry ELL status impact.....	186
Summary of ELL MCAS performance.....	190
Performance on Content Domains .....	192
English proficiency impact .....	197
First language family impact .....	207
First language orthography impact .....	210
Late-entry ELL status impact.....	212
Summary of content domain performance.....	215
Performance at Cognitive Skill Levels .....	217
English proficiency impact .....	219
First language family impact .....	222
First language orthography impact .....	224
Late-entry ELL status impact.....	226
Summary of cognitive skill level performance .....	227
Performance on Item Linguistic Complexity Levels .....	228
English proficiency impact .....	231

CHAPTER	Page
First language family impact .....	238
First language orthography impact .....	241
Late-entry ELL status impact.....	243
Summary of linguistic complexity performance.....	244
Summary of Results .....	245
ELL Biology MCAS performance.....	245
English proficiency impact .....	246
First language family impact .....	248
First language orthography impact .....	249
Late-entry ELL status impact.....	250
5. DISCUSSION .....	251
Summary of Findings .....	252
Sample Characteristics.....	254
English proficiency .....	255
First languages .....	256
Late-entry ELL status .....	258
The June 2012 MCAS Instrument .....	258
ELL Performance .....	261
Content domain performance.....	262
Cognitive skill level performance .....	263
Linguistic complexity performance .....	265
English Proficiency Impact.....	267
When do ELLs have meaningful participation on the Biology MCAS?.....	268
English proficiency was a strong predictor of Biology MCAS performance .....	271
Impact on content domain performance .....	272
Impact on cognitive skill level performance.....	272
Impact on linguistic complexity performance .....	274
English proficiency was a construct-irrelevant variance .....	276
ELL Biology MCAS performance supported some aspects of cognitive load theory, the cognition hypothesis, and a Goldilocks effect .....	277
Cognitive load theory.....	278
The cognition hypothesis .....	280
The Goldilocks effect.....	282
First Language Characteristics Impact.....	282
Lower performance for ELLs with a Latinate L1 .....	283
Weak impact of L1 orthography .....	287
Late-Entry ELL Impact.....	290
Implications of the Study .....	291

CHAPTER	Page
Policy implications.....	292
Practice implications.....	298
Future research implications.....	302
Validity, Reliability, and Limitations .....	305
Threats to construct validity.....	305
Limitations .....	306
Summary .....	307
APPENDIX	
A. JUNE 2012 BIOLOGY MCAS .....	309
B. MEPA PERFORMANCE-LEVEL DESCRIPTORS .....	334
C. MASSACHUSETTS HIGH SCHOOL BIOLOGY STANDARDS.....	336
D. COGNITIVE AND QUANTITATIVE SKILLS DESCRIPTIONS FOR SCIENCE AND TECHNOLOGY/ENGINEERING MCAS TESTS.....	340
E. LEXILE ANALYZER® RESULTS.....	341
F. FIRST LANGUAGE CHARACTERISTICS: LANGUAGE FAMILY AND ORTHOGRAPHY.....	342
G. TEXTUAL ANALYSES OF JUNE 2012 BIOLOGY MCAS MULTIPLE-CHOICE ITEMS.....	350
H. ITEM COMPOSITE LINGUISTIC COMPLEXITY .....	352
I. MCAS RAW-TO-SCALED SCORE CONVERSION FOR JUNE 2012 BIOLOGY MCAS .....	354
J. MCAS PERFORMANCE BY MEPA SCORE .....	355
K. CONTENT DOMAIN PERFORMANCE BY ENGLISH PROFICIENCY .....	356
REFERENCE LIST .....	359



## LIST OF TABLES

Table	Page
1.1 MCAS by Subject and Grade.....	8
1.2. 2012 MCAS Results by Performance Level (%).....	12
3.1. Performance Variable Summary.....	101
3.2. ELL Variable Summary.....	102
3.3. Age-Related Variable Summary.....	104
3.4. Item Attribute Variable Summary.....	105
3.5 Descriptions of Linguistic Complexity Variables.....	106
4.1. ELL Grade and Age Demographics for the June 2012 Biology MCAS .....	129
4.2. Age Demographics by Grade Level.....	130
4.3. ELL Programs.....	132
4.4. English Proficiency Level.....	133
4.5. English Proficiency Means by Subgroups .....	134
4.6. Ten Most Common First Languages for ELLs Who Took the June 2012 Biology MCAS .....	135
4.7. Ten Most Common First Languages with Combined Chinese Languages .....	135
4.8. Demographics by L1 Language Family.....	137
4.9. Demographics by L1 Orthography .....	138
4.10. ELL Demographics by Late-Entry Status.....	140
4.11. Domain and Cognitive Skill Level Item Attributes of the Multiple- Choice Items on the June 2012 Biology MCAS .....	144

Table	Page
4.12. Multiple-Choice Cognitive Skill Level across Content Domains on the June 2012 Biology MCAS .....	145
4.13. Measures of Central Tendency of Linguistic Elements of the Multiple-Choice Items on the June 2012 Biology MCAS.....	147
4.14. Summary of Scheffé Post Hoc Analysis on Total Answer Lexical Density (TALD) across Content Domains.....	148
4.15. Summary of Scheffé Post Hoc Analysis on Linguistic Elements across Cognitive Skill Levels.....	150
4.16. Measures of Central Tendency for Linguistic Complexity Variables by Domain and Cognitive Skill Level .....	151
4.17. Measures of Central Tendency for Normalized Linguistic Variables and Composite Linguistic Complexity of the Multiple-Choice Items on the June 2012 Biology MCAS .....	153
4.18. Percent Composite Linguistic Complexity Items across Domains.....	154
4.19. Item Composite Linguistic Complexity across Cognitive Skill Levels.....	155
4.20. ELL Performance on the June 2012 Biology MCAS .....	158
4.21. Summary of ELL June 2012 Biology MCAS Performance by English Proficiency .....	160
4.22. Summary of Scheffé Post Hoc Analysis on MCAS Score across English Proficiency Levels .....	164
4.23. Summary of Linear Regression between English Proficiency and Biology MCAS Score .....	166
4.24. ELL Performance Levels by First Language Family.....	170
4.25. ELL Scores on the June 2012 Biology MCAS Score by Language Family .....	170
4.26. Summary of MCAS Performance Level by L1 Language Family and English Proficiency Levels .....	175

Table	Page
4.27. Impact of First Language Characteristics and Late-Entry ELL Status on Biology MCAS Score .....	177
4.28. ELL Performance Levels by L1 Orthography .....	179
4.29. ELL Scores on the June 2012 Biology MCAS by L1 Orthography .....	180
4.30. Summary of MCAS Performance Level by L1 Orthography and English Proficiency Levels .....	182
4.31. ELL MCAS Scores and Performance Levels by Late-Entry ELL Status.....	187
4.32. Summary of MCAS Performance Level by Late-Entry ELL Status and English Proficiency Levels .....	188
4.33. ELL Multiple-Choice Percent Correct by Content Domain and Statewide Average Percent Correct .....	194
4.34. Domain Mean Percent Correct by English Proficiency (MEPA) Levels .....	199
4.35. Expected Domain Performance by English Proficiency Levels .....	201
4.36. Ranking of Domains by Mean Percent Correct for Each English Proficiency Level .....	203
4.37. Summary of Scheffé Post Hoc Analysis on Domain Percent Correct across Adjacent English Proficiency Levels.....	206
4.38. Domain Mean Percent Correct by Language Family .....	208
4.39. Domain Percent Correct by First Language Orthography .....	210
4.40. Domain Percent Correct by Late-Entry ELL Status .....	212
4.41. Domain Mean Percent Correct by Late-Entry ELL Status for MEPA Levels 3 to 5.....	214
4.42. Multiple-Choice Percent Correct by Cognitive Skill Level.....	219
4.43. Comparison of Cognitive Skill Mean Percent Correct by English Proficiency .....	220

Table	Page
4.44. Summary of Scheffé Post Hoc Analysis on Cognitive Skill Percent Correct across Adjacent English Proficiency Levels.....	221
4.45. Cognitive Skill Level Percent Correct by Language Family .....	223
4.46. Cognitive Skill Level Percent Correct by L1 Orthography .....	225
4.47. Cognitive Skill Level Percent Correct by Late-Entry ELL Status.....	227
4.48. Percent Correct by Item Linguistic Complexity .....	230
4.49. Linguistic Complexity Percent Correct by English Proficiency Levels.....	232
4.50. Summary of Scheffé Post Hoc Analysis on Linguistic Complexity Percent Correct.....	234
4.51. Summary of Linear Regression between English Proficiency and Item Linguistic Complexity Performance.....	236
4.52. Summary of Linear Regression between English Proficiency and Item Linguistic Complexity Performance for MEPA Levels 3 to 5 .....	237
4.53. Linguistic Complexity Percent Correct by First Language Family .....	239
4.54. Linguistic Complexity Percent Correct by First Language Orthography .....	242
4.55. Linguistic Complexity Percent Correct by Late-entry ELL Status.....	244

## LIST OF FIGURES

Figure	Page
1.1 Biology MCAS Item Cognitive Skills .....	10
2.1 Krashen's Monitor Model of L2 Acquisition .....	33
2.2. Cummins' Four-Quadrant Model .....	42
2.3. Language, Content, and the Biology MCAS .....	84
3.1. Study Design for ELL Biology MCAS Performance .....	90
3.2. 2012 Biology MCAS Questions with Diagram and Label Answer Options .....	112
4.1. MEPA Performance Levels by Late-Entry ELL Status .....	139
4.2. Percent Distribution of Item Cognitive Skills across Domains on the June 2012 Biology MCAS .....	146
4.3. Percent Composite Linguistic Complexity Items across Domains.....	155
4.4. Mean MCAS Score across English Proficiency Levels.....	161
4.5. Percent of ELLs Failing or Passing the June 2012 Biology MCAS by English Proficiency Levels .....	162
4.6. Percent at Performance Levels by English Proficiency for ELLs Who Passed the June 2012 Biology .....	163
4.7. MCAS Performance Level Percentages by First Language Family .....	171
4.8. MCAS Performance Level Percentages by English Proficiency and L1 Language Family .....	176
4.9. MCAS Performance Level Percentages by First Language Orthography .....	181
4.10. MCAS Performance Level Percentages by English Proficiency and L1 Orthography.....	185
4.11. MCAS Performance Level Percentages by English Proficiency and Late-Entry ELL Status .....	189

Figure	Page
4.12. ELL Mean Percent Correct by Content Domain Multiple-Choice Items.....	195
4.13. ELL Domain Mean Percent Correct Compared to State Average .....	196
4.14. Mean Percent Correct by Domain by English Proficiency .....	202
4.15. Domain Mean Percent Correct by L1 Language Family .....	209
4.16. Domain Mean Percent Correct by L1 Orthography.....	211
4.17. Mean Percent Correct for Cognitive Skill by First Language Family .....	224
4.18. Mean Percent Correct for Cognitive Skills by First Language Orthography .....	226
4.19. Mean Percent Correct for Item Linguistic Complexity by First Language Family .....	239
4.20. Mean Percent Correct for Item Linguistic Complexity by First Language Orthography .....	242

## CHAPTER 1

### INTRODUCTION

The number of English language learners (ELLs) nationwide grew by 50% between 2000 and 2010 (Cook, Boals, & Lundberg, 2011).<sup>1</sup> Although the Massachusetts student population has remained relatively stable at approximately one million students, the student demographics are changing (Massachusetts Department of Elementary and Secondary Education [MA DESE], 2011a; Zubrzycki, 2011). In Massachusetts, ELL students grew 64% between the 2000-2001 and 2011-2012 school years, and they represent the fastest growing student population (Office of English Language Acquisition & Academic Achievement [OELA&AA], 2013). In the 2013-2014 school year, ELLs represented approximately 7.9% of public school students in Massachusetts (MA DESE, 2011a), and projections indicate that ELLs will represent 20% of the students in Massachusetts public schools by 2021 (OELA&AA, 2012).<sup>2</sup>

The 1999-2000 school year found 70% of ELLs concentrated in 10% of schools that were predominantly urban (Cosentino de Cohen, Deterding, & Clewell, 2005). This urban concentration pattern has continued; Boyson and Short's (2012) survey of 63 U.S.

---

<sup>1</sup> An English language learner, or ELL, refers to a student whose native language is not English and who is not English proficient. Massachusetts state law defines an English learner as "a child who does not speak English or whose native language is not English, and who is not currently able to perform ordinary classroom work in English" (M.G.L. c. 71A § 2 English Language Education in Public Schools, 2002).

<sup>2</sup> In school year (SY) 2011-2012, ELLs represented 7.3% of the students in Massachusetts public schools (Massachusetts Department of Elementary and Secondary Education, 2012a).

ELL newcomer programs found that 52% self-identified as urban. In Massachusetts, over two-thirds (68%) of ELL students are concentrated in 12 school districts (OELA&AA, 2012), and the seven largest urban school districts in Massachusetts have ELL populations ranging from 14.1% in Springfield to 33.1% in Lowell (MA DESE, 2011a).

### **Problem Statement**

Urban schools in the United States have long wrestled with the educational needs of students who do not speak English (Stuftt & Brogadir, 2011), and today they also face additional challenges brought on by increased enrollment of ELLs with high transiency rates and the need to meet state and federal accountability requirements (Nevárez-La Torre, 2012). High schools face additional challenges in providing meaningful education to ELLs who enter the U.S. after the age of 12 years with minimal or no English. These “late-entry” ELLs may spend as little as one to two years in U.S. high schools, yet in that time they are expected to acquire the language and content knowledge needed to pass wide-scale assessments for post-secondary economic and educational opportunities.

ELLs need to acquire the language and skills not only to negotiate their world outside of school, but also to succeed academically in secondary and post-secondary education. Schools need to provide ELLs with the opportunity to develop the language skills that will let them access the same educational opportunities as their native-English-speaking peers. One widely accepted distinction in second language development is that which Cummins (1979) draws between basic interpersonal communication skills (BICS), the language used for everyday communication, and cognitive academic language proficiency (CALP), the language and discourse requisite for academic success (as cited in Cummins, 2003). Beyond this basic distinction, however, there is only consensus that



academic language is the language needed for academic success (Bailey & Huang, 2011; N. Lee, 2011).

In Massachusetts, high school graduation requirements include passing the Massachusetts Comprehensive Assessment System (MCAS) in English language arts (ELA), mathematics, and science (MA DESE, 2011c).<sup>3</sup> Each MCAS content area has its own discourse, which includes discipline-specific vocabulary and syntactic structures (Bailey & Huang, 2011; Cook et al., 2011; Tan, 2011). Historically, ELLs have not performed on par with native English speakers on the standardized MCAS exams (Abedi & Dietel, 2004; Cook et al., 2011). Therefore, accelerating the acquisition of academic language for ELLs who enter at the secondary level is critical, and schools need to implement second language acquisition (SLA) pedagogies that develop academic language to close the ELL achievement gap.

Late-entry ELLs, who enter the public education system after the age of 12 years, find it even more difficult to gain the language needed to succeed on the standardized assessments. SLA research shows that conversational English or BICS is acquired in one to three years; however, a significant body of research puts the acquisition of cognitive academic language proficiency, or CALP, at five to nine years on the high end and four to seven years on the low end (Cook et al., 2011; Cummins, 2008). Cook and Zhao (2011) found that students' starting ELL level impacted whether they attained English proficiency in five years; two-thirds of higher level ELLs reached English proficiency in five years, but only 10% of beginning ELLs did (as cited in Cook

---

<sup>3</sup> Students must take a Grade 10 science MCAS in one of the following content areas: physics, biology, chemistry, or technology/engineering.

et al., 2011, p. 68). If it takes a minimum of four to five years to acquire academic language in English, then how can late-entry secondary ELLs, who enter with minimal or no English, graduate from high school in four years (or even less time)? For secondary ELLs, there appears to be an incongruity between SLA research and the reality of the U.S. public education system.

### **Background and Context**

**English language learners.** The ELL population in U.S. schools is heterogeneous, and its heterogeneity presents challenges in serving all ELLs equally (Boyson & Short, 2012; Hersi, 2012). Boyson and Short (2012) studied 63 “newcomer programs” for newly arrived immigrant students across the U.S. and found that the ELLs in secondary newcomer programs ( $n = 45$ ) represented over 90 countries and 55 languages and dialects. Although the majority (75%) of ELLs in U.S. schools are native Spanish-speakers, this language group is not monolithic (Boyson & Short, 2012; Xu & Drame, 2008). Some ELLs enter U.S. schools literate and on grade level; they are well-educated in their first language. Some enter U.S. schools with basic emergent literacy in their first language and without any formal schooling. Others fall somewhere in between.

***Early-entry ELLs.*** Many ELLs enter public schools in kindergarten or in the primary grades. There is a broad body of research on two-way and one-way immersion programs for these young English learners (YELs), including policies and pedagogies to foster literacy development in one or both languages. If it takes five to seven years to acquire academic language, then in theory these YELs have time to catch up to their native-English-speaking peers. This is not to say that they do; rather, there is enough time to accommodate the linguistic time requirements.

***Late-entry ELLs.*** Secondary late-entry ELLs present distinct challenges to schools, especially in the development of academic literacy (Boyson & Short, 2012). Late-entry ELLs (whose age of arrival is 12 years or older) encounter time constraints to achieve academic success. Some ELLs enter the public school system in Grade 11 or even Grade 12 with barely enough time to acquire conversational English. Yet in Massachusetts, they need to take the Mathematics and science MCASs at the next test administration. In 2006, a student newly arrived from the Dominican Republic was required to take the Biology MCAS on his second day in a U.S. school, even though he spoke no English—because he was in Grade 11.

This shortened time is even more problematic for late-entry students with interrupted or limited formal education. These students often have minimal literacy skills in their first language. Boyson and Short's (2012) survey of newcomer programs in the United States (n = 63) found that 27% of the ELL students had limited or interrupted schooling; many secondary ELL newcomers were becoming literate for the first time. Most high school teachers have neither the classroom time nor the training to teach beginning or developing literacy skills, further compounding this problem (Boyson & Short, 2012).

**ELL designation in Massachusetts.** In Massachusetts, state law (English Language Education in Public Schools, 2002) defines an English learner as “a child who does not speak English or whose native language is not English, and who is not currently able to perform ordinary classroom work in English.” From 2004 to 2012, Massachusetts used the MEPA, which had two parts, to determine English proficiency. The first part was the MEPA-Reading/Writing test (MEPA R/W), a wide-scale assessment given to all

ELLs each spring to assess their progress in reading and writing English. The second part was the Massachusetts English Language Assessment-Oral (MELA-O), an informal assessment over a period of time by the classroom teacher to assess listening and speaking skills (MA DESE, 2011e). Spring 2012 was the last administration of the MEPA. In 2012, Massachusetts became the 28<sup>th</sup> state to join the World-Class Instructional Design and Assessment (WIDA) consortium (WIDA, n.d.). Massachusetts began using ACCESS for ELLs, the English proficiency assessment tool developed by WIDA, in the 2012-2013 school year. It is too early to know whether the switch from an English proficiency assessment developed specifically for Massachusetts to one developed for use in multiple states will impact ELL designation.

The MEPA had five performance levels ranging from 1 to 5 (MA DESE, 2012g). Students who tested at Level 5 were recommended to be reclassified from Limited English Proficient (LEP) to Former Limited English Proficient (FLEP); these FLEPs were considered able to perform classwork in English (MA DESE, 2011e). The spring 2012 MEPA results illustrated some aspects of the heterogeneity of Massachusetts ELLs. In spring 2012, the MEPA tests were administered to ELLs at the secondary level (n = 11,814), and ELLs who had been in Massachusetts schools for five or more years represented the largest segment (31.1%), followed by second-year ELLs (20.2%), first-year ELLs (19.9%), third-year ELLs (16.8%), and fourth-year ELLs (11.2%; MA DESE, 2013d).<sup>4</sup> Of the secondary ELLs who have been in Massachusetts schools for five or

---

<sup>4</sup> These percentages are for ELLs who took the MEPA in spring 2012. Participation rates for ELLs in their first through fourth year in Massachusetts ranged from 92% to 93%. Participation for ELLs in the fifth or later years in Massachusetts only had a participation rate of 79%, which indicated that the number of ELLs in Massachusetts for five or more years was actually a higher percentage of the total ELL population. The overall participation rate of ELLs in the MEPA was 88%.

more years, only 33% achieved a MEPA score of 5; 67% were still considered limited English proficient, with 34% still at a Level 3 performance (MA DESE, 2013d). ELLs in their first year in Massachusetts showed the greatest diversity: 29% tested at Level 1, but 51% tested at Level 3 or higher (MA DESE, 2011e, 2012g).

### **The MCAS.**

***History.*** In 1993, the Massachusetts legislature passed the Education Reform Act “to provide immediately for the improvement of public education in the commonwealth” (Education Reform Act, 1993). The Act amended M.G.L. c. 69 §1 under the intent, among others, of ensuring “a deliberate process for establishing and achieving specific educational performance goals for every child.” It further amended M.G.L. c. 69 §1D as follows:

The “competency determination” shall be based on the academic standards and curriculum frameworks for tenth graders in the areas of mathematics, science and technology, history and social science, and English, and shall represent a determination that a particular student has demonstrated mastery of a common core of skills, competencies and knowledge in these areas, as measured by the assessment instruments described in section one. Satisfaction of the requirements of the competency determination shall be a condition for high school graduation.

The development of the MCAS was one response to this call for educational reform (MA DESE, 2011d). The high school MCAS exams—Grade 10 ELA, Grade 10 Math, and one science content area—are high-stakes tests; passing scores on all three are required for a high school diploma (MA DESE, 2011c).

***Who takes the MCAS?*** The MCAS exams are standardized assessments given to all students who are educated with Massachusetts public funds (MA DESE, 2011d) in Grades 3 through 8 and in high school according to the schedule in Table 1.1.

Table 1.1  
*MCAS by Subject and Grade*

<u>Subject Matter</u>	<u>Grade</u>
ELA	
Reading Comprehension	3, 4, 5, 6, 7, 8, and 10
Composition	4, 7, and 10
Mathematics	3, 4, 5, 6, 7, 8, and 10
Science and Technology/Engineering (STE)	
STE	5 and 8
Biology, Chemistry, Intro to Physics, or Technology/Engineering	High school

---

Adapted from: 2012-2013 MCAS/Access for ELLs testing schedule and administration deadlines (MA DESE, 2012b)

At the secondary level, the MCAS exams must be taken by all students in Grades 10, 11, or 12 who have not taken the Grade 10 MCAS exams and were not included in a school’s annual yearly progress (AYP). Grade 9 students are eligible to take a science MCAS exam if they have studied the Massachusetts curriculum for that content area. For example, if a student studies biology in Grade 9 rather than Grade 10, then the student is eligible to take the Biology MCAS exam in Grade 9 (MA DESE, 2011d).

In keeping with the Commonwealth of Massachusetts’ goal “to provide a public education system of sufficient quality to extend to all children ... the opportunity to reach their full potential and to lead lives as participants in the political and social life of the commonwealth and as contributors to its economy” (Powers and Duties of the Department of Elementary and Secondary Education, 1993), ELLs and students with


disabilities take the MCAS exams. For the Grade 10 ELA MCAS, schools have the option of not testing ELLs who have been in U.S. schools for less than one year; however, all ELLs are required to take the Grade 10 Math and science MCAS exams regardless of time in U.S. schools (MA DESE, 2011d). There are ELL accommodations for the MCAS; common accommodations include word-to-word dictionaries and content-specific glossaries (MA DESE, 2011d; 2012e).

**The Biology MCAS.** The June 2012 Biology MCAS assessed knowledge across six content domains: anatomy and physiology, biochemistry, cell biology, ecology, evolution and biodiversity, and genetics. There are 40 multiple-choice items and 5 constructed-response items.<sup>5</sup> Items assessed five cognitive skills levels: foundational, conceptual, application, constructive, and quantitative.<sup>6</sup> Foundational items were the least cognitively demanding, and constructive items were the most cognitively demanding.

---

<sup>5</sup> The Biology MCAS scores are based on 40 multiple-choice items and 5 constructed-response items. The instrument, however, has a total of 60 multiple-choice items and 7 constructed-response items because it includes 20 multiple-choice items and 2 constructed-response items that are being field-tested. Performance on field tested items are used only in the item development process; they do not affect a student's score and are not released to the public.

<sup>6</sup> See Appendix D for a full description of the Biology MCAS cognitive skill levels.

Basic Skills	Cognitive Skill	Description
<div style="text-align: center;">  </div>	Foundational	Declarative knowledge; recall facts; definition/vocabulary
	Conceptual	Concept recognition; descriptions of principles or processes
	Application	Procedural knowledge; application of conceptual knowledge to a novel situation; use predetermined models; classify diverse objects into unifying groups
	Constructive/ Synthetic	Synthesis of a novel response; multi-step problem solving; experimental design and critique; predictive reasoning; scientific inquiry or engineering design process and data analysis interpretation
More Demanding Skills		
-----		
Other Skills	Quantitative	Data analysis; computation of numerical solution; graphical and data table interpretation; predictive calculations

*Figure 1.1. Biology MCAS Item Cognitive Skills. Adapted from Massachusetts Department of Elementary and Secondary Education. (n.d.). Cognitive and quantitative skills descriptions for science and technology/engineering MCAS tests. Malden, MA: Author.*

**Pedagogy.** In Massachusetts, sheltered English instruction (SEI) is a common pedagogy in schools where ELL numbers can sustain separate SEI classrooms. Sheltered English instruction is a form of one-way immersion in which content classes consisting only of ELLs use English as the language of instruction (Rossell, 2005) and content is “sheltered” for language learners by modifying the language of instruction, materials, and assessments (Short, 1991). In Massachusetts, 92% of ELLs are in SEI classrooms (OELA&AA, 2012) where language is modified and ELLs do not compete academically with native speakers (Freeman & Freeman, 1988). ELLs, however, find themselves in a different context on the high-stakes MCAS. On the MCAS, ELLs do compete with native speakers; there is no modification of language, and proficiency categories are normed on native speaker benchmarks. ELLs can use a word-to-word dictionary and content



glossaries; however, if they lack academic or content language in their first language, these are of no help.

**ELL achievement gap.** ELLs do not perform as well as native English speakers on standardized assessments (Abedi & Dietel, 2004; Cook et al., 2011; Duran, 2008; Xu & Drame, 2008). This is not unexpected since by their very designation as ELLs these students are not proficient in English, the language of assessment. Duran (2008) states that “ELLs as a whole lag behind other students,” quantifying the gap as “0.5 to 2 standard deviations in magnitude on standardized tests” (p. 297).

This ELL achievement gap is seen in the 2012 high school MCAS test scores and performance levels.<sup>7</sup> Statewide, 88% of students scored Proficient or higher on the Grade 10 ELA MCAS; however, only 35% of ELLs scored Proficient or higher, 49% scored at Needs Improvement, and 16% failed (MA DESE, 2012i). A similar achievement gap is seen on the Grade 10 Mathematics MCAS: 78% of students statewide scored Proficient or higher; 32% of ELLs scored Proficient or higher; 33% scored Needs Improvement; and 35% failed (MA DESE, 2012i). The gap widened with the 2012 science MCAS. Statewide, 69% of students scored Proficient or higher, but only 17% of ELLs scored Proficient or higher; 50% of ELLs scored at Needs Improvement and 34% failed (MA DESE, 2012i). In other words, 84% of secondary ELLs scored below Proficient in science. These results are summarized in Table 1.2.

---

<sup>7</sup> At the secondary level, MCAS performance categories are Advanced, Proficient, Needs Improvement, and Failing.

Table 1.2  
*2012 MCAS Results by Performance Level (%)*

	<u>Advanced</u>	<u>Proficient</u>	<u>Needs Improvement</u>	<u>Failing</u>
Grade 10 ELA				
Statewide	37	51	9	3
ELLs	1	34	49	16
Grade 10 Math				
Statewide	50	28	15	7
ELLs	13	19	33	35
STE <sup>8</sup>				
Statewide	24	45	25	6
ELLs	2	15	50	34

Source: Spring 2012 MCAS tests: Summary of state results (MA DESE, 2012i)

### **Research Purpose and Rationale**

The purpose of this study was to analyze ELL performance on the June 2012 Biology MCAS and determine linguistic factor (English proficiency, first language family, first language orthography) and item parameter (content domain, cognitive skill level, and linguistic complexity) impact on ELL performance. Identification of factors that confound or mitigate ELL status can drive policy and pedagogical changes to close the ELL achievement gap on the Biology MCAS. Analysis of ELL performance on the Biology MCAS is exigent because this assessment is a high-stakes exam with consequences for both students and schools. Students must pass a science MCAS exam, such as the Biology MCAS, as a high school graduation requirement (MA DESE, 2011c). Previously, only student performance on the ELA MCAS and Mathematics MCAS were

<sup>8</sup> STE MCAS results only include students who were continuously enrolled in Massachusetts schools for Grades 9 and 10 (MA DESE, 2012i)

included in a school's AYP, the accountability measure under the No Child Left Behind Act (NCLB). That has changed. Beginning with the 2012-2013 school year, the composite performance index (CPI), the new accountability measure for Massachusetts public schools under the NCLB waiver, includes student performance on the Biology or other Grade 10 STE MCAS (MA DESE, 2012d).

Late-entry secondary ELLs are “vulnerable to academic failure” (Boyson & Short, 2012, p. 5). The research agrees that acquiring academic language takes significantly longer than acquiring everyday language. The low end of the time required to acquire academic language is four years, and nine years is the high end of the range. Since results from high-stakes testing can “misguide educators and policymakers in setting appropriate solutions to eliminate the academic gap” (Kim, 2008, p. 47), it is important to analyze the performance of ELLs, especially late-entry ELLs, on standardized assessments. This study disaggregated this subgroup and analyzed their Biology MCAS performance and the impact of linguistic factors and item parameters.

Historically, the United States has lacked a national curriculum. Thus, national datasets for science achievement could have construct validity issues, especially for ELLs, who have not studied the construct under assessment in English. The variance of science curricula by state, including the sequence of secondary science disciplines, argues for using state-level data. Likewise, ELL designations and proficiency levels historically have been determined by each state, and this also argues for using state-level data for exploring ELL performance. The educational landscape is changing with the move to the Common Core, which has been adopted by 45 states (“Standards in your state,” 2012) and the WIDA consortium for ELLs, which has 31 member states and territories (WIDA,

n.d.); however, the need to examine the academic performance of secondary ELLs, especially late-entry ELLs, cannot wait several years for future data.

Massachusetts is recognized as a leader in educational reform and as having a rigorous statewide assessment system, the MCAS. Massachusetts students have been recognized as being first in the United States based on National Assessment for Educational Progress (NAEP) data (Blume, 2012) and the Program for International Assessment mathematics data (Ripley, 2010). Massachusetts has long been proactive in developing its curricula standards and high-stakes assessments. Although Massachusetts adopted the national Common Core standards for English language arts and mathematics in 2010, it still retains its biology standards (MA DESE, 2010). Massachusetts has retained its rigorous content area assessments, and passing the ELA, mathematics, and a science MCAS remains a graduation requirement (MA DESE, 2010). The Biology MCAS afforded an opportunity to analyze secondary ELL performance on a rigorous content exam in the context of uniform ELL designations and proficiency levels, uniform content standards, an instrument designed to measure those specific content standards, and a uniform pedagogy of English-only instruction.<sup>9</sup>

### **Research Questions**

This study analyzed ELL Biology MCAS performance (total score and performance level) as well as ELL performance across domains,<sup>10</sup> cognitive skill levels,<sup>11</sup>

---

<sup>9</sup> In 2002, Massachusetts voters overwhelmingly passed Question 2, an “English-only immersion” initiative. Massachusetts General Laws c. 71A §4 states that “Subject to the exceptions provided in Section 5 of this chapter, all children in Massachusetts public schools shall be taught English by being taught in English and all children shall be placed in English language classrooms” (English Language Education in Public Schools, 2002).

<sup>10</sup> Domain refers to the content strand or standard. The six content domains assessed on the Biology MCAS are anatomy and physiology, biochemistry, cell biology, ecology, evolution and biodiversity, and genetics.

and item linguistic complexity. This study also explored the impact of the linguistic factors of English proficiency, Latinate/non-Latinate first language, and alphabetic/non-alphabetic first language. In addition, disaggregation of the late-entry ELL sample examined performance for this subgroup. Specific research questions were:

1. How did ELLs perform on the total score and on the performance level of the Biology MCAS?
  - (a) To what extent did English language proficiency impact total score and performance level on the Biology MCAS?
  - (b) To what extent did the first language family (Latinate or non-Latinate) impact total score and performance level on the Biology MCAS?
  - (c) To what extent did the first language orthography (alphabetic or non-alphabetic) impact total score and performance level on the Biology MCAS?
  - (d) To what extent did the late-entry ELL status impact total score and performance level on the Biology MCAS?
2. How did ELLs perform on the six content domains of the Biology MCAS?
  - (a) To what extent did English language proficiency impact performance on each of the six content domains of the Biology MCAS?
  - (b) To what extent did the first language family (Latinate or non-Latinate) impact performance on each of the six content domains of the Biology MCAS?

---

<sup>11</sup> The 2012 Biology MCAS assessed content knowledge across five cognitive skill levels: foundational, conceptual, application, quantitative, and constructive; however, multiple-choice items on the June 2012 Biology MCAS only assessed content knowledge at the foundational, conceptual, and application cognitive skill levels.

- (c) To what extent did the first language orthography (alphabetic or non-alphabetic) impact performance on each of the six content domains of the Biology MCAS?
  - (d) To what extent did the late-entry ELL impact performance on each of the six content domains of the Biology MCAS?
- 3. How did ELLs perform on the different cognitive skill levels of the Biology MCAS?
  - (a) To what extent did English language proficiency impact performance on the different cognitive levels of the Biology MCAS?
  - (b) To what extent did the first language family (Latinate or non-Latinate) impact performance on the different cognitive levels of the Biology MCAS?
  - (c) To what extent did the first language orthography (alphabetic or non-alphabetic) impact performance on the different cognitive levels of the Biology MCAS?
  - (d) To what extent did the late-entry ELL impact performance on the different cognitive levels of the Biology MCAS?
- 4. How did ELLs perform on the different levels (high, medium, low) of item linguistic complexity on the Biology MCAS?
  - (a) To what extent did English language proficiency impact performance on each of the three levels of item linguistic complexity of the Biology MCAS?

- (b) To what extent did the first language family (Latinate or non-Latinate) impact performance on each of the three levels of item linguistic complexity of the Biology MCAS?
- (c) To what extent did the first language orthography (alphabetic or non-alphabetic) impact performance on each of the three levels of item linguistic complexity of the Biology MCAS?
- (d) To what extent did the late-entry ELL impact performance on each of the three levels of item linguistic complexity of the Biology MCAS?

The first research question described ELL Biology MCAS performance. MCAS data report ELLs as a single cohesive group; however, ELLs are not monolithic. The first research question further described ELL Biology MCAS performance for subgroups disaggregated by (1) English proficiency level, (2) first language family (Latinate/non-Latinate), (3) first language orthography (alphabetic/non-alphabetic), and (4) late-entry ELL status. English proficiency levels among ELLs ranged from non-existent among newcomers to relatively proficient. This research question explored whether the Biology MCAS performance of ELLs with low English proficiency masked the performance of ELLs at intermediate and advanced English proficiency levels. Disaggregation of the ELL sample by English proficiency level showed a narrowing of the gap as English proficiency increases; it also addressed validity and reliability issues of Biology MCAS performance at lower levels of English proficiency.<sup>12</sup> The disaggregation of late-entry ELLs lent insight on Biology MCAS performance for ELLs who arrive after the age of

---

<sup>12</sup> See Abedi (2002).

12 years, a subgroup that faces unique challenges vis à vis time constraints in acquiring grade-level academic language.

The second research question explored ELL performance on each of the six content domains of the Biology MCAS. This study explored whether ELL performance is uniform across the six content domains or whether some domains are more accessible or more difficult. Like the first research question, the second research question further explored the impact of (1) English proficiency level, (2) first language family (Linate/non-Linate), (3) first language orthography (alphabetic/non-alphabetic), and (4) late-entry ELL status on ELL content domain performance.

Building on the exploration of differential performance across domains, the third research question explored ELL performance across the cognitive skill levels. Like the previous research questions, the third research question further explored the impact of (1) English proficiency level, (2) first language family (Linate/non-Linate), (3) first language orthography (alphabetic/non-alphabetic), and (4) late-entry ELL status on cognitive skill level performance for ELLs. Since cognitive skill levels represent different cognitive load and abilities, it was expected that lower cognitive skill level questions would have higher performance.

The fourth research question explored the impact of item linguistic complexity on ELL Biology MCAS performance. Several studies (Abedi, 2002, 2009; Abedi & Gándara, 2006; Abedi & Hejri, 2004; Abedi, Hofstetter, & Lord, 2004; Abedi & Lord, 2001; Martiniello, 2008; Menken, 2008, Chapter 4; Solano-Flores & Li, 2009) have shown that linguistic complexity impacts ELL performance. These studies have used various elements of linguistic complexity and methodologies. This study operationalized



elements of linguistic complexity from a textual analysis of the June 2012 Biology MCAS and explored ELL performance for items with low, medium, and high linguistic complexity. It further explored the impact of (1) English proficiency level, (2) first language family (Latinate/non-Latinate), (3) first language orthography (alphabetic/non-alphabetic), and (4) late-entry ELL status on performance at three levels of linguistic complexity.

### **Significance of the Study**

Language is the primary means through which knowledge is assessed; however, it is the most under-analyzed aspect of learning (Schleppegrell, 2004). Previous studies identified a need for further research into the role of language in ELL content assessment (Abedi, 2008b; Abedi et al., 2004; Abedi & Lord, 2001; Martiniello, 2008; Solorzano, 2008). A review of the literature identified a need to study the extent of the impact of English proficiency on ELL performance on standardized assessments (Solorzano, 2008) in order to inform the construction of reliable and valid content assessments for ELLs (Abedi, 2008b) and to serve as a source of differential item functioning (Martiniello, 2008). This study contributes to this literature by examining English language proficiency impact on performance and across the item parameters of domain, cognitive skill level, and linguistic complexity. It further contributes to this literature by disaggregating late-entry ELLs and exploring English language proficiency impact for this at-risk subgroup.

This study also adds to the body of prior research demonstrating that the linguistic complexity of items impacts ELL achievement on wide-scale assessments. This study contributes to the literature by examining linguistic complexity and differential item functioning for ELLs from all language backgrounds, a need for further study

identified by Martiniello (2008). It further augments this literature by disaggregating the late-entry ELL population and exploring differential impact for this subgroup.

By using state-level data, this study addressed validity and reliability concerns raised by Abedi et al. (2004) and Solorzano (2008) with respect to the varying ways states define English proficiency and re-designate ELLs as English proficient. The ELL data in this study were based on English proficiency level designations, including re-designation as English proficient, used uniformly throughout Massachusetts. This study analyzed the impact of linguistic factors (English proficiency and first language characteristics) and late-entry ELL status for ELLs at Massachusetts English Proficiency Assessment (MEPA) scores of 3 and higher. This addressed reliability concerns at the lower English proficiency levels and the limitation of uncertain English proficiency levels that Abedi and Hejri (2004) encountered using NAEP data.

This study contributes to the literature by considering ELL and disaggregated late-entry ELL performance not only across six biology domains but also across three cognitive skill levels and three levels of item linguistic complexity. To my knowledge, no study has been done on: (1) ELL performance on the Biology MCAS; (2) ELL performance across the six domains of the Biology MCAS; (3) ELL performance across the cognitive skill levels of the Biology MCAS; (4) ELL performance across three levels of item linguistic complexity on the Biology MCAS; or (5) the impact of English proficiency level, a Latinate or non-Latinate first language, an alphabetic or non-alphabetic first language, and late-entry ELL status. This study contributes to the literature on secondary ELLs, including the late-entry ELL subgroup and wide-scale

science assessments in general, and serves as a springboard for further studies on ELL and late-entry ELL achievement on the Biology MCAS in particular.

### **Definition of Terms**

The following definitions are provided to ensure uniform understanding of the terms used throughout the study.

“Academic language” is the language used in academic contexts such as instruction, textbooks, and assessment. This study interpreted academic language as the lexical (vocabulary) and syntactic (grammatical) elements situated in discourse communities—language features can be classified as general or content-specific. This conforms to the notions of academic language proposed by Bailey and Huang (2011), Schleppegrell (2001), and Snow and Uccelli (2009). This study used academic language interchangeably with *academic register* and *academic discourse*, and treated the language of the Biology MCAS as a discourse community.

“Basic interpersonal communication skills,” or BICS, comprise the everyday social language as defined by Cummins (1979, as cited in Cummins, 2003).

“Cognitive academic language proficiency,” or CALP, is the language of schooling as defined by Cummins (1979, as cited in Cummins, 2003).

“Cognitive skill level” refers to the foundational, conceptual, application, quantitative, and constructive cognitive skills tested on the Biology MCAS.

An “English language learner,” or ELL, refers to a student whose native language is not English and who is not English proficient; rather, ELLs are in the process of becoming English proficient. In some literature, ELLs are also referred to as English learners (ELs) or as limited English proficient (LEP).

“First language acquisition,” or FLA, refers to the subfield of linguistics that studies the acquisition of a first language.

“L1” refers to first language.

“L2” refers to any language learned subsequent to the first language (i.e., second, third, fourth languages, etc.).

“Late-entry ELL” refers to an ELL who arrived in the United States at the age of 12 years or older.

“Lexeme/lexical item,” in linguistic terms, refers to a unit in a language’s compendium of all words and word variations (Finegan, 2004).<sup>13</sup> This study used *lexical* in a more general sense that is congruent with its use and expression in classrooms and pedagogies. This study used lexical to refer to the word or vocabulary level of language.

The “Massachusetts Comprehensive Assessment System,” or MCAS, exams are the standardized tests developed by the Commonwealth of Massachusetts to assess students. The MCAS exams in Grade 10 English language arts (ELA), mathematics, and science are high-stakes tests; passing all three is a high school graduation requirement. Secondary level science MCAS content areas are physics, biology, chemistry, and technology/engineering.

The “Massachusetts English Language Assessment-Oral,” or MELA-O, was a subtest of the MEPA. It was an informal assessment administered by the classroom teacher over a period of few weeks to assess an ELL’s listening and speaking skills.

---

<sup>13</sup> Variations of a lexeme or word would include singular, plural, possessive forms, tense, and aspect forms, among others.

MELA-O scores ranged from 1 (minimal) to 5 (native-like). The ACCESS for ELLs instrument replaced the MELA-O in 2013.

The “Massachusetts English Proficiency Assessment-Reading/Writing,” or MEPA R/W, was a subtest of the MEPA. It was a standardized exam that assessed reading and writing skills. The ACCESS for ELLs instrument replaced the MEPA R/W in 2013.

The “Massachusetts English Proficiency Assessment,” or MEPA, was developed for the Commonwealth of Massachusetts to assess the English proficiency level of ELLs in the language skills of reading, writing, listening, and speaking; it was also a measure of annual growth in English proficiency. The MEPA consisted of two subtests: (1) the MEPA R/W, which assessed reading and writing skills, and (2) the MELA-O, which assessed listening and speaking skills. ELLs new to Massachusetts public schools took the MEPA in October for a baseline score. In the spring, all ELLs took the MEPA to measure English proficiency and growth since the prior test administration. March 2012 was the last administration of the MEPA; the ACCESS for ELLs instrument replaced the MEPA in 2013.

The “Massachusetts Department of Elementary and Secondary Education,” or MA DESE, is the Commonwealth of Massachusetts’ governmental agency that oversees education.

“Science MCAS” refers to the science and technology/engineering MCAS exams in the following areas: biology, chemistry, physics, and technology/engineering. A passing score on one of these science content MCAS exams is one of the requirements for high school graduation in Massachusetts.

“Second language acquisition,” or SLA, refers to the subfield of linguistics that studies the acquisition of a language subsequent to first language acquisition. In this context, a second language refers to any language that is not a first language (i.e., second, third, fourth languages, etc.).

“Sheltered English instruction,” or SEI, is the common pedagogy for ELLs in Massachusetts. It is a one-way English immersion pedagogy for ELL content classrooms whereby language is modified to make content accessible (Short, 1991).

“Syntax/syntactic,” in linguistic terms, refers to how words are structured into patterns to convey meaning (Finegan, 2004). This study used the word *syntax* in a more general sense that is congruent with its use and expression in classrooms and pedagogies. This study used syntax to mean a language’s rules of grammar and sentence structure.

“Total Biology MCAS Score” means the total raw score on the June 2012 Biology MCAS.

“World-Class Instructional Design and Assessment,” or WIDA, is a consortium of 31 U.S. states and territories headquartered at the Wisconsin Center for Education Research. WIDA focuses on academic achievement for ELLs, including development of academic language (<http://www.wida.us>).

### **Limitations and Delimitations**

ELLs who enter U.S. schools are heterogeneous in their prior education and first language literacy. Some of these ELLs have academic language in their first language, and some have also studied biology in their first language. The common underlying proficiency model allows for positive transfer between L1 (first language) academic language and L2 (second language) academic language (Cummins, 2000). The use of

existing statewide MCAS data did not permit disaggregation of the ELL population into subgroups based on L1 academic language or L1 biology content knowledge. A limitation of this study was the impact of L1 academic language or L1 biology knowledge on Biology MCAS performance.

The impact of poverty and other socioeconomic factors was beyond the scope of this study; however, it is noted that in 2011, 79% of ELLs in Massachusetts were low-income (MA DESE, 2012j). The notion of cultural validity in assessments was also beyond the scope of this study. Performance and portfolio assessments as alternatives to wide-scale standardized assessments were beyond the scope of this study. Likewise, the construction, validity, and reliability for wide-scale assessments designed to measure Biology and English language proficiency were beyond this study's purview.

### **Assumptions**

Transiency in the ELL population was a limitation in this study. MEPA data track the number of years a student has been enrolled in Massachusetts public schools. In the current study, the number of years enrolled in Massachusetts public schools served as a proxy for years in the United States, which in turn was used to calculate age of arrival for the disaggregation of late-entry ELLs ( $\text{Age} - \text{Years in Massachusetts schools} = \text{Age of arrival}$ ). Equating years in Massachusetts schools with years in the United States posed a limitation. For example, a student who is 16 years old and has been enrolled in Massachusetts public schools for two years would be considered a late-entry ELL under the construct in this study ( $16 \text{ years} - 2 \text{ years} = 14 \text{ years}$ ). If that same student, however, had spent three years in another state before coming to Massachusetts, the age of arrival becomes 11 years; this student would then be part of the disaggregated late-entry ELL

population even though he or she would not meet the criteria of the late-entry ELL construct used in this study. It was not feasible to review individual student records for each ELL who took the Biology MCAS. This study assumed that years in Massachusetts public schools was equivalent to time in the United States.

Another limitation of the study related to determining uniform access to the biology curriculum. Massachusetts has developed biology standards; however, it is left to the individual districts and schools to determine when and how these biology standards are implemented. Some districts and schools teach the full biology curriculum in Grade 9; others teach it in Grade 10. Still others teach the biology curriculum over two years in Grade 9 and Grade 10. It was assumed that secondary ELLs took biology at the same time as their English-speaking peers (i.e., in Grade 9, Grade 10, or both Grade 9 and Grade 10). A limitation of this study was that some ELLs, especially those who enter in Grades 11 or 12, may not have had exposure in English to the full biology curriculum assessed by the Biology MCAS. This study assumed that the ELLs who took the Biology MCAS had been exposed to the complete Massachusetts biology standards.

In 2002, Massachusetts mandated English-only instruction for ELLs. A limitation of this study was that it was impossible to know what went on behind closed doors in every ELL biology classroom in Massachusetts or whether each ELL was placed appropriately. This study assumed that ELLs are in sheltered English instruction (SEI) or mainstream biology classrooms where the language of instruction is English.<sup>14</sup> It also was

---

<sup>14</sup> In SY 2011-2012, Massachusetts ELL program enrollment was 91% in SEI, 5% in no program, 1.5% opted out, 1.5% in other programs, and 1% in dual language (OELA&AA, 2013).



assumed that teachers in those classrooms have a biology teaching license and either a separate license, certification, or training to teach ELLs.

Massachusetts allows ELL accommodations in the form of word-to-word bilingual dictionaries and glossaries. A limitation of this study was the inability to track whether an ELL had use of word-to-word bilingual dictionaries, glossaries, or both. This study assumed that ELLs received all permitted accommodations on the Biology MCAS.

### **Summary**

A science MCAS, such as the Biology MCAS, must be taken by all Grade 10 students enrolled in Massachusetts public schools on the date of test administration; unlike the Grade 10 ELA MCAS, there is no exception for ELLs who have been in the United States for less than one year. In practice, this means that ELLs with minimal language skills and who may have not been exposed to all the Massachusetts biology standards still must take the test. ELL secondary students in Massachusetts experience the curriculum through English. Much of the work on one-way English immersion programs in the U.S. has focused on elementary or middle school students. The ELL population whose entry is at the secondary level, however, must be differentiated because they have fully acquired the phonetic, morphologic, syntactic, and pragmatic elements of their first language (Ellis, 2003). They also face time constraints in acquiring the academic language requisite for achievement on wide-scale standardized assessments.

This study analyzed the Biology MCAS performance of secondary ELLs and explored factors impacting their achievement. There is a gap in second language acquisition, academic language, and assessment literature with respect to an analysis of late-entry ELL performance on the Biology MCAS. Most of the literature on high-stakes

testing and secondary ELLs has focused on ELA and mathematics, and the few studies on secondary ELLs and science achievement were not for the Biology MCAS. Likewise, most of the studies have focused on Hispanic or Latino students. This study analyzed performance for the ELL sample, as a whole, and for subgroups disaggregated by (1) English proficiency level, (2) first language family (Linate/non-Linate), (3) first language orthography (alphabetic/non-alphabetic), and (4) late-entry ELL status. This identified ELL subgroups that were closing the achievement gap on the Biology MCAS and subgroups that were still at risk. In addition, this study analyzed ELL performance across the six content domains and the cognitive skill levels assessed on the June 2012 Biology MCAS, as well as across three levels of item linguistic complexity.

## CHAPTER 2

### LITERATURE REVIEW

This chapter describes the current literature on secondary ELLs and wide-scale assessments, particularly in content areas. The literature review is divided into the following fields: (1) second language acquisition theories, (2) academic language, (3) critical issues on cognitive load and item attributes, and (4) critical issues on ELLs and high-stakes testing. The section on second language acquisition (SLA) theories provides the reader with a brief overview of the three major schools of SLA. It then focuses on two aspects of SLA with respect to secondary ELLs: (1) Krashen's five hypotheses of SLA and (2) the critical period hypothesis (CPH). Krashen's five hypotheses of SLA comprise a conceptual framework on how second languages are acquired. The CPH posits that there is a critical period for second language acquisition, and beyond this critical period, second language acquisition falls short of native-like proficiency. This section concludes with a rationale for differentiating ELLs who enter after the age of 12 years from ELLs who arrive at an earlier age.

The second section discusses the literature on academic language. It begins with a discussion of Cummins' seminal work demarcating language into basic interpersonal communication skills (BICS) and cognitive academic language proficiency (CALP), including a discussion of his four-quadrant model. Next follows a discussion of how

academic language has been defined and characterized in the literature, including academic language particular to science texts. After discussing the relation between academic language proficiency and achievement, this section concludes with a discussion of the research indicating that ELLs need four to seven years to acquire academic language.

The third section discusses critical issues on cognitive load and item attributes on wide-scale assessment items. It begins with a discussion of item difficulty, followed by a discussion of item cognitive complexity. Since cognitive load is related to item complexity, this section then discusses types of cognitive processing and cognitive load. It also includes a discussion of language factors that may affect cognitive load. It then discusses cognitive load in relation to second language learning and ends with a brief discussion on differential item functioning.

The fourth section highlights critical issues surrounding ELLs and high-stakes assessments. It examines the appropriateness of using standardized assessments developed for and normed on English proficient students. A discussion follows on how ELL language factors may affect the validity and/or reliability of these assessments. This section concludes with a discussion on testing accommodations used to address language factors in assessing ELLs on standardized tests.

Drawing on these bodies of literature, the literature review concludes with a conceptual framework for analyzing ELL performance on the Biology MCAS. The conceptual framework draws on Cummins' CALP and four-quadrant model and Krashen's input hypothesis, and posits that learner characteristics impact Biology MCAS performance through their intermediary effect on academic language. The learner

characteristics explored in this study were English proficiency level, a Latinate L1, an alphabetic L1, and whether age of entry was 12 years or later.

### **Review of Second Language Acquisition Theories**

Second language acquisition covers many disciplines, including, among others, linguistics, psycholinguistics, neurolinguistics, sociology, anthropology, cultural studies, and pedagogies. “The field of [second language acquisition] lacks a uniformly accepted theory of how [second languages] are acquired” (Marinova-Todd, Marshall, & Snow, 2000, p. 14). This part of the literature review briefly describes the major schools of SLA theory and focuses on aspects of SLA that impact ELLs whose age of arrival is 12 year or later.

**Second language acquisition theoretical schools.** Second language acquisition theories can be broadly categorized into three schools that parallel the schools in first language acquisition (FLA): behaviorism, innatism, and interactionism.

***Behaviorism.*** The earliest school of language acquisition theory, behaviorism, holds that language is acquired through imitation and repetition. In second language pedagogy, behaviorism gave rise to the audio-lingual method, which is characterized by dialogue memorization and hours spent in language labs listening and repeating (Larsen-Freeman, 2000). Although some aspects of behaviorism may account for second language acquisition, this school has been superseded by innatism and interactionism (Lightbown & Spada, 2004).

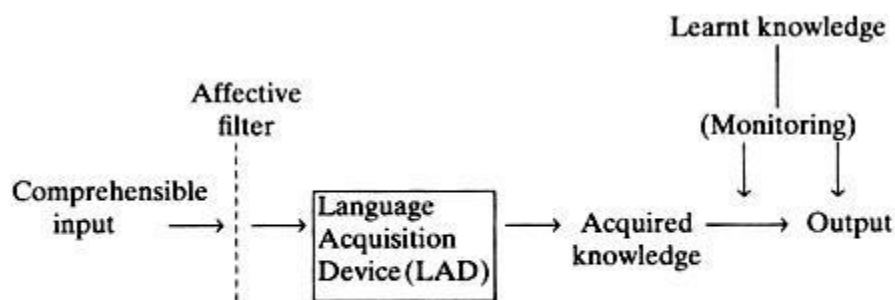
***Innatism.*** In FLA, the innatist school believes that humans are born with an innate ability to acquire language. That is, since human brains are pre-wired for language, language acquisition is just another biologically programmed ability. Innatists theorize

that the process of language acquisition commences upon exposure to language, and thereafter it follows a prescribed sequence until acquisition is complete. The precise brain region where language acquisition takes place has not been identified; however, Chomsky named it the language acquisition device (LAD), and this is widely accepted by innatists and interactionists (Lightbown & Spada, 2004). In SLA, innatists believe that this LAD is also responsible for learning a second language, though the LAD may not be as available with age.

***Interactionism.*** Like the innatist model, the interactionist model of language acquisition is based on an innate language ability; however, interactionists place more importance on the linguistic environment. In first language acquisition, interactionism draws on the work of cognitive psychologist Piaget and Vygotsky's sociocultural theory of mental processing (Lightbown & Spada, 2004). According to interactionists, a child's innate ability to acquire language is mediated by interaction with proficient speakers who modify linguistic input to the child's level—that is, child-directed speech. Modified speech is not limited to linguistic simplification; it also includes speech that is context-embedded, at a slower rate, elaborated, and supplemented by other communicative acts such as gesticulation (Lightbown & Spada, 2004).

**Krashen's monitor model and five hypotheses.** Stephen Krashen, a prominent SLA researcher, developed the monitor model and five hypotheses of SLA (Finegan, 2004). The monitor model is an adaptation of Chomsky's LAD and integrates Krashen's five hypotheses of SLA (see Figure 2.1). Krashen's model begins with comprehensible input (input hypothesis), which is then filtered (affective filter hypothesis) before entering the LAD. The LAD then processes the input to produce acquired knowledge (acquisition-

learning hypothesis), which is then monitored (monitor hypothesis) using learned knowledge (acquisition-learning hypothesis) to produce observable linguistic output (see Lightbown & Spada, 2004). Although Krashen's monitor model and five hypotheses are widely accepted by applied linguists, they are subject to criticism. Many critics believe that Krashen's model and hypotheses use imprecise terminology and are either untestable or poorly supported by empirical evidence (Lightbown & Spada, 2004; Zafar, 2009).



*The Input Hypothesis Model of L2 learning and production (adapted from Krashen, 1982, pp. 16 and 32; and Gregg, 1984)*

*Figure 2.1.* Krashen's Monitor Model of L2 Acquisition. This model illustrates the role of comprehensible input in acquiring knowledge for linguistic output.  
Source: (Isabelli, n.d.)

In his first hypothesis, the acquisition-learning hypothesis, Krashen demarcates language into acquired language and learned language for adult second language learners. According to Krashen, understanding the language of exposure results in acquired language, much in the same way children acquire their first language, and this is what leads to fluency. In contrast, learned language is the result of a conscious effort and attention to form and rules (Lightbown & Spada, 2004). Krashen believes that acquired and learned languages are distinct and that learned language cannot become acquired

language. The acquisition-learning hypothesis is one of Krashen's more provocative constructs about SLA (see Zafar, 2009).

Krashen's second hypothesis, the monitor hypothesis, involves productive language (speaking and writing) and also incorporates the demarcation between acquired language and learned language. The second language learner has an internal monitoring system in which learned language "monitors" acquired language when the focus is not only on conveying meaning but also on correctness (see Lightbown & Spada, 2004). The majority of the Biology MCAS is multiple-choice (40 out of 60 possible points) and not productive language; however, its nature as a high-stakes assessment would bring "correctness" to the fore of a student's mind, and thus the monitor would be activated.

Whereas the monitor hypothesis deals with productive language, Krashen's fourth hypothesis, the input hypothesis, involves receptive language (listening and reading). The input hypothesis emphasizes the importance of comprehensible ( $i + 1$ ) input in SLA, and it is similar to Vygotsky's zone of proximal development in FLA (Lightbown & Spada, 2004). Since understanding the question is essential for achievement on assessments, comprehensible input contributes to ELL achievement on the Biology MCAS.

Krashen's fifth hypothesis, the affective filter hypothesis, states that mental dispositions and other factors can raise barriers to language acquisition despite comprehensible input (Lightbown & Spada, 2004). The affective filter hypothesis has been criticized as untestable, especially with respect to establishing not only causality between second language acquisition and affective filters but also the direction (Lightbown & Spada, 2004; Zafar, 2009).



**Critical period hypothesis.** Lenneberg, an innatist, believed that after brain lateralization (specialization) of language function, first language acquisition becomes difficult (Collier, 1987b; Marinova-Todd, et al., 2000; Snow & Hoefnagel-Höhle, 1978). Lenneberg proposed the critical period hypothesis (CPH) in FLA, which states that if language acquisition does not begin before a certain age, full language acquisition will not occur (Collier, 1987b; Marinova-Todd et al., 2000). Research has shown that there does indeed appear to be a critical period for acquiring a first language, and the CPH is widely accepted in the field of first language acquisition (Brown, 2000; Lightbown & Spada, 2004). Researchers, however, do not agree on when the critical period for first language acquisition ends; some believe it is by age 5, and others, including Lenneberg, believe it is by the start of puberty (Brown, 2000; Lightbown & Spada, 2004).

Since SLA often parallels FLA, the question of a similar critical period for SLA has been raised. Some hypothesize that first language acquisition occurs through universal grammar, the innate language ability, which has a critical period for activation, but that adult second language acquisition occurs through the problem-solving aspect of cognition (Long, 2007, Chapter 3). Marinova-Todd et al. (2000) stated that it is now generally agreed that there is not a critical period for second language acquisition, except in pronunciation (also see Collier, 1987a, 1987b). The issue, however, has not been settled definitively. Long (2007) reviewed oppositional studies on critical or sensitive periods for second language acquisition and found that even with more than 100 empirical studies, there is no consensus on the existence or nature of a critical or sensitive period in second language acquisition. He further stated:

Researchers on both sides of the critical period issue agree about the facts to be explained, that marked differences in ultimate attainment constitute one of the most salient features—perhaps the most salient—distinguishing child and adult language acquisition and hence one of the empirical (cf. conceptual) “problems” (L. Laudan, 1977, 1996b, and elsewhere) a viable SLA theory needs to solve. (p. 45)

Long (2007, Chapter 3) used the term “sensitive periods” rather than critical periods since adults can acquire a second language, but often their ultimate acquisition is less native-like proficiency than those who began at a young age. He made a strong argument for one or more sensitive periods, citing two unpublished studies of former students. Long (2007, Chapter 3) concluded from his review of the literature that there are multiple sensitive periods in second language acquisition. The sensitive period for acquiring native-like pronunciation appears to be up until the age of 6 years, and the sensitive period for acquiring native-like morphology and syntax appears to extend into the mid-teens (DeKeyser, 2000; DeKeyser, Ravid, & Alfi-Shabtay, 2004, as cited in Long, 2007, Chapter 3).

Snow and Hoefnagel-Höhle (1978) studied the acquisition of Dutch by English speakers living in the Netherlands across age and proficiency levels. Age levels studied were 3- to 5-year-olds ( $n = 10$ ), 6- to 7-year-olds ( $n = 8$ ), 8- to 10-year-olds ( $n = 13$ ), 12- to 15-year-olds ( $n = 8$ ), and adults ( $n = 11$ ). The two Dutch proficiency levels were Beginner and Advanced, distributed across the five age groups. The study controlled for test bias against younger subjects by designing an instrument that was content appropriate for the younger subjects; however, they stated that as a result, “most of the test material

was quite childish for the older subjects” (p. 1123). Two groups of monolingual Dutch speakers were tested to establish age norms. The Advanced group was tested only once; the Beginner group was tested at three intervals over a one-year period.

The 12- to 15-year-old group exhibited the most rapid acquisition of the skills tested, and the 3- to 5-year-old group scored the lowest. Snow and Hoefnagel-Höhle (1978) reasoned that positive L1 to L2 transfer accounted for the 12- to 15-year-olds outperforming the younger age groups. At the six-month mark, the authors found that the 12- to 15-year-old group approached native-Dutch-speaker performance faster than the other groups except in the skill of sentence judgment. At the one-year mark, the 12- to 15-year-old group continued to score higher in most skills, and all the 6- to 15-year-old subjects had acquired sufficient Dutch to be considered bilingual. Based on these findings, the authors rejected a critical period for SLA and the belief that younger children are better second language learners than adolescents. In his criticism of studies disproving a critical period in second language acquisition, Long (2007) claimed that the Snow and Hoefnagel-Höhle study is flawed due to “confusion between rate and ultimate attainment” (p. 70).

Collier (1987b) listed nine studies conducted between 1962 and 1984 that showed that ELLs arriving between the age of 8 and 12 years acquired academic language in a second language faster than students who arrive at younger ages. It is generally thought that older ELLs have the advantage of a more fully acquired first language and higher meta-linguistic awareness. Long (2007) pointed out that several studies do show that older learners acquire a second language faster than young children; however, rate of

acquisition is not the same as ultimate attainment, and it is in the latter that young learners eventually surpass learners who begin after puberty.

There is consensus in applied linguistics that a second language can be acquired at any age, as is evidenced by second language speakers worldwide. There are differing opinions, however, as to what constitutes acquisition, proficiency, and native-like ability, among other measures. Marinova-Todd et al. (2000) discussed the misconceptions about age and learning a second language, including the misconception that age precludes native-like second language proficiency:

Age does influence language learning, but primarily because it is associated with social, psychological, education, and other factors that can affect L2 proficiency, not because of any critical period that limits the possibility of language learning by adults. (p. 28)

The current study disaggregated late-entry ELLs and explored whether English proficiency and first language characteristics had differential impact on their Biology MCAS performance.

**Late-entry ELLs.** The ELL population whose age of entry is 12-plus years must be differentiated because they have fully acquired their first language. The research shows that the phonetic, morphologic, syntactic, and pragmatic elements of a first language are fully acquired by the age of 12 years, if not earlier. Lightbown and Spada (2004) state that “In fact, it is generally accepted that by age four, children have mastered the basic structures of the language or languages which have been spoken to them in these early years” (p. 2). Finegan (2004) puts the majority of first language acquisition at six years but leaves room for additional linguistic acquisition:

By the age of 6, barring several mental or physical impairments, children the world over have acquired most of what they need to know to speak their language fluently. By the time a child arrives in school, perhaps 80% of the structures of its language and more than 90% of the sound system has been acquired. (p. 541)

Collier (1987b, as cited in McLaughlin, 1984, pp. 41-43) believes that first language acquisition continues through the elementary and middle school years (age 6 to 12 years) as children are developing their reading and writing skills. This conforms to Cummins' demarcation of language into the language of everyday social interaction and the language encountered in the context of school. Children arrive at school fully competent in the former, but it is the latter that develops throughout schooling (Cummins, 2003).

Although there is no consensus on the exact age when first language acquisition is complete, most agree that it is acquired by the age of 12 years. Thus, late-entry ELLs have fully acquired the phonetic, morphologic, syntactic, and pragmatic elements of their first language, and they bring this knowledge to second language acquisition (Ellis, 2003). Older second language learners are aware of morpho-syntactic linguistic elements such as parts of speech and word order even if they have not been explicitly taught terms like noun, verb, and direct object (Finegan, 2004). Some researchers believe this conscious familiarity with grammar, an aspect of metalinguistic awareness, promotes second language acquisition; however, to what extent remains unknown (Finegan, 2004). Literacy (reading and writing) increases metalinguistic awareness (Lightbown & Spada, 2004), and this can facilitate second language acquisition. Literacy, however, is a language skill, and the linguistic elements of a first language are still acquired in its absence. "In the case of ALA (adult language acquisition) all of the relevant cognitive

machinery is in place” (Robinson & Gilabert, 2007, p. 165, as cited in Slobin, 1993, p. 243). Therefore, as successful first language learners, late-entry ELLs have conscious and unconscious metalinguistic abilities to help facilitate their second language acquisition.

### **Review of Academic Language Literature**

“Schooling is primarily a linguistic process” (Schleppegrell, 2004, p. 2). Although there may not be a consensus on the specific details, academic language is the register or discourse requisite for academic success (Abedi, 2008b; Bailey & Huang, 2011; Snow & Uccelli, 2009), though Bunch (2006) argues for broadening the definition. Schleppegrell (2004) believes that defining academic language requires a level of abstraction because not all features of academic language are present in all instances. Rather, academic language is a “constellation of features” (Schleppegrell, 2004, p. 5) that constitute a language register, which is defined as the “lexical and grammatical features that characterizes particular uses of language” (Schleppegrell, 2001, pp. 431, as cited in Halliday & Hassan, 1989; Martin, 1992).

In linguistics, the notion of register is based on the use of particular linguistic features in different contexts and purposes (Johnstone, 2002). Register has been described as specialized language use (Hymes, 2003) or as the language associated with occupations or social groups (Johnstone, 2002; Wardhaugh, 2006). As Johnstone (2002) notes, not all linguists are uniform in their use of term register.

Likewise, the term discourse has two general but different meanings in linguistics. Discourse can mean any communicative act utilizing language; it is used as a non-count noun describing any and all language acts (Johnstone, 2002). Other linguists, however, use discourse as a count noun, defining a discourse as a set of linguistic features used in

specific contexts or by specific social groups (see Johnstone, 2002).<sup>15</sup> Both discourse definitions include all the factors needed for communicative competence such as knowledge of register, sociolinguistic factors, and purpose.

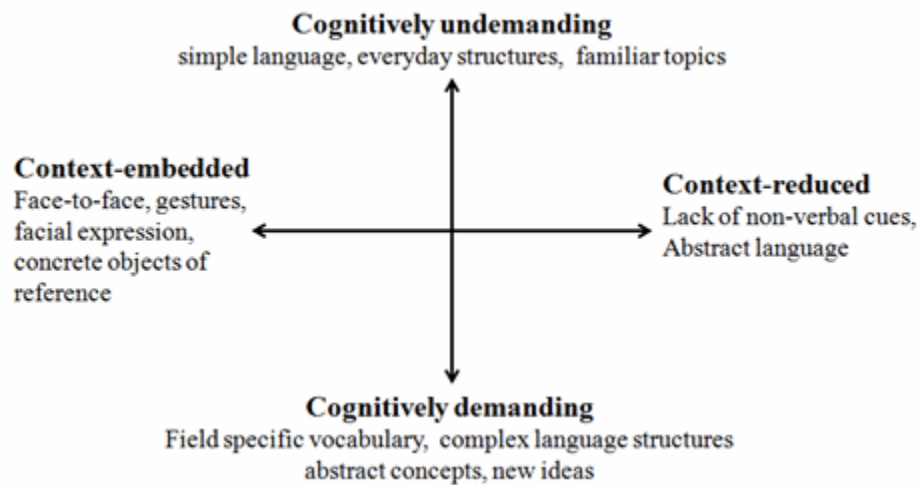
This study analyzed the relationship between the language of the June 2012 Biology MCAS, specifically item linguistic complexity, and ELL performance. The language of the Biology MCAS includes language used by biologists (an occupation) and by those who are considered “educated” in the greater context of our society (social group). Thus, the Biology MCAS has its own discourse. The linguistic features of the Biology MCAS can be viewed as the dual registers of biology and academia situated within their respective larger discourse communities. The delineation between which linguistic elements are features of a register, a discourse, or both was not addressed in this study. Instead, this study adopted Schleppegrell’s (2001) definition of a register as lexical and grammatical features situated in the larger context of the discourse community.

**Cummins’ model of BICS and CALP.** Cummins (1979) coined the term cognitive academic language proficiency (CALP) to describe the language of school (cited in Cummins, 2003). Cummins contrasted CALP with everyday conversational language, which he termed basic interpersonal communication skills (BICS). These terms have been widely accepted in the applied linguistics and education fields to classify language according to context and purpose, and as such, these terms describe different language registers (see Schleppegrell, 2004). The distinction between BICS and CALP

---

<sup>15</sup> James Gee (1999) differentiated these as discourse and Discourse. When discourse is in lowercase letters (or “little d” discourse), it refers to communicative language acts (the non-count noun). Discourse, when it is capitalized (or “big D” discourse), refers to the particular linguistic features used by specific groups of speakers or Discourse communities (the count noun).

underscored the difference in how long it takes second language learners to acquire these different registers and identified a source of academic challenge for second language learners (Cummins, 2003). Cummins further categorized language into four quadrants along two continua: (1) from highly contextualized to decontextualized, and (2) from highly cognitively demanding to cognitively undemanding (see Figure 2.2).



*Figure 2.2.* Cummins' Four-Quadrant Model. This model illustrates the intersection of context and cognitive demands of language.

Source: Madyarov (2009)

Using the four-quadrant model, BICS is contextualized and cognitively undemanding, whereas CALP moves toward greater decontextualization and higher cognitive demands. In contextualized language, the negotiation of meaning is aided by visual, tonal, gesticulatory, and other situational elements.<sup>16</sup> Decontextualized language, on the other hand, is language taken out of context with minimal or no situational clues to

<sup>16</sup> For example, the question “Is there any salt?” can have different meanings depending on the context. If this question is spoken by someone eating a meal, it could be interpreted as a request to pass the salt or as a criticism of the cook (among other interpretations). If it is spoken by someone when car wheels are spinning on ice, it can act as a suggestion in problem-solving.



aid negotiation of meaning.<sup>17</sup> The assumption that language can be either contextualized or decontextualized is one criticism of Cummins' BICS/CALP four-quadrant model (see Aukerman, 2007). Aukerman (2007) maintained that the belief that decontextualized language proficiency is requisite for academic achievement is "ultimately destructive" (p. 18) and contributes to a deficit model for ELLs. Cummins (2003, as cited in Edelsky, 1990; Edelsky et al., 1983; Martin-Jones & Romaine, 1986) acknowledged the criticisms of CALP as a deficit model and responded that these criticisms arise from interpretations of CALP within sociopolitical constructs; no critic "has disputed the basic realities from which the distinction [of CALP] derives" (p. 326). Schleppegrell (2004) argued for a functional linguistics perspective focusing on the linguistic choices in academic texts rather than on categorizing language as decontextualized. She argued that all language is contextual and that problems with academic language arise from its contextualization in a discourse unfamiliar to many students. Snow and Uccelli (2009), however, believed that academic language is "less obvious and less accessible" (p. 123), a view congruent with Cummins' decontextualized and cognitively demanding language quadrant.

Aukerman (2007) also questioned the cognitively undemanding and demanding dimension of BICS and CALP; he used the example of a conversation with a friend who has cancer as social language use but not necessarily cognitively undemanding (i.e., difficulty in knowing what to say). The directional arrows in Cummins' four-quadrant model, however, represent a language continuum where some BICS is more cognitively demanding than other BICS and some CALP is less cognitively demanding than other CALP. As is the nature of continua, boundaries are not always clear (Bailey & Huang,

---

<sup>17</sup> For example, "The salinity of the ocean surface varies around the globe."

2011), and Schlepppegrell (2004) cautions that cognitive demand is dependent on the context and the student's background knowledge rather than intrinsic to the task or text.

Although influential, Cummins' model has been criticized for conflating proficiency and academic achievement, as well lacking the social, political, and power-dynamics elements of language (Bunch, 2006). The BICS/CALP distinction, however, was not intended to be an overarching language theory but rather an aspect of second language acquisition in the context of schooling (Cummins, 2003). This study adopted Cummins' distinction between BICS and CALP as different language registers situated along the language continua of contextualized/decontextualized and cognitively undemanding/demanding. It also adopted the view that CALP is needed for academic success, including performance on the Biology MCAS.

**Characteristics of academic language.** Academic discourse not only includes linguistic knowledge but also the sociolinguistic competence to make correct lexical and syntactic choices (Schlepppegrell, 2001, 2004). Lexical choices can be broadly thought of as choosing one word over another to convey meaning; syntactic choices can be broadly thought of as the grammatical constructions.<sup>18</sup> The purpose of the text (spoken or written) influences linguistic choices (i.e., register), and academic texts "typically structure information so that it can be presented efficiently and arguments can be hierarchically constructed for a non-interacting audience" (Schlepppegrell, 2001, p. 435). That is, linguistic choices function to relate particular text to prior text (Schlepppegrell, 2004). Snow (2010) adds the following features to academic language: "high density of

---

<sup>18</sup> Although this is a linguistic simplification of the concepts of lexemes and syntax, it approximates how these linguistic terms are translated into classroom pedagogy.

information-bearing words, ensuring precision of expression; relying on grammatical processes to compress complex ideas into few words” (p. 450); and an “impersonal, authoritative voice,” which adolescents may find perplexing (p. 451). This results in “significant grammatical differences” (p. 454) between academic and non-academic texts even for native speakers (Schleppegrell, 2001) because the learner must discern how the new register differs from familiar registers (Conrad, 1996). Cook et al. (2011) offer a less technical and more classroom-friendly definition of academic language as having “more complex grammatical forms, more technical vocabulary, less use of slang and idioms, clearer references, and a more objective sense” (p. 67).

Biber and Gray (2010) challenged the stereotype that although academic writing is grammatically complex and decontextualized, it is explicit in meaning. The authors conducted a large corpus-based analysis of academic research articles in four disciplines at three intervals over a 60-year period. The authors found that academic language was explicit with respect to clear referents; however, at the sentence or grammatical level, academic discourse was inexplicit in the “logical relations among elements in the text” (p. 18). Based on their analysis, the authors concluded that academic writing (at the research and university levels) is a distinct English register that differs from other written registers (such as prose). While not problematic for experienced readers, the inexplicit aspects of this register can make comprehension difficult for novices to academic discourse. Schleppegrell (2004) noted that although studies characterize explicitness by linguistic features such as the lexicalization of referents (i.e., use of full noun phrases and avoidance of pronouns or words that require situational context), these linguistic features are not synonymous with clarity. Whether a text is clear or explicit depends on the

reader's background knowledge, including familiarity with the discourse structure; that is, explicitness is relative.

Academic language can be categorized along two dimensions: use and linguistic level. The first dimension, use, divides academic language into general and discipline-specific. General or non-content academic language includes the vocabulary (e.g., *nevertheless*) and structures (e.g., passive voice) found across disciplines (Schleppegrell, 2001). In addition to general academic language, disciplines have their own particular academic discourse that includes discipline-specific vocabulary and syntactic structures (Bailey & Huang, 2011; Cook et al., 2011; Liu, 2012; Schleppegrell, 2001, 2004; Tan, 2011). Different disciplines have different discourse requirements for academic success (Hyland, 2002; Mackiewicz, 2004; North, 2005, as cited in Baik & Greig, 2009, pp. 403-404), and learning content cannot be separated from academic discourse (Schleppegrell, 2001).

Academic language can also be categorized by linguistic level: word (lexical), sentence (syntactic or grammatical), and discourse (Bailey & Huang, 2011). At the word level, Bailey and Huang (2011) divided academic vocabulary into general and discipline-specific (which they call specialized academic vocabulary), but they also added a third category of context-specific academic vocabulary. Context-specific academic vocabulary is vocabulary that has a different meaning when used in a content area—for example, the everyday word *by* has a different meaning (multiplication) when used in mathematics (Bailey & Huang, 2011). Researchers have used frequency in corpus studies to identify academic language at the vocabulary level. Coxhead (2000) developed the Academic Word List (AWL), a compilation of cross-disciplinary (general) academic words. The

AWL consists of 540 word families<sup>19</sup> and account for 10% of the words in textbooks, but only 1.4% of the words in fiction of comparable length (Coxhead, 2000). Lawrence, White, and Snow (2010), however, maintained that the AWL provides little insight into academic words encountered in middle school texts because it was developed using adult texts. I believe, however, that high school texts approach or reach adult level, especially in the higher grades. Thus, the AWL is salient for the texts secondary school students encounter. More recently, Simpson-Vlach and Ellis (2010) developed the Academic Formula List (AFL). The AFL is a corpus-derived list not of individual academic words but of collocations, lexical bundles, or formulaic language that appears in academic texts.<sup>20</sup> Both the AWL and AFL are important contributions to defining academic language at the lexical level.

At the sentence level, academic texts are lexically dense with linguistic features such as nominalizations and more ideas per clause (Schleppegrell, 2001).<sup>21</sup> Elaboration of nominal elements (nouns) with adjectives, prepositional phrases, embedded clauses, and subordinate clauses lead to increased complexity (Schleppegrell, 2004). Schleppegrell (2004) noted, however, that educational and linguistic studies need to define accurately the notion of linguistic complexity.

Acknowledging some controversy over findings on academic language and syntactic structures, Snow and Uccelli's (2009) academic language model also went beyond the word level (lexical) and sentence level (grammatical or syntactical) to a meta-

---

<sup>19</sup> A word family consists of a head word and its variations. For example, the word family for the headword *coincide* would include its variations *coincidence*, *coincidental*, and *coincidentally* (Coxhead, 2000).

<sup>20</sup> Collocation, lexical bundle, and formula refer to words that are commonly found together—e.g., *on the other hand*, *due to the fact that*, *in the case of* (see Simpson-Vlach & Ellis, 2010).

<sup>21</sup> Nominalization is the process of turning a clause, a verb, or another part of speech into a noun—e.g., *deforestation*, *habitat loss*, *measurement* (see Schleppegrell, 2001; Turk & Kirkman, 1989).

communicative or discourse level that encompass discipline-specific genres, reasoning, taxonomies, and salient relations. At the discourse level, academic language includes the language functions of “explaining, informing, comparing, describing, classifying, proving, debating, persuading, and evaluating” (Chamot & O'Malley, 1994, as cited in Bailey & Huang, 2011, p. 351). Lemke (1990) described science discourse as having the additional language functions of “hypothesizing, questioning, challenging, designing experiments, comparing, analyzing, evaluating, and generalizing” (cited in Bailey & Huang, 2011, p. 351).

This study adopted a notion of academic language as the lexical (vocabulary) and syntactic (grammatical) elements situated in discourse communities, and these language features can be classified as general or content-specific. This conformed to the notions of academic language proposed by Bailey and Huang (2011), Schleppegrell (2001), and Snow and Uccelli (2009). This study's notion of academic language also incorporated Bieber and Gray's (2010) finding that complex syntax may obscure the explicitness and clarity of academic language's lexical level and Schleppegrell's (2004) notion that clarity is relative to a student's background knowledge and experience.

**Academic language and science.** Reading is inferring meaning from text (Norris & Phillips, 2009), and reading comprehension is tied to academic language (Carlo et al., 2008; Dalton, 2011). Academic language is the “expertise in understanding and using literacy-related aspects of language” (Cummins, 2000, as cited in Aukerman, 2007, p. 7). Students are novices in their content area and its academic discourse, and full participation requires mastery of the discipline's discourse (Baik & Greig, 2009; Schleppegrell, 2001; Snow & Uccelli, 2009; van Goor & Heyting, 2008). Middle and

secondary school students who are unable to transition their word-reading skills into comprehension may find science texts challenging (Snow, 2010).

Norris and Phillips (2009) argued that science content is not enough; teachers must also focus on developing literacy in the discourse of science. Science teachers, however, often lack the skills and training to help students access academic discourse beyond content-specific vocabulary (Snow, 2010), and, historically, literacy instruction has been absent in the science classroom (Norris & Phillips, 2009). Snow (2010) suggested incorporating explicit teaching of general academic discourse into the science curriculum, including the general academic language that is used to define content-specific terms. I agree with Norris and Phillips (2009) and Snow (2010): Secondary science classes, especially SEI science classes, must incorporate the dual dimensions of content and academic literacy, which includes both general and discipline-specific academic language.

Since reading skills and strategies can differ across disciplines, Norris and Phillips (2009) examined the notion of scientific literacy (or discourse) as distinct from general academic literacy. Citing Myers (1991), the authors pointed out that science discourse requires scientific meta-language and that a dictionary for content-specific vocabulary is insufficient for comprehension since “much of the difficulty interpreting scientific text lies in grasping the connections of one statement to another” (p. 276). Bailey, Butler, Stevens, and Lord (2007) found that fifth-grade science texts contain more prepositional phrases, noun phrases, and passive constructions than mathematics texts (cited in Bailey & Huang, 2011). Several other linguistic features of scientific texts—their rhetorical nature, their distinction between facts and interpretations, their recognition of human

agency and personal attitudes—also present challenges to novices (Conrad, 1996; Norris & Phillips, 2009).

Conrad (1996) examined the academic language in two university-level ecology textbooks and found that purpose, topic, and audience influenced linguistic features. She concluded that the variation of linguistic elements between the texts illustrated not only the complexity of academic language but also the need to consider multiple linguistic elements. The author acknowledged that a limitation of the study was that only two ecology texts were analyzed. Stapp (2003) listed three characteristics of science concepts that may make general ELL reading comprehension strategies insufficient; one is language-related, and two are discipline-related. The language-related characteristic is at the lexical or word level. Science terminology is precise with little room for guessing or approximation of meaning. In addition, some science terms are context-specific with meanings that differ from their common usage. Comprising the discipline-related characteristics, (1) many science concepts are abstract, and (2) science concepts can be counter-intuitive. Language surrounding abstract or counter-intuitive concepts is both decontextualized and cognitively demanding, and Stapp cited several studies criticizing science textbooks for “failing to anticipate the much more concrete perspective of the naïve learner” (p. 32).

**Academic language and ELLs.** Cummins’ (2000) common underlying proficiency (CUP) model allowed for positive transfer between L1 academic language and L2 academic language. Notwithstanding CUP’s positive transfer of academic language elements between languages, academic language is challenging for ELLs, and much of the empirical research on academic language is in the context of second



language learners (Snow & Uccelli, 2009). The WIDA consortium currently focuses on ELLs, academic language, and assessments. In 2010, WIDA, together with WestEd, established the Madison Academic Language Working Group (MALWG) at the University of Wisconsin to develop “core components of the academic language construct to support student growth” (“Academic language,” 2011), and WIDA has defined English proficient as the point where language becomes “less related to academic achievement” (Cook et al., 2011, p. 67).

ELL students may meet the minimum English proficiency requirements for post-secondary education; however, they struggle to meet the increased demands of academic language (Birell, 2006; Bretag, 2007; Pantelides, 1999, as cited in Baik & Greig, 2009). Baik and Greig (2009) studied the effectiveness of discipline-specific interventions for ESL university students ( $n = 37$ ) studying architecture, building, and planning at the University of Melbourne. The authors found a positive relationship between discipline-specific language support and student performance, which was consistent with earlier studies of the effectiveness of content-based ESL programs.

It is widely accepted in the literature that acquiring academic discourse in a second language takes between four and seven years, and some studies suggest that it may take up to nine years. Collier (1987a) studied ELL students ( $n = 1,548$ ), whose age of arrival ranged from 5 to 15 years, and who were on grade level but had no previous exposure to English. Collier found that students who arrived between the ages of 8 and 12 years were the first to reach native-speaker norms on content assessments, taking four to five years. Students who arrived at earlier ages (5 to 7 years) took five to eight years to reach native-speaker norms. The 12- to 15-year-old group, however, exhibited the most

difficulty and took the longest—six to eight years. These results suggest that late-entry ELLs need the most time to achieve academic language parity with native speakers.

Hakuta, Butler, and Witt (2000) used existing data from two Canadian ELL studies and new data from two U.S. schools to explore English proficiency as a function of length of exposure to English. The authors acknowledged that ideal data would be longitudinal from the time of arrival to the acquisition of English proficiency; however, there were no existing databases with this data. The two Canadian datasets confounded length of immigration with age of immigration since the participants were studied according to grade level; the U.S. datasets did not. The findings indicated that it takes between three and five years for ELL students to acquire oral English but significantly longer to acquire academic language proficiency. For ELL students who enter in kindergarten, the data indicated that it takes between four and seven years to achieve academic language proficiency in English. Ninety percent of the students achieved academic language proficiency by the end of Grade 6 (seven years of schooling). The Canadian datasets did not address the issue of socioeconomic status (SES). The U.S. datasets, however, did show that SES was an important variable in predicting the rate of becoming proficient in English.<sup>22</sup> The authors cautioned that their findings may underestimate the time it takes to become proficient in English and that the ultimate effects of age of arrival on English proficiency were beyond the scope of the study.

---

<sup>22</sup> Martiniello (2008) also found a relationship between SES and English proficiency level.

**Academic language and achievement.** Development of academic language is by definition requisite for academic achievement, and a major goal of NCLB Title III is to help ensure that limited English proficient (LEP) children attain English proficiency, develop high levels of academic competence in English, and meet the same challenging state academic content and student academic achievement standards that all children are expected to meet. (cited in Abedi, 2008b, p. 202)

Without academic language, students may have problems not only accessing and communicating knowledge in the classroom but also performing on standardized assessments. At the higher grades, students need advanced literacy skills as they encounter unfamiliar concepts presented in dense and abstract language (Schleppegrell, 2004). Lack of academic language can not only reduce text comprehension and create problems learning content (Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006) but also create barriers to demonstrating content knowledge on standardized assessments (Martiniello, 2008). Thus, students with limited academic language are at risk for academic failure, especially at the secondary level where ideas become more abstract.

Secondary ELLs, however, do not have the “luxury” of waiting until they acquire academic language before they learn content (Bunch, 2006). Therefore, content classes for secondary ELLs must provide opportunities to acquire both content and academic English (Bunch, 2006). This, however, in turn presents another problem: The time spent developing academic language in the content classroom reduces the time spent on content (Abedi & Gándara, 2006). The amount of time late-entry ELLs have in high school, plus the time they need to acquire academic language, contributes to the tension between content and language instruction in the SEI science classroom.

## **Review of Literature on Critical Issues for Cognitive Load and Wide-Scale Assessments**

Assessment items have attributes, or characteristics. Some item attributes are the domain or content that the item is assessing as well as what skills are needed in that domain for a correct response. For example, an item may assess whether a test-taker can recall a fact, understand a concept, apply a concept, or integrate knowledge from different domains to solve the problem. Thus, items can be thought of in terms of complexity; some items require more complex reasoning than others. How an item is worded is also an item attribute. For example, an item could test the science concept of osmosis using language that is easily understood by most test-takers or it could use low frequency words that might interfere with some test-takers' understanding of the question.

**Item difficulty.** In classical test theory statistics, item difficulty refers to how many test-takers answer the item correctly (i.e., what percentage got it right). In item response theory (IRT) statistics, item difficulty refers to the probability that at a given level (competence) 50% will answer the item correctly. Item difficulty and item cognitive level are not the same. Within a cognitive skill level, some items are more difficult than others. Schneider, Huff, Egan, Gaines, and Ferrara (2013) raised the possibility that current test construction processes may not be capturing an interaction between item difficulty and item cognitive complexity. They also cautioned that item difficulty estimates are not without error. For example, a difficult item may appear less difficult because students have practiced this skill to the point where the item is no longer assessing the challenging aspect of the skill. Likewise, an item may appear more difficult if students have not learned the concept under assessment.

**Cognitive complexity.** Cognitive demands are “related to the number and strength of the connections within and between mental networks” (Webb, 1997, 1999, as cited in Jirka & Hambleton, 2005, pp. 6-7). Cognitive level is a dimension assigned to items during development (Ćurković, 2012), and cognitive frameworks can be used to explain or predict standardized test performance, including test score inferences and differential performance by subgroups (Gierl, Alves, & Majeau, 2010). In the attribute hierarchy model (AHM), test performance depends on hierarchal cognitive attributes (processes and skills requisite for correct responses), and these attributes are sources of cognitive complexity (Gierl et al., 2010). Jirka and Hambleton (2005) evaluated various models of cognitive complexity to recommend a model for the MCAS; models evaluated included Bloom’s original taxonomy, Bloom’s revised taxonomy, Webb’s depth of knowledge, NAEP’s mathematics framework, and NEAP’s reading frameworks. The authors concluded that more than four categories of cognitive skills could decrease reliability because of subtle differences among levels.<sup>23</sup>

Ćurković (2012) used structural equation modeling to explore the existence of different cognitive levels on the Croatian state-level summative high school math exam (n = 9,626). The study used an abbreviated version of Bloom’s taxonomy with only three cognitive levels (knowledge, comprehension, and application) and assumed that the cognitive levels were hierarchal. Results indicated strong unidimensionality but did not support the existence of different cognitive levels. The author concluded that cognitive

---

<sup>23</sup> Beginning in 2013, the Biology MCAS item development process assigns one of five cognitive skills that mostly correlate to Bloom’s revised taxonomy. Prior to 2013, the item cognitive skill levels were foundational, conceptual, application, constructive, and quantitative.

levels should remain part of item development and test construction; however, there is a need for better operationalization of cognitive levels.

**Cognitive load.** Generally, human cognitive processing is thought of as having two components: working memory and long-term memory. Working memory includes those aspects of cognition that are currently being attended to and noticed; working memory has capacity and duration constraints (Paas, Van Gog, & Sweller, 2010). Humans transform information from how it is presented into schemas that are stored in long-term memory (Sweller & Chandler, 1991), which has unlimited storage capacity (Paas et al., 2010). The schemas stored in long-term memory determine performance in a given area or domain, and they reduce demands on working memory because they are processed as one element (Paas & Sweller, 2012; Paas et al., 2010).

Tasks differ in information processing load, and more complex tasks require different cognitive processes (Dodonova & Dodonov, 2013). Cognitive load comprises the amount of mental processing needed for a task and its demands on working memory (Shank, 2007). Working memory can process 7 +/- 2 units of information at a time; however, some recent studies indicate that this number may be too high, especially for novice learners (Paas et al., 2010; Shank, 2007). Ozinar's (2009) examination of research trends found that the "cognitive load theory" was the second most used phrase (as cited in Paas et al., 2010, pp. 115-116).

Cognitive load theory (CLT) focuses on capacity-limited working memory (Van Gog, Paas, & Sweller, 2010). In learning, cognitive load can only be managed, not eliminated, because it is part of the learning process (Shank, 2007). Congruent with Vygotsky's zone of proximal development, CLT posits that there is an optimal level of

intrinsic cognitive load for learning; too many interactive elements can overwhelm learners and negatively impact understanding (Paas et al., 2010; Shank, 2007).

Unnecessary cognitive load should be eliminated when a learner is new to a domain (Paas et al., 2010; Shank, 2007). In assessments, vertical alignment of achievement level descriptors and cognitive complexity is becoming more common (Schneider et al., 2013).

Cognitive load is affected by several factors, including the content and the number of items that require attending (Shank, 2007). The higher the interactivity between elements (items), the more demand on working memory resources, and the higher the cognitive load (Paas et al., 2010). Experienced learners can retrieve schemas from long-term memory, which can be treated as one element in working memory; however, novices use means-end analysis that requires cognitive resources that may be irrelevant to the task (Sweller & Chandler, 1991). “[T]he more difficult one factor is (e.g. the language), the less attention can be dedicated to another (e.g., the content)” (Dickey, 2004, p. 11). Novice L2 readers experience high cognitive load because they must attend to and integrate the interacting elements of vocabulary, sentence structure, and syntactic rules for comprehension, which may or may not be successful (C. H. Lee & Kalyuga, 2011). As L2 readers gain automaticity, this frees up working memory resources and reduces cognitive load (C. H. Lee & Kalyuga, 2011).

There are three types of cognitive processing that can contribute to cognitive load: (1) extraneous, (2) intrinsic, and (3) germane or generative (DeLeeuw & Mayer, 2008). Extraneous cognitive load can be increased if there is redundancy in the task, such as two information streams that must be processed (e.g., text and animation, or text and diagrams). Sentence complexity is one form of intrinsic load because more complex

sentences require holding more information in working memory for comprehension. Germane or generative cognitive load is related to making connections between the task and prior knowledge. Extraneous and intrinsic cognitive loads are additive and problematic only if the combined cognitive load overwhelms the learner. A task can have high extraneous cognitive load if the intrinsic cognitive load is low, and tasks with a high intrinsic cognitive load should have low extraneous cognitive load (Van Merriënboer & Sweller, 2005).

DeLeeuw and Mayer (2008) explored redundancy and sentence complexity as sources of cognitive load for college students ( $n = 58$ ). They used a  $2 \times 2$  mixed design with redundancy as the between-subjects factor and sentence complexity as the within-subject factor. Participants were divided into redundant ( $n = 28$ ) and non-redundant ( $n = 28$ ) groups, and each group was given tasks with four low and four high complexity sentences defined by the number of interacting concepts. Response time was the measure of extraneous cognitive load based on the hypothesis that redundant words waste cognitive resources and thus lead to a slower response time. Self-reported effort rating was the measure for intrinsic cognitive load based on the hypothesis that a learner has to work harder for comprehension when sentence complexity is high. Thus, if response time is a valid measure of cognitive load, then the non-redundant group should have a shorter response time than the redundant group for the higher complexity sentences.

The authors only found a marginally longer response time for higher complexity sentences in the redundant group, which was not statistically significant ( $p = .06$ ), and there was no interaction between redundancy and sentence complexity. With respect to the main effect of sentence complexity, the results indicated a significant ( $p = .001$ ) effect



in the rating of mental effort, that is, intrinsic cognitive load; there was no interaction between sentence complexity and redundancy. A second experiment likewise divided college students ( $n = 99$ ) into non-redundant ( $n = 49$ ) and redundant ( $n = 50$ ) groups. Again, there was a significant ( $p < .001$ ) effect of sentence complexity on mental effort. The authors concluded that manipulation of intrinsic cognitive processing through sentence complexity had a significant effect on mental effort ratings. They suspected that prior knowledge could have been a factor and suggested that future studies take this into account.

Schneider, Huff, Egan, Gaines, and Ferrara (2013) investigated the relationships among item cognitive complexity, contextual demand, and item difficulty. The authors looked at the reading load (intrinsic) demand of mathematics multiple-choice items for Grade 4 ( $n = 64$ ) and Grade 8 ( $n = 64$ ); reading load was the amount and complexity of textual and visual information a student needed to process for comprehension and a correct response. Reading load can be a source of construct-irrelevant variance in content assessments. The results did not indicate that reading load was a consistent predictor of item difficulty; however, the authors did note that a 2011 study by Ferrara showed a consistent predictive value of reading load on item difficulty. The authors concluded that more research is needed on the relationship between cognitive complexity and reading load.

Item features can affect the cognitive process of mental representation, which follows comprehension and sets the stage for problem-solving (Leighton & Gokierto, 2005). The cognitive effects of item vocabulary has emerged as an area of interest for item developers because items with polysemous words or irrelevant introductory

information can result in incorrect problem-solving assumptions and thus to incorrect responses; however, these item features are only problematic if students are unable to derive meaning from item context and surrounding text. (Leighton & Gokiart, 2005). These item features can be classified into as either radical or incidental elements. Radical elements are words that, if changed, alter the construct under assessment, whereas changing an incidental element does not alter the construct (Leighton & Gokiart, 2005). The study of item features such as vocabulary can identify and distinguish between construct-relevant and construct-irrelevant variances (Leighton & Gokiart, 2005). Leighton and Gokiart (2005) used the example of *clarify*, a word that might be unknown to some test-takers. Differential performance on an item between two test-takers could be due not to differences in content knowledge (the construct under assessment) but rather to understanding the meaning of *clarify* (a construct not under assessment).

Leighton and Gokiart (2005) studied 30 publically released practice test items from the School Achievement Indicators Program (SAIP) science assessment, Canada's national standardized assessment. The SAIP science has three broad content domains and five ability levels. Their study evaluated: (1) words and phrases, (2) context, and (3) item format. Their criteria for misalignment of words and phrases included non-scientific words with ambiguous (e.g., *spoiled*) or multiple meanings, and low frequency non-scientific words that a Grade 8 or Grade 11 student might not know (e.g., *site*) and which were not defined in the item. Their method was a one-hour think-aloud while solving problems (Grade 8 n = 30; Grade 11 n = 24).

Results indicated that items with problematic features, such as words with multiple meanings, can slow cognitive processing and lead to faulty assumptions and

incorrect responses. Thus, items with these features can generate construct-irrelevant variances. The authors drew three conclusions from this study. First, the majority of students had problems in paraphrasing the item. “Just over 80% of the students ... relied heavily on words and direct phrases from the item stem and question” (p. 15).

Paraphrasing is an important metacognitive skill in comprehension. Second, although most students could identify the knowledge and skills required to answer an item, most could not identify the general concept of the item. Third, item ability level was the only item feature that influenced students’ uncertainty about their item performance. The authors believed that this third conclusion supports their finding that students can access the knowledge and skills for a correct answer even in the absence of full comprehension of the item.

Believing that ambiguity is the central problem in sentence processing, Hale (2003) studied how sentence ambiguity resolution (entropy reduction) characterized a range of cognitive load patterns. The study adopted Chomsky’s competence hypothesis that linguistic knowledge is directly related to comprehension. The study assumed a combinatory relationship in incremental sentence processing and the reduction in cognitive load with ambiguity resolution. That is, the level of ambiguity is directly related to the amount of cognitive load. It also assumed that the producer (item writer) and the comprehender (test-taker) shared the same grammar (language). The author proposed that readers identify combinatory relationships and perform the maximum ambiguity resolution at each word; however, he did not propose this as a cognitive process but rather as a consequence of top-down parsing theoretical claims.

Paas and Sweller (2012) revisited CLT from an evolutionary perspective of biologically primary and biologically secondary information. Biologically primary information is information that humans have evolved to acquire such as language and facial recognition. Biologically secondary information, on the other hand, is cultural knowledge, including reading, among others. The authors posited that working memory limitations may be more critical for biologically secondary information because human evolution does not include how to process this type of information; it requires effort and cognitive resources. They further suggest that CLT is only relevant for learning biologically secondary information.

**Cognitive load and second language learning.** “Culture and language influence cognitive processes. Consequently, they may affect a person’s performance in cognitive tests” (Schaap & Vermeulen, 2008, as cited in Schaap, 2011, p. 138). Automaticity, such as in reading, reduces cognitive load by freeing up working memory. Thus, novice L2 readers have a higher cognitive load than more proficient L2 readers.

Robinson (2005) examined predictions made by his cognition hypothesis, which complements and extends Krashen’s comprehensible input hypothesis, and posited that SLA tasks should be sequenced by levels of increasing complexity (also see Robinson & Gilabert, 2007). His first prediction was that higher complex tasks would lead to greater accuracy and complexity of L2 output because increased cognitive demands would focus second language learners on the grammatical differences between the L1 and the L2. The second prediction was that tasks with higher cognitive demands would promote heightened attention and noticing, and increased interaction and negotiation. The third prediction was that individual learner characteristics such as cognitive ability and

affective factors would have increasing impact as task complexity increased. Robinson examined four studies and concluded that task complexity affected turn taking and interaction, but he did not find the predicted positive effects on L2 output. In a later study, Robinson and Gilabert (2007) found that increased task complexity resulted in fewer errors by second language learners, increased self-repair, and increased use of higher frequency words.

Michel, Kuiken, and Vedder (2007) tested Robinson's cognition hypothesis in a study of Dutch second language learners ( $n = 44$ ), whose first L2 contact was post-puberty. Using Guiraud's index of lexical complexity and the percentage of lexical words to total words, the authors found that more complex tasks produced higher accuracy in output, decreased structural complexity, and increased lexical complexity. The study included both monologic and dialogic task conditions, and results indicated that task condition had significant effects on accuracy, complexity, and fluency. Dialogic task conditions produced fewer errors and omitted words. Monologic task conditions resulted in a higher percentage of self-repair. There was no robust interaction between task complexity and task condition. These findings replicated earlier findings that increased task complexity had a positive effect on accuracy, a minor positive effect on lexical complexity, and a negative impact on fluency.

**Differential item functioning.** When groups have a statistically significant performance difference on an item, it is referred to as differential item functioning (DIF). Differential item functioning is not intrinsic to an item; it relates to how an item functions in the context of a particular assessment (Zenisky, Hambleton, & Robin, 2003). Differential item functioning studies are routine in item development; however, studies

exploring potential sources of DIF are less common (Zenisky et al., 2003). The conventional approach to studying DIF—that is, submitting statistically identified DIF to content reviewers—can result in inconclusive reasons for group differences (Gierl et al., 2010).

Using a variation of Doran and Kulick’s standardization procedure, Zenisky, Hambleton, and Robin (2003) employed six sets of data from wide-scale science assessments (each  $n = 60,000$ ) to explore DIF and non-DIF trends for the item attributes of content, cognitive demands, and text, among others. The authors found that item content, cognitive demand, and negative wording could be potential sources of DIF; however, since DIF does not equal bias, the authors were unsure how their findings should be addressed in testing situations. In later psychometric analyses of the 2006 Biology MCAS, Hambleton, Zhao, Smith, Lam, and Deng (2008) found no item DIF for gender and ethnicity; they did not explore DIF for ELLs.

Linguistic differences can also be a source of DIF, even in non-verbal instruments. Schaap (2011) investigated DIF on a non-verbal instrument, the PIB/SpEEEx Observance Test (401), which requires respondents to identify differences or similarities among shapes, figures, and pictures in 22 items. Participants were adult speakers representing five first languages ( $n = 5,971$ ); the mean age was 20 years, and the age range was 17 to 59 years. Classical test analyses showed clear differences among first language groups and raised the issue of non-equivalence and bias. Factor item analyses showed unidimensionality, and further analyses indicated statistically uniform DIF for most items and non-uniform DIF for two items. The author, however, stated that the overall effect size was small and, from a practical perspective, negligible. Nine items

(40%) showed practically important DIF for African language speakers versus Afrikaans speakers. At the group level, however, uniform DIF was not consistent across different first language groups. There were five uniform DIF items with negative impact for IsiZulu speakers. There were four DIF items, both negative and positive, for the Setswana-speaking group, and three items for the Northern Sotho-speaking. This study showed that for this particular instrument, different first language groups similarly interpreted and processed visual images; however, first language group had different effects on item functioning for a number of items.

### **Review of Critical Issues for ELLs and High-Stakes Testing**

A high-stakes test is an assessment whose outcome has major consequences for the test taker and, under NCLB accountability measures, for the school and district as well (Solorzano, 2008). Under this definition, the Biology MCAS is a high-stakes exam: Students must pass a science MCAS exam as a graduation requirement, and the measure of a school's success under the composite performance index measure (CPI) of the NCLB waivers includes student performance on the Biology MCAS (MA DESE, 2012d). This section discusses the assessment of ELLs on content area standardized tests constructed and normed for an English-proficient population. Since academic language is needed to demonstrate knowledge on standardized tests, this section also includes a discussion of language impact on ELL achievement, including issues of validity and reliability. The section finishes with a discussion of accommodations for ELLs on wide-scale assessments.

**Assessment of ELLs.** “Since 1905, it has been clear that one can link the results of psychometric tests to class and/or culture. The fair testing of people from highly

dissimilar backgrounds therefore poses a challenge for those who apply tests” (Schaap, 2011, p. 137). Historically, ELLs have not performed on par with native English speakers on standardized exams (Abedi, 2002, 2009; Abedi & Dietel, 2004; Abedi et al., 2004; Cook et al., 2011; Menken, 2008, Chapter 4; Solorzano, 2008). ELL performance on standardized exams, however, may not be an accurate measure of content knowledge (Abedi, 2002, 2008b; Abedi & Gándara, 2006; Martiniello, 2008). Since standardized assessments are written for and normed on native English speakers, these assessments become de facto dual assessments for content knowledge and academic English proficiency (Abedi, 2002, 2008b; Abedi & Gándara, 2006; Solorzano, 2008). In criticizing Cummins’ BICS/CALP model, Edelsky (1996) described academic language as nothing more than “test-wiseness” (as cited in Cummins, 2003, p. 324), and Bailey and Huang (2011) questioned whether standardized assessments “have ignored the unique characteristics” of academic language (p. 349).

The American Educational Research Association (2000) recommends that an assessment should not be used with a student who does not understand “the language of the test” (as cited in Solorzano, 2008, p. 262). An ELL designation indicates lack of English proficiency, which in turn indicates potential problems in understanding the language of the assessment. Although other factors such as poverty and parent education impact ELL achievement, language has the greatest impact, with increasing gaps as language demands increase (Abedi, 2002, 2008b, 2009; Abedi & Gándara, 2006; Abedi et al., 2004). ELLs may have problems accessing the academic language of standardized assessments at all three levels: lexical (word), syntactical (grammar), and discourse (discipline conventions).



Native speakers have a vocabulary between 5,000 and 7,000 words when they begin formal reading instruction (August, Carlo, Dressler, & Snow, 2005). Beginning-level, late-entry ELLs begin to read in English knowing only a few English words, and even though intermediate ELLs have some English vocabulary, it is inferior to that of native speakers when they begin to read, let alone to that of their native-speaking peers in secondary school. Vocabulary is linked to reading comprehension, and ELLs' limited breadth and depth of vocabulary leads to a persistent reading achievement gap between them and native English speakers (August et al., 2005). The ability to read a standardized exam is a factor in performance, and Abedi and Lord (2001) found that low frequency vocabulary contributed to the ELL performance gap. Lawrence, White, and Snow (2010) found that middle school students who improved their vocabulary scores in Word Generation, an academic vocabulary development program, also improved their MCAS scores; however, the authors did not suggest that improvement in academic vocabulary by itself leads to higher MCAS performance. These findings on the impact of vocabulary on ELL test scores support Krashen's input hypothesis that the test must be comprehensible.

At the sentence or syntactic level, linguistic complexity can impact ELL performance on standardized assessments (Abedi & Gándara, 2006; Abedi & Lord, 2001). Geva, Yaghoub-Zadeh, and Schuster (2000) found that ELLs lacked the reading skills of syntactic awareness that their native-speaking peers possessed (as cited in Abedi & Gándara, 2006, p. 38). Abedi, Lord, and Plummer (1997) listed some syntactic elements that increased linguistic complexity for the reader: "long noun phrases, long question phrases, passive voice constructions, comparative structures, prepositional

phrases, sentence and discourse structure, subordinate clauses, conditional clauses, relative clauses, concrete versus abstract or impersonal presentations, and negation” (as cited in Abedi & Gándara, 2006, p. 41). These linguistic features are the same elements used to describe academic language at the syntactic level. In other words, the linguistic complexity that impacts ELL performance on wide-scale assessments is academic language.

**Impact of linguistic complexity.** Abedi and Lord (2001) explored the impact of test item linguistic complexity on performance on the Grade 8 Mathematics NAEP exam. They conducted two studies: (1) examination of student perceptions of test items, and (2) investigation of the impact of linguistic complexity, ELL status, SES, gender, and course of mathematics study on mathematics word problem performance and whether these factors interacted to affect performance. Both studies used 69 released math items from the 1992 NAEP main math assessment with linguistic modifications to non-content vocabulary and structures. There was no modification of content-specific vocabulary, and two experts found that the mathematics content was parallel between the original and revised test items. Thus, mathematics construct validity was maintained.

The first study ( $n = 36$ ) found that students’ preference for the linguistically simplified items was statistically significant. This finding indicated that non-content linguistic modifications could make test items more accessible to students. The second study ( $n = 1,174$ ) examined the impact of the linguistically simplified items on mathematics performance. The authors found that ELLs ( $n = 372$ ) scored lower than non-ELLs ( $n = 802$ ;  $p < .000$ ) and that high-SES ( $n = 725$ ) students performed better than low-SES students ( $n = 449$ ;  $p < .000$ ). ELL status explained 4.1% of the variance, and the

findings suggested that ELL status and SES are confounding factors ( $p < .01$ ). The mathematics class the students were enrolled in, however, had the greatest impact on performance.

A possible limitation of the study is that the authors only selected 20 of the 69 NAEP test items for linguistic simplification based on judgments of which items were most likely to impede performance. Duran and Moreno (2004) found that cognitive load can be a factor in ELL performance (as cited in Duran, 2008); the shorter revised tests could have had a lower cognitive load demand than the original NAEP exam and a hidden impact on ELL performance. Another possible limitation is differential impact of an item's linguistic structures because of heterogeneity in the ELL population.<sup>24</sup> Since this study used released NAEP items, it is also possible that some teachers had incorporated these in their instruction, which might explain the finding that the mathematics course the student took had the greatest impact on mathematics performance. That is, the more advanced mathematics courses may have incorporated more academic discourse into their curriculum either through implicit or explicit teaching, and this could have been a hidden factor.

Abedi and Lord's (2001) study confirmed earlier studies that there is an ELL achievement gap on the NAEP. Their study was exploratory, but it showed that the linguistic complexity of test items impacts mathematics performance, especially for ELL and low-achieving students. The authors called for continued research into the role of language in content assessment. This study explored the relationship between three levels of item linguistic complexity and ELL performance on the Biology MCAS.

---

<sup>24</sup> See Solano-Flores and Li (2009) for a discussion of the heterogeneity among Spanish-speaking ELLs.

Abedi (2002) explored whether language factors could explain the ELL achievement gap. The study examined ELL achievement data from four sites; one site provided data from the Iowa Tests of Basic Skills, and the other three sites provided data from the Stanford Achievement Test, 9<sup>th</sup> edition. Each of the sites provided data for different grade levels, but when taken together, the data ranged from Grade 1 to Grade 11. The results indicated that the impact of language factors increased as the content area's language demands increased.<sup>25</sup> Higher grade levels have increased linguistic demands for academic English, and the author found that the ELL achievement gap increased as grade level increased.

Martiniello (2008) explored differential item functioning for the 2003 Grade 4 Mathematics MCAS (n= 68, 839; ELLs = 3,179) in terms of linguistic complexity and strand for Spanish-speaking students. Ten items were identified as having a differential favoring non-ELLs,<sup>26</sup> and the author studied six of these items. Martiniello found that measures of lexical and syntactic complexity could estimate item difficulty for ELLs. Vocabulary knowledge is related to reading comprehension, and the author found that when an item was lexically dense (i.e., too many unknown words), this impacted comprehension of the question. The author also found lexical complexity in polysemous words<sup>27</sup> and words that required cultural knowledge, such as *chores* or *coupon*. In one part of the study, the author conducted think-aloud interviews with ELLs and found that only the most common meaning of polysemous words was known. She also found that

---

<sup>25</sup> In this study, however, Abedi (2002) classified science as a content area with less language demand.

<sup>26</sup> A differential favoring non-ELLs means that for students with similar overall scores, non-ELLs were more likely to do better on a particular test item than ELLs.

<sup>27</sup> Polysemous words have more than one meaning; meaning is context-dependent. Martiniello (2008) gives the example of *one*, which depending on context, can be a pronoun or a numeral.

ELLs used knowledge of Spanish-English cognates for unknown vocabulary; however, this was only a useful strategy if the Spanish cognate was known. The author believed that the finding on cultural references was probably generalizable to all ELLs but that the finding on using cognates was generalizable only to ELLs from Romance and Germanic language backgrounds. The author also found that syntactic elements such as multiple clauses and long phrases limited syntactic transparency and likewise affected the comprehensibility of the test item for ELLs.

Martiniello's (2008) study confirmed linguistic complexity as one source of differential item functioning and showed that language and content were intertwined for ELLs, who need "sustained linguistic scaffolding" (p. 363).<sup>28</sup> The study also found that some content strands (data analysis, statistics, and probabilities) had more items flagged for differential functioning, though the author only speculated as to several possible explanations, including greater lexical and syntactic complexity in these strands. The author called for further studies to examine the interaction between different learning strands and linguistic complexity as a source of differential performance for ELLs.

A limitation of Martiniello's (2008) study is that it only explored differential item functioning for Spanish-speaking ELLs; thus, the author called for further studies to examine linguistic complexity and differential item functioning for ELLs from other language backgrounds. The current study built on Martiniello's findings by analyzing ELL Biology MCAS performance across three levels of item linguistic complexity for ELLs with Latinate and non-Latinate first languages and with alphabetic and non-

---

<sup>28</sup> The study also found that higher socioeconomic status was linked to higher English language proficiency; Hakuta et al. (2000) found a similar relationship.

alphabetic first language orthographies. Martiniello also called for future studies that account for ELL proficiency differences. The current study addressed this by analyzing ELL performance by English proficiency levels on the Biology MCAS and for each of the five content domains and three item cognitive skill levels for the multiple-choice items.

**Validity and reliability.** No Child Left Behind accountability measures require the testing of ELLs; however, the rush to include them in high-stakes testing may have been at the expense of sound psychometric measures (Solorzano, 2008). Wide-scale standardized assessments have a lengthy and reiterative test construction process that creates an assumption of validity and reliability (Solorzano, 2008). Although this may be true for native or proficient English speakers, the testing of ELLs presents challenges to validity and reliability assumptions. The notion of validity refers to the soundness of inferences drawn from the data (Solorzano, 2008; also see Duran, 2008), and the assumption that the test-taker understood the item underlies inferential validity (Leighton & Gokiert, 2005). Any assessment that uses language is also assessing language skills, and validity requires that the language of the assessment reflect the language of learning (Abedi, 2002; Abedi & Gándara, 2006; Martiniello, 2008; Solorzano, 2008). The National Research Council (NRC) has warned that results for ELLs on tests written for and normed on native speakers may lack validity because “if a student is not proficient in the language of the test, her performance is likely to be affected by construct-irrelevant variance—that is, her test score is likely to underestimate her knowledge of the subject matter” (as cited in Abedi & Gándara, 2006, p. 39). Abedi echoes the NRC’s position and believes that designing and norming content assessments on non-ELLs not only “could

seriously undermine the validity of content-based assessments for ELLs” (2008b, p. 203) but could also “lead to inappropriate instruction and create invalid inferences about ELL academic achievement” (2008a, p. 28).

The impact of language factors brings into question the appropriateness of using assessments developed for native speakers to test ELLs (Abedi, 2008b; Abedi & Gándara, 2006; Solorzano, 2008; also see Duran, 2008). A construct-irrelevant factor introduces an irrelevant variable that has the potential of affecting the interpretation of results (Solorzano, 2008). Linguistic factors in content exams are a source of construct-irrelevant variance because they are unrelated to the construct being assessed (content knowledge), and they are potential sources of measurement error in estimating the assessment’s reliability (Abedi, 2002; Abedi & Gándara, 2006). Using data from four sites and across grade levels ranging from Grade 3 to Grade 11, Abedi (2002) found low test item reliability for ELLs with low English proficiency. He also found a lower statistical fit for ELL models compared to non-ELL models and hypothesized that language factors introduced construct-irrelevant variance.

Building on their 2006 study of Haitian Creoles, Solano-Flores and Li (2009) explored the assumption of linguistic homogeneity as a source of measurement error for Spanish-speaking ELLs. The study examined the performance of Grade 4 and Grade 5 Spanish-speaking students at three sites ( $n = 90$ ) in bilingual transitional programs on an instrument based on Grade 4 NAEP mathematics items. They used two dual-language versions of the instrument: (1) standard English and standard Spanish, and (2) standard

Spanish and a Spanish dialect.<sup>29</sup> The authors found no statistically significant difference in performance between the two versions; however, they found inconsistent ELL performance across items and languages. The authors believed that the inconsistent performance resulted from a combination of each ELL's unique strengths and weaknesses in both languages and the different linguistic challenges in the items. Score dependability varied among standard Spanish, Spanish dialect, or standard English; however, standard Spanish produced more dependable scores than the Spanish dialect. The number of items for a dependable score also varied across and within language groups. The authors found the greatest variation in the interaction among language, item, and student (41% to 48%). The study confirmed the construct-irrelevant aspect of language in ELL testing and indicated that even within broad language groups, generalizations can vary in appropriateness.

Research has shown that English proficiency impacts ELL achievement on standardized tests (Solorzano, 2008). The varying definitions of English proficiency and how ELLs are re-designated as English proficient create additional concerns for norming, validity, and reliability (Abedi et al., 2004; Solorzano, 2008). Another concern is that ELL curricula may limit access to academic language, which would also raise validity issues if the language of instruction is not the language of assessment (Solorzano, 2008). Abedi (2008b) argued that ELLs should participate in standardized content assessments only when English proficiency assessments show that language proficiency matches the language of the content assessment. As discussed in Chapter 1, this is not an option for

---

<sup>29</sup> In linguistic terms, standard English is a dialect of English, and standard Spanish is a dialect of Spanish. Standard English and standard Spanish refer to the dialects of those languages that are used in formal contexts, mass communication, schooling, and in the wider society by those with power.



secondary ELLs in Massachusetts public schools. For Massachusetts ELLs, enrollment in Grade 10 or higher and not having been included in a previous high school science AYP calculation—not language proficiency—determines whether they take the Biology or other science MCAS.

Research is needed to determine the extent of the impact of English proficiency on ELL achievement on standardized tests (Solorzano, 2008) and to inform the construction of reliable and valid content assessments for ELLs (Abedi, 2008b). The current study analyzed the relationship between ELL Biology MCAS performance and English proficiency to inform whether validity concerns exist at all levels of English proficiency. The current study adopted Abedi's (2002) findings that standardized, wide-scale assessments have reliability issues for ELLs at the lower English proficiency levels. To eliminate low English proficiency as a source of validity and reliability issues, this study analyzed ELL performance by English proficiency level, which allowed performance analyses for ELLs with MEPA scores at level 3 or above.<sup>30</sup>

**Accommodations.** Accommodations are one way to address validity and reliability issues with subpopulations. Accommodations are changes in the test or its administration, and they should be designed so that they do not affect the assessment's validity or reliability (Hakuta et al., 2000). "An ideal accommodation must be effective, valid, and appropriate to the background of recipients, while at the same time feasible" (Abedi et al., 2004, p. 16). An accommodated assessment is not the same as a modified assessment. A modified assessment is an alternative assessment that differs from the one

---

<sup>30</sup> In Massachusetts, the MEPA assesses English proficiency levels on a scale of 1 to 5, where 1 is the lowest English proficiency level.

given to the general test population. Compared to the regular assessments, a modified assessment may have reduced linguistic complexity, items of less difficulty, fewer items, assess fewer skills (Duran, 2008), or have alternate ways of demonstrating knowledge, such as a cumulative portfolio of work. The distinction between accommodations and modifications has been blurred in the literature pertaining to ELLs, and linguistic modification of test items is routinely referred to as an ELL accommodation. The current study likewise treated linguistic modification as an accommodation.

ELL accommodations can be broadly categorized into two types: test environment accommodations and linguistic accommodations. Test environment accommodations include changes in the setting, scheduling, timing, or response format (Duran, 2008). Linguistic accommodations can take a variety of forms such as bilingual dictionaries, content-specific glossaries, non-content glossaries, bilingual assessments, and translated versions of the assessments (Duran, 2008).

ELL accommodations are intended to make an assessment accessible to ELLs, not to give ELLs an advantage. If an accommodation gives an advantage over non-accommodated test-takers, then the accommodation may affect the validity of the assessment (Abedi & Gándara, 2006; Abedi et al., 2004). If an ELL accommodation is given to non-ELLs and there is no increase in non-ELL performance, then the accommodation does not affect validity (Abedi & Gándara, 2006).

Many commonly-used ELL accommodations have been found to be ineffective or to affect test validity (Abedi, 2009; Abedi & Gándara, 2006). Glossaries and longer testing time have both been found to increase native speaker performance, so their use as ELL accommodations questions the validity of the accommodated assessment (Abedi &

Gándara, 2006). Notwithstanding feasibility issues, bilingual and English dictionaries have also raised validity issues because they give an advantage to ELLs who can look up the definitions of content words (Abedi, 2009; Abedi & Gándara, 2006). Extra time is a common accommodation; however, research as to its effectiveness and its effect on validity is inconclusive (Abedi et al., 2004). Since the MCAS is an untimed assessment, the issues surrounding this accommodation of extended time are moot. Reading the test to test-takers is another accommodation, but Abedi et al. (2004) noted that the test administrator could unintentionally introduce non-verbal cues, which could affect the validity and reliability. Since Massachusetts does not allow this accommodation on the Biology MCAS based on a student's lack of English proficiency, the issues surrounding this accommodation are also moot.

Accommodations that address language barriers are more effective and valid for ELLs than accommodations designed for students with disabilities, such as small group testing and more time (Abedi & Gándara, 2006). An accommodation can be effective for some ELLs but not for others depending on student background and other factors (Abedi et al., 2004). Duncan et al. (2005) found that dual-language tests were effective for Grade 8 students in mathematics without affecting validity (as cited in Abedi & Gándara, 2006), but other research has shown that assessments translated into a first language may not be effective for ELLs, especially if content instruction was in English (Abedi et al., 2004). The Biology MCAS only has an English language version, and most ELLs in Massachusetts are in English-only immersion biology classrooms.

Abedi, Courtney, Mirocha, Leon, and Goldberg (2001) studied the use of English and bilingual dictionaries as accommodations for both ELL and non-ELL students on

science assessments (as cited in Abedi et al., 2004). The authors concluded that the dictionaries were not effective and had feasibility issues. They further questioned whether dictionary use gave accommodated students an advantage, thereby raising validity issues. ELLs are allowed use of MCAS-approved word-to-word dictionaries (MA DESE, 2012e). Such dictionaries do not raise validity issues because they only give the word, not the definition, in the first language. This accommodation, however, does not help a student who has not learned content-specific or general academic vocabulary in their first language or who cannot read in their first language.

Like Abedi et al.'s (2001) finding, this MCAS accommodation raises feasibility issues. Some languages have more than one word-to-word dictionary on the MCAS-approved dictionary list (MA DESE, 2012e). The approved word-to-word dictionaries differ in the number of entries both inter- and intra-languages. Which approved dictionary a student is given as an accommodation or whether he or she is given this accommodation at all can vary by district, by school, and even within the same test administration room. There was no way to determine which dictionary a student used or even if the accommodation was made available to the student; however, the current study assumed that ELLs had access to this accommodation on the Biology MCAS.

Another language accommodation is use of content or non-content glossaries that provide an ELL with a word in his or her first language. Some studies have shown that glossaries, plus extra time, are effective in increasing ELL performance; however, one study showed that these accommodations also increased the performance of non-ELLs (Abedi et al., 2004). In Massachusetts, ELLs can use approved word-to-word bilingual content glossaries on the MCAS, and since the MCAS is an untimed test, there is no issue

of providing the former accommodation without the latter (MA DESE, 2012e). Although there are MCAS-approved word-to-word content glossaries, these only exist for some of the first languages of Massachusetts ELLs.<sup>31</sup> The current study assumed that if there was a word-to-word glossary in an ELL's L1, the student had access to it for the June 2012 Biology MCAS.

Since academic language may impact ELL performance, reducing an assessment's linguistic complexity is a possible accommodation for ELLs. Several studies have shown that this is an effective accommodation without affecting validity (Abedi & Gándara, 2006). Abedi (2009) studied the effect of four ELL accommodations on a Grade 8 mathematics test devised from a combination of publically released items from the NAEP and the Third International Mathematics and Science Study. Two language accommodations were used: a customized English dictionary<sup>32</sup> and a pop-up glossary in a computerized test version. The results showed that Grade 8 ELLs with a language accommodation (n = 170) performed better than Grade 8 ELLs without an accommodations (n = 86); however, only the accommodation of a computer pop-up glossary reached significance. The results also indicated that the computer pop-up glossary was only significant for linguistically complex test items and that ELLs glossed nearly three times that of non-ELLs. Since neither the customized English dictionary nor the pop-up glossary increased the performance of non-ELLs, these accommodations had no effect on validity.

---

<sup>31</sup> At the high school level, there are MCAS-approved glossaries for biology in the following languages: Arabic, Bosnian, Chinese Simplified, Russian, and Spanish. In addition, there is also a Burmese glossary for high school science in general, not specifically for biology (MA DESE, 2012e).

<sup>32</sup> The customized English dictionary had all the content-related words removed; thus, the authors believed that this accommodation had no effect on validity.

Abedi and Hejri (2004) studied accommodations and ELL performance on the 1996 NAEP main assessment in science and math and the 1998 assessment in reading, writing, and civics at the Grade 4 and Grade 8 levels.<sup>33</sup> The study examined three aspects of accommodations and ELLs: (1) effect, (2) validity, and (3) whether linguistic complexity of the items interacted with any effect of the accommodations. The authors computed a “percent of overachievement” (POA) as a comparison index between ELL and non-ELL achievement where the magnitude of the POA suggested the magnitude of the achievement gap.<sup>34</sup> Their results indicated that the accommodations did not have a significant effect on ELL NAEP performance across the content areas in either Grade 4 or Grade 8. In general, it appeared that the accommodations did not affect the validity of the assessment; however, the authors suggested caution in interpreting this finding given the design limitations and small sample size for ELLs with accommodations on the NAEP (Grade 4 reading  $n = 41$ ; Grade 8 reading  $n = 30$ ).

Confirming earlier studies, Abedi and Hejri (2004) found an ELL achievement gap in all subjects at both grade levels. They also found that reduced linguistic complexity yielded a more valid assessment for ELLs, though there was variation across subjects and grade level. The ELL achievement gap was similar for both linguistically simple and linguistically complex items for Grade 4 science, Grade 8 math, and writing in both Grade 4 and Grade 8. The ELL achievement gap for Grade 4 math and Grade 8 science, however, widened when items became linguistically complex. For Grade 4 math, the POA or ELL achievement gap was 50.7% for items with linguistic simplicity, but this

---

<sup>33</sup> The authors also looked at Grade 12 NAEP data, but they did not report the findings and cited space limitations and more available research at the Grade 4 and Grade 8 levels.

<sup>34</sup> A large POA suggested a large gap between ELL and non-ELL achievement and vice versa.

nearly doubled to 96.1% when the items had more linguistic complexity. Similar results were found for Grade 8 science where the POA for linguistically simple items was 52.5% and 93.1% for the more linguistically complex items, a 77% increase.

Abedi and Hejri (2004) acknowledged several limitations that arose from the NAEP data. Their major concern was the reliability and validity of individual student performance, but other noted limitations were the small number in some subgroups and the randomness of ELL accommodations. The instances of some accommodation types were so small that accommodations, regardless of type, were treated as one variable. This could have masked an accommodation's differential effectiveness across ELL subpopulations. The authors also noted difficulty in knowing the language proficiency between accommodated ELLs and non-accommodated ELLs; language proficiency differences could have affected an accommodation's effectiveness and/or validity. Another limitation of Abedi and Hejri's study was that NAEP ELL selection criteria included both ELLs with more than three years of English instruction and ELLs with less than three years of English instruction (unless their school thought they could not participate). The ELL selection criteria could have hidden intra-group differences of how English proficiency levels impacted performance. The lack of a uniform definition of ELL or English proficiency levels among the states compounded the problem. Although Abedi and Hejri tried to compensate for limitations in the NAEP data, they believed it limited their ability to study the validity of NAEP accommodations.

After analyzing several studies, Abedi and Gandara (2006) concluded that (1) content assessments for ELLs should have reduced linguistic complexity and (2) reduced linguistic complexity did not alter the construct validity (also see Abedi et al., 2004). In

an earlier paper, Abedi et al. (2004) addressed the issue of whether linguistic modification was equivalent to dumbing down an assessment. They authors felt that the goal of academic language proficiency for all students and linguistic modification of content assessments were not mutually exclusive. The former can be developed in other areas of schooling, while the latter ensures that wide-scale standardized tests only assess the construct of content.

Although I believe this recommendation is sound for assessing ELL content knowledge, I believe it would present feasibility issues in terms of cost, as well as raise some equity issues. With respect to ELLs, a reduced linguistic complexity content exam would not assess their mastery of grade-appropriate discipline discourse and could potentially keep them in a linguistic ghetto that would hinder their post-secondary education. With respect to non-ELLs, this approach would also raise equity issues for struggling readers if they were not also given the opportunity to demonstrate content knowledge on assessments that presented items in below grade-level language. This in turn raises the issue of whether the instrument is designed to assess content knowledge or content knowledge in appropriate grade-level discourse (see Martiniello, 2008).

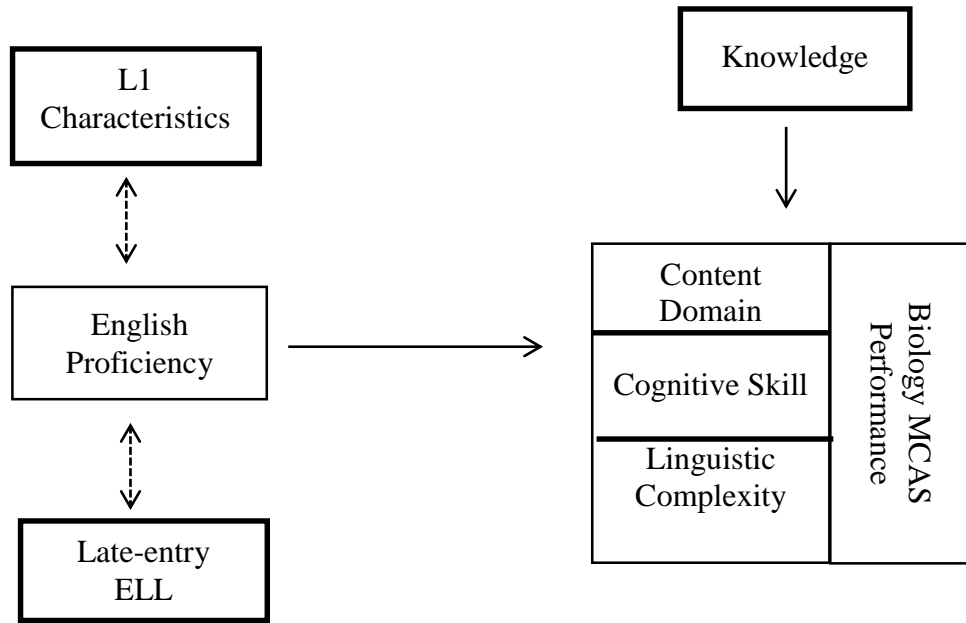
The current study explored the impact of item linguistic complexity on ELL performance on the Biology MCAS and whether the linguistic complexity findings of studies at lower grade levels and in different discipline areas hold at the secondary level in the content area of biology. It further explored whether the linguistic complexity findings hold across English proficiency levels and language groups, thereby addressing one of the limitations Abedi and Hejri (2004) encountered using NAEP data in studying accommodations.



## **Conceptual Framework**

Since language conveys meaning, content and language are not isolated realms. Content learning requires academic language (Boyson & Short, 2012; Schleppegrell, 2001; Snow, 2010; van Goor & Heyting, 2008). Tan's (2011) findings reinforced earlier research that it is not enough to teach content in the target language and hope that language learning is taking place, or to teach language and hope that content learning is happening (also see Kong & Hoare, 2011). In SEI classrooms, content knowledge is presented through modified language for comprehensible input. Biology MCAS items, however, have the lexical and syntactic linguistic complexity characteristic of academic language, and the absence of grade-level academic language impacts an ELL's ability to demonstrate content knowledge.

The current study's conceptual framework is that the learner characteristics of English proficiency level (linguistic knowledge), L1 language family, L1 orthography, and late-entry ELL status impact ELL Biology MCAS performance through their intermediary effect on academic language (Figure 2.3). This study analyzed ELL performance on the whole instrument and multiple-choice performance for each of the six content domains, three levels of item cognitive skill, and three levels of item linguistic complexity. It further analyzed the impact of English proficiency on these performance areas as well as the impact of two L1 characteristics: (1) Latinate or non-Latinate, and (2) alphabetic or non-alphabetic. In addition, disaggregation of late-entry ELLs explored performance for a subgroup that is considered at risk for acquiring academic language in four or fewer years.



*Figure 2.3. Language, Content, and the Biology MCAS. ELLs need both comprehensible input and academic language in learning content. Item linguistic complexity and cognitive skill level also impact achievement.*

## Summary

Krashen's five hypotheses of SLA illustrate the importance of comprehensible input in second language acquisition. In second language content instruction, comprehensible input is usually achieved through modified (i.e., simplified) language. Research indicates that a second language can be acquired at any age; however, there is some disagreement in the literature as to what constitutes acquisition. One view is that native-like proficiency can rarely be achieved if second language acquisition begins after the age of puberty. Another view is that except in the area of pronunciation, proficiency approaching native-speaker levels is possible; yet, several factors may impact or limit the ultimate proficiency attained. Irrespective of the ultimate proficiency possible, second language learners who begin after the age of 12 years face challenges not present for

younger second language learners. Despite the time constraint challenge, older second language learners have some advantages over younger learners. Older second language learners have fully acquired their first language and have metalinguistic awareness, which can positively transfer to the second language.

Academic language is the biggest challenge facing second language learners who start after 12 years old. Social language or BICS is usually acquired within three years. Academic language or CALP, on the other hand, requires between four and nine years for native-like proficiency. This creates a challenge to second language learners who enter U.S. public schools in high school. The research indicates that there is simply not enough time for them to become proficient in the English language register needed to access the same educational opportunities as their English-speaking peers.

Academic language differs from social language in vocabulary, syntax, and discourse conventions. Compared to social language, academic language is complex. Wide-scale assessments use academic language, and lack of academic language contributes to the ELL achievement gap. If an instrument uses language that obfuscates meaning for ELLs, then this raises questions of validity and reliability for its use to assess ELL content knowledge. Accommodations are one way to make an assessment accessible to ELLs. The literature, however, is inconclusive on the effectiveness of several accommodations and whether these accommodations affect validity. Linguistic modification appears to be an effective ELL accommodation without altering construct validity. This underscores the role of linguistic complexity in ELL performance on content assessments.

The current study analyzed ELL performance on the June 2012 Biology MCAS and the impact of English proficiency, first language characteristics, and late-entry ELL status. It further analyzed the impact of these factors on ELL performance for the multiple-choice item attributes of content domain, cognitive skill level, and linguistic complexity.

## CHAPTER 3

### METHODS

The previous two chapters established that English proficiency impacts ELL performance on standardized assessments, as does the linguistic complexity of the instrument, resulting in an achievement gap between ELLs and non-ELLs (Abedi, 2002, 2009; Abedi & Gándara, 2006; Abedi & Hejri, 2004; Abedi et al., 2004; Abedi & Lord, 2001; Martiniello, 2008; Menken, 2008, Chapter 4; Solano-Flores & Li, 2009; Solorzano, 2008). Previous research has identified a need for further study into the role of language in ELL content assessment (Abedi, 2008b; Abedi et al., 2004; Abedi & Lord, 2001; Martiniello, 2008; Solorzano, 2008) and the extent of impact of English proficiency (Solorzano, 2008) in order to inform the construction of reliable and valid content assessments for ELLs (Abedi, 2008b) and as a source of differential item functioning (Martiniello, 2008). This study analyzed Biology MCAS performance for ELLs at five levels of English proficiency. In addition, it also analyzed ELL Biology MCAS performance by first language characteristics and the impact of item linguistic complexity—needs identified by Abedi (2008b), Martinello (2008), and Solorzano (2008).

This study explored the impact of English language proficiency, Latinate first language, first language orthography, and late-entry ELL status not only on Biology

MCAS performance but also across its six biology domains and the three cognitive skill levels of its multiple-choice items. To my knowledge, no study has been done on:

(1) ELL performance on the Biology MCAS, (2) ELL performance across the six domains of the Biology MCAS, (3) ELL performance across the cognitive skill levels of the Biology MCAS, and (4) item linguistic complexity impact on ELL Biology MCAS performance.

This chapter first discusses the research design and research questions. A discussion of data sources and variables used in this study follows. This study consisted of two phases. Phase I operationalized multiple-choice item linguistic complexity and analyzed the multiple-choice item attributes of content domain, cognitive skill level, and linguistic complexity. Phase II analyzed ELL Biology MCAS performance. The Phase I data analysis section discusses the operationalization of linguistic complexity, data transformations, and descriptive and comparative analyses for the attributes of content domain, cognitive skill level, and linguistic complexity. The data analysis section for Phase II (ELL performance) follows. This section includes a discussion of data management, data preparation, data transformations, and the descriptive, comparative, and inferential statistical analyses for ELL Biology MCAS performance.

### **Research Design and Data Sources**

The nature of the study was a secondary data analysis utilizing statewide data to describe and analyze secondary ELL performance on the Biology MCAS. This study first analyzed ELL performance by total scores and performance levels on the June 2012 Biology MCAS. This study then further analyzed performance by the multiple-choice

item attributes of content domain,<sup>35</sup> cognitive skill,<sup>36</sup> and linguistic complexity. This allowed for exploration of ELL performance on: (1) six content domains, (2) three cognitive skills, and (3) three levels of linguistic complexity.<sup>37</sup>

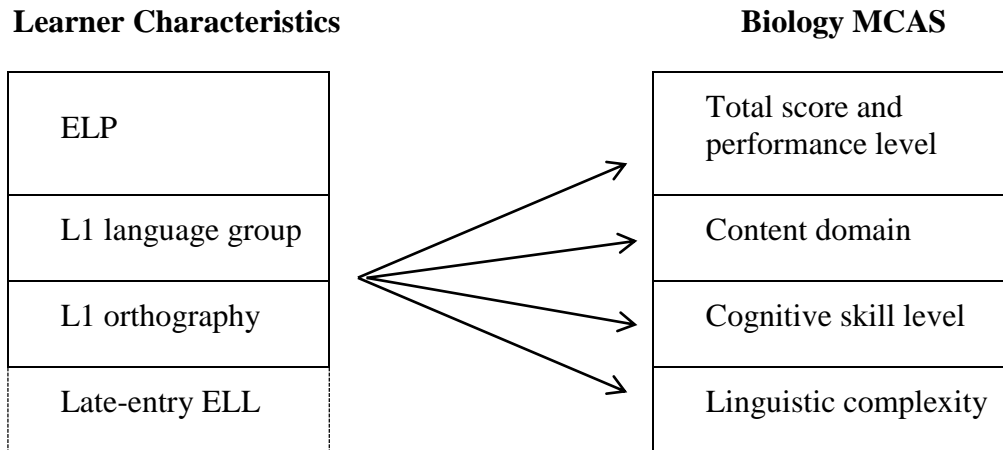
Each ELL student brings his or her own unique and dynamic interlanguage to the assessment (Ellis, 2003); however, this heterogeneity can mask the impact of language factors on Biology MCAS performance. This study explored the impact of three language factors on ELL performance: (1) English language proficiency (ELP), (2) first language family (Linate or non-Linate), and (3) first language orthography (alphabetic or non-alphabetic). It also explored the impact of late-entry ELL status (ELLs who enter the U.S. at 12 years of age or later) on performance. This study's conceptual framework posited that language factors (ELP, L1 language family, and L1 orthography) and age of entry impact ELL Biology MCAS performance (Figure 3.1).

---

<sup>35</sup> Domain refers to the content strand or standard. The six domains assessed on the Biology MCAS are: (1) anatomy and physiology, (2) biochemistry, (3) cell biology, (4) ecology, (5) evolution and biodiversity, and (6) genetics.

<sup>36</sup> The June 2012 Biology MCAS assessed content knowledge across foundational, conceptual, application, constructive, and quantitative cognitive skill levels (Appendix D).

<sup>37</sup> The multiple-choice items on the June 2012 Biology MCAS assessed knowledge at three cognitive skill levels: foundational, conceptual, and application.



*Figure 3.1.* Study Design for ELL Biology MCAS Performance. Learner characteristics that impact ELL Biology MCAS performance.

**Research questions.** This study was conducted in two phases. Phase I was a textual analysis of the June 2012 Biology MCAS that operationalized and analyzed linguistic complexity and the item attributes of domain and cognitive skill level. Phase II consisted of descriptive and comparative analyses of ELL performance on the June 2012 Biology MCAS. Specific research questions were:

1. How did ELLs perform on the total score and on the performance level of the Biology MCAS?
  - (a) To what extent did English language proficiency impact total score and performance level on the Biology MCAS?
  - (b) To what extent did the first language family (Latinate or non-Latinate) impact total score and performance level on the Biology MCAS?
  - (c) To what extent did the first language orthography (alphabetic or non-alphabetic) impact total score and performance level on the Biology MCAS?



- (d) To what extent did the late-entry ELL status impact total score and performance level on the Biology MCAS?
2. How did ELLs perform on the six content domains of the Biology MCAS?
- (a) To what extent did English language proficiency impact performance on each of the six content domains of the Biology MCAS?
  - (b) To what extent did the first language family (Linate or non-Linate) impact performance on each of the six content domains of the Biology MCAS?
  - (c) To what extent did the first language orthography (alphabetic or non-alphabetic) impact performance on each of the six content domains of the Biology MCAS?
  - (d) To what extent did the late-entry ELL impact performance on each of the six content domains of the Biology MCAS?
3. How did ELLs perform on the different cognitive skill levels of the Biology MCAS?
- (a) To what extent did English language proficiency impact performance on the different cognitive levels of the Biology MCAS?
  - (b) To what extent did the first language family (Linate or non-Linate) impact performance on the different cognitive levels of the Biology MCAS?
  - (c) To what extent did the first language orthography (alphabetic or non-alphabetic) impact performance on the different cognitive levels of the Biology MCAS?

- (d) To what extent did the late-entry ELL impact performance on the different cognitive levels of the Biology MCAS?
- 4. How did ELLs perform on the different levels (high, medium, low) of item linguistic complexity on the Biology MCAS?
  - (a) To what extent did English language proficiency impact performance on each of the three levels of item linguistic complexity of the Biology MCAS?
  - (b) To what extent did the first language family (Latinate or non-Latinate) impact performance on each of the three levels of item linguistic complexity of the Biology MCAS?
  - (c) To what extent did the first language orthography (alphabetic or non-alphabetic) impact performance on each of the three levels of item linguistic complexity of the Biology MCAS?
  - (d) To what extent did the late-entry ELL impact performance on each of the three levels of item linguistic complexity of the Biology MCAS?

**Data sources.** Phase I of this study operationalized linguistic complexity and determined multiple-choice item parameters. A textual analysis of the publically released June 2012 Biology MCAS yielded linguistic complexity elements for each multiple-choice item. Item content domain is reported with MCAS results, and item cognitive skill level was provided by Massachusetts Department of Elementary and Secondary Education (MA DESE).

For Phase II, an application was made to MA DESE for access to statewide data. Phase II utilized three sources of statewide data that the Commonwealth of Massachusetts

makes available to researchers: (1) June 2012 Biology MCAS data (MCAS data), (2) March 2012 MEPA data (MEPA data), and (3) student demographic information from the Student Information Management System (SIMS data). Each student in Massachusetts public schools is designated by a 10-digit, state-assigned identification (SASID) number that is a unique identifier. The SASID linked MCAS data, MEPA data, and student demographic data; however, the data were de-identified by MA DESE before making them available for this study.

***The Biology MCAS instrument.*** The Biology MCAS exam is a summative assessment for the Massachusetts biology frameworks. The instrument is intended for all students—including students with disabilities and English language learners (ELLs)—in Massachusetts secondary schools (MA DESE, 2011b).

***Reliability.*** In their psychometric testing of the 2006 Biology MCAS, Hambleton, Zhao, Smith, Lam, and Deng (2008) found that the instrument's reliability statistics were high for: (1) all items ( $\alpha = 0.91$ ), (2) multiple-choice items ( $\alpha = 0.88$ ), and (3) constructed-response items ( $\alpha = 0.81$ ).<sup>38</sup> The authors also found that the test items were somewhat difficult for students as evidenced by a mean score of approximately 50% correct ( $n = 55,673$ ;  $M = 29.3$ ;  $SD = 12.5$ ).<sup>39</sup> Using a criterion of .10, the authors found no significant differential item functioning for the following groups: males/females, Whites/Blacks, Whites/Hispanics, and Whites/Asians. Item difficulty and discrimination analyses confirmed that the Biology MCAS had excellent technical quality.

---

<sup>38</sup> In 2012, the Biology MCAS had a Cronbach's  $\alpha$  of 0.91 and SEM of 3.32 (MA DESE, 2013a).

<sup>39</sup> Possible score range is 0 to 60.

Using a random sample of approximately 5,000, Hambleton et al. (2008) used two approaches—structural equation modeling and Eigenvalue plots—to explore test dimensionality. The results suggested a strong first factor (biology competence as measured by the instrument), which met the unidimensionality requirement for item response theory (IRT) models.<sup>40</sup> The authors fit a three-parameter logistic IRT model for the multiple-choice items and the graded response model for constructed-response items.<sup>41</sup> They found an excellent IRT model fit, which confirmed the unidimensionality of the instrument as well as accurate predictions of test score distribution. These findings supported the appropriateness of using IRT item statistics in test development, test score equating, and reporting. The item parameter estimates confirmed that the test items were on the difficult side for students (b parameter  $M = 0.35$ ;  $SD = 0.70$ ) and that the discrimination levels were high (a parameter  $M = 0.98$ ;  $SD = 0.21$ ).

*Validity.* There appear to be no threats to internal content validity. Biology MCAS test items undergo a rigorous item development cycle, which includes determining whether each item tests content knowledge in the Massachusetts Biology Curriculum Frameworks (MA DESE, 2011b). The Biology MCAS has a pre-equating design, so there is no issue of content and statistical match between linking and operational items (Deng, Sukin, & Hambleton, 2009). In addition, Hambleton et al. (2008) found a correlation between 2006 Biology MCAS scores and prior performance on the Grade 8 science MCAS ( $r = 0.80$ ) and the Mathematics MCAS ( $r = 0.77$ );

---

<sup>40</sup> Percent variance on first factor was 31%. Eigenvalues were: factor 1 = 13.8, factor 2 = 1.7, factor 3 = 1.4, and factor 4 = 1.2.

<sup>41</sup> The 2011 MCAS and MCAS-Alt Technical Report states that a one-parameter logistical model was used for the high school STE MCAS exams.

however, the authors urged caution in interpreting the significance because 2006 was the first year of the Biology MCAS.

*The MEPA instrument.* In SY 2011-2012, ELLs in Massachusetts public schools took the spring administration of the MEPA, which was based on the Massachusetts English Language Arts Curriculum Framework and aligned to English Language Proficiency Benchmarks and Outcomes for English Language Learners.<sup>42</sup> The MEPA assessed four language skills: reading, writing, listening, and speaking. The MEPA was composed of two assessments: (1) a written assessment for reading and writing skills (MEPA-R/W) and (2) the Massachusetts English Proficient Assessment-Oral (MELA-O), which was an observational assessment for listening and speaking skills (MA DESE, 2013d). The MEPA level (1 to 5) was determined by a combination of the MEPA-R/W and MELA-O scores.

*MEPA-R/W.* The secondary level MEPA-R/W, administered in two sessions, contained both multiple-choice and constructed-response items (short answer and open response depending on the session level). The reading items addressed the skills relating to vocabulary and syntax in print, beginning to read in English, comprehension, literary elements and techniques, and informational/expository text; the writing items focused on the skills relating to prewriting, writing, and editing (MA DESE, 2013d). The MEPA-R/W had multiple forms and levels; depending on ELP or previous MEPA score, a student took either Forms 1 and 2 or Forms 2 and 3. Employing a common item-linking technique using IRT scaling, within year and across year equating maintained the measurement scale established in 2009 (MA DESE, 2013d).

---

<sup>42</sup> See [www.doe.mass.edu/ell/benchmark.pdf](http://www.doe.mass.edu/ell/benchmark.pdf).

*MELA-O.* The MELA-O score was based on informal classroom observations by a qualified MELA-O administrator, who completed the MA DESE-sponsored training and passed the qualifying test (MA DESE, 2013d). Using a scoring matrix, MELA-O administrators assessed a student's listening and speaking skills (fluency, vocabulary, pronunciation, and grammar) with scores ranging from 0 to 5 (MA DESE, 2013d). There were five possible points for listening and 20 possible points for speaking, and the total points were added to the MEPA-R/W score. For item calibration, the listening scores were treated as one item, and the speaking scores were treated as four items (MA DESE, 2013d).

*Classical test theory.* Classical statistical analyses, which included difficulty and discrimination indices, found the MEPA-R/W and MELA-O to be sound instruments (MA DESE, 2013d). The difficulty indices (P-value) for the spring 2012 MEPA-R/W were within acceptable ranges. For Forms 1 and 2, the P-value for all items ( $n = 40$ ) was  $M = 0.60$  and  $SD = 0.12$ , and for Forms 2 and 3, the P-value for all items ( $n = 39$ ) was  $M = 0.63$  and  $SD = 0.11$  (MA DESE, 2013d). Similarly, the discrimination indices for all items were within acceptable ranges. For Forms 1 and 2, all items ( $n = 40$ ) had discrimination  $M = 0.48$  and  $SD = 0.18$ ; for Forms 2 and 3, all items ( $n = 39$ ) had discrimination  $M = 0.41$  and  $SD = 0.20$  (MA DESE, 2013d).

*Item response theory.* Standardization DIF procedures (minimum  $n = 200$ ) evaluated differential item functioning for the following groups: males/females, Whites/Blacks, Whites/Hispanics, Whites/Asians, not low-income/low-income, and no disability/disability (MA DESE, 2013b). Where found, differential item functioning was classified as low ( $-0.10$  to  $-0.05$  and  $0.05$  to  $0.10$ ) or high ( $< -0.10$  and  $> 0.10$ ) DIF;

values between -0.05 and 0.05 were considered negligible. DIF analyses indicated that few multiple-choice and constructed-response items had either low or high differential item functioning.

*Reliability.* English proficiency was the primary dimension assessed by the MEPA; however, item content subcategories introduced potential threats to unidimensionality (MA DESE, 2013d). Two nonparametric IRT-based dimensionality analyses, DIMTEST and DETECT, were conducted; the former detected violations of local independence (i.e., presence of multidimensionality), and the latter measured multidimensionality effect size (MA DESE, 2013d). DIMTEST analyses for the Grade 9-12 MEPA R/W and MELA-O rejected the null hypothesis of unidimensionality at level 0.01, and DETECT was employed to measure effect-size of all three subtests—reading, writing, and listening/speaking (MA DESE, 2013d). Grade 9-12 Forms 1 and 2 and MELA-O exhibited a moderate (0.37) effect size, and Forms 2 and 3 and MELA-O exhibited a strong (0.46) effect size (MA DESE, 2013d). Further analyses indicated that MELA-O was the primary cause of multidimensionality because it measured a construct different from reading and writing. When DETECT analyses were conducted only on the MEPA-R/W subtests, the effect size for multidimensionality was very weak (0.19) for Grade 9-12 Forms 1 and 2 or weak (0.21) for Forms 2 and 3 (MA DESE, 2013d)

**Sample.** The study sample comprised secondary students designated as ELLs by the Commonwealth of Massachusetts in June 2012 and who took the June 2012 Biology MCAS and the 2012 MEPA. In addition to looking at the ELL sample as a whole, this study disaggregated subgroups by English proficiency, first language characteristics

(language family and orthography), and arrival in the United States at 12 years or older, defined herein as the late-entry ELL subgroup.

The ELL data in this study are based on ELP-level designations, including re-designation as English proficient, used uniformly throughout Massachusetts. This addressed validity and reliability concerns that arise from the differing ELL designations and proficiency levels across states (Abedi et al., 2004; Solorzano, 2008). This study also disaggregated the ELL sample by ELP level. This allowed not only the exploration of ELP impact on performance but also the exploration of performance and factor impact at MEPA Levels 3 and above, which addressed reliability concerns at the lower English proficiency levels (Abedi & Hejri, 2004).

**Variables.** The variables in this study fell into three general categories: (1) performance, (2) student demographics, and (3) item attributes. State-level datasets provided the values for performance variables and most of the student demographic information, as well as the item attributes of domain and cognitive skill level. The item linguistic complexity attribute resulted from a textual analysis of the June 2012 Biology MCAS instrument. Table 3.1 summarizes the variables for Biology MCAS performance and English proficiency (MEPA performance). Tables 3.4 and 3.5 summarize demographic variables, and Table 3.4 summarizes the item attribute variables.

***Performance variables.*** Performance variables included the raw score and performance level on the June 2012 Biology MCAS and the scaled score and performance level on the spring 2012 MEPA instruments. Table 3.1 summarizes the dependent variables for Biology MCAS and MEPA English proficiency performance.



***Biology MCAS performance.*** This study described and analyzed ELL performance on the June 2012 Biology MCAS (Appendix A). The instrument consisted of 40 dichotomously scored items (multiple-choice) and five polytomously scored items (constructed-response). The maximum raw score was 60 points: 1 point for each correct multiple-choice question (40), and 0 to 4 possible points for each of the five open response questions (20). All MCAS items are calibrated using the IRT one-parameter logistic model for multiple-choice items and the graded response model for constructed responses (MA DESE, 2013e). After calibration and the identification of item parameters,  $\theta$  (true score) is calculated for each student (MA DESE, 2011a). The  $\theta$  scores are transformed into a reporting scale for ease of understanding by all stakeholders (MA DESE, 2013e).<sup>43</sup> The scaled scores range from 200 to 280 points and are used to determine performance levels: Failing (200-218), Needs Improvement (220-238), Proficient (240-258), and Advanced (260-280).<sup>44</sup> Performance levels were recoded as follows: Failing = 1, Needs Improvement = 2, Proficient = 3, and Advanced = 4. Performance levels represent a range of skills at that level—i.e., not all students at the Proficient level can answer correctly all the items at this level (Schneider et al., 2013).

This study limited analysis of item attribute performance to the multiple-choice items because there is no subjectivity or ambiguity in dichotomous scoring. The instrument's multiple-choice items included 36 stand-alone items and four items that were part of a module. A module consists of a reading passage, which sets the context, followed by four multiple-choice items and one constructed-response item. This study did

---

<sup>43</sup> Because of the scaling process, means and standard deviations should be reported only on raw scores.

<sup>44</sup> Scaled scores are reported at even number intervals—i.e., 200, 202 ... 278, 280.

not differentiate between the 36 stand-alone multiple-choice items and the four multiple-choice items in the module because the latter could have been answered through content knowledge even with limited passage comprehension.

***English language proficiency (ELP).*** In March, 2012, Massachusetts used the MEPA-R/W and MELA-O instruments to assess ELP for ELLs. MEPA scores were reported on a scale of 400 to 550, which were scaled to the following ELP levels: Level 1 = 400-449, Level 2 = 450-463, Level 3 = 464-488, Level 4 = 489-499, and Level 5 = 500-550 (MA DESE, 2012f).<sup>45</sup>

---

<sup>45</sup> MEPA performance-level descriptors for Grades 9-12 are included in Appendix B.

Table 3.1  
*Performance Variable Summary*

<u>Variable name</u>	<u>Type</u>	<u>Values</u>	<u>Data source</u>
Biology MCAS performance			
MCAS raw score	Interval	0-60	MCAS
MCAS scaled score	Interval	200-280	MCAS
MCAS performance level	Ordinal	Failing, Needs Improvement, Proficient, Advanced	MCAS
English language proficiency (ELP)			
MEPA scaled score	Interval	400-550	MEPA
MEPA performance level	Ordinal	1-5	MEPA

***Demographic variables.*** Demographic variables fell into two categories: ELL and age-related. ELL demographics are summarized in Table 3.2, and these included the first language, ELL status, and programs. Age-related demographics were used to calculate age and age of arrival for the disaggregation of the late-entry ELL subpopulation. Age-related demographic variables are summarized in Table 3.3.

***First language characteristics.*** Reading involves processing cues, and language characteristics (e.g., inflections) have greater influence on the processing of informative cues than on less informative cues (Chitiri & Willows, 1994). “Underlying cognitive resources are tapped differentially, to the degree demanded by the orthographic or linguistic characteristics of L1 and L2” (Geva, 1999, as cited in Birch, 2002, p. 38). As discussed in Chapter 2, Schaap (2011) found that differential item functioning was not the same across first language groups. This study examined the impact of two first language characteristics on ELL Biology MCAS performance. The first characteristic

was whether the first language was a Latinate language, and the second characteristic was whether first language orthography (writing system) was alphabetic.

Table 3.2  
*ELL Variable Summary*

<u>Variable name</u>	<u>Type</u>	<u>Values</u>	<u>Data source</u>
First language	Categorical	Alphanumeric, 3-digit; 177 codes	SIMS
ELL status (limited English proficient)	Categorical	00 = non-ELL 01 = ELL	SIMS
LEP program status	Categorical	00 = no program 01 = SEI 02 = 2-way bilingual 03 = other bilingual 04 = LEP opted out of ELL programs	SIMS

To categorize a first language as Latinate or non-Latinate, the SIMS first language variable was recoded into Latinate = 1 and non-Latinate = 2. Latinate first languages were defined as the language family derived from Latin (Finegan, 2004), and they included Latin (SIMS 480), Italian (SIMS 005), French (SIMS 003), Spanish (007), Portuguese (006), Rumanian (SIMS 655), Romanisch (SIMS 660), Cape Verdean (SIMS 001),<sup>46</sup> Catalan (SIMS 205), Valencian (SIMS 820), and French Patois (SIMS 305); all other first languages were recoded as non-Latinate.

Birch (2002) categorized reading strategies as: (1) visual meaning-based, (2) partial alphabetic, and (3) fully alphabetic. There is evidence that reading logographic orthographies is more similar to processing images than to the reading process for an alphabetic language where visual processing is only for sight words (Birch, 2002);

<sup>46</sup> Cape Verdean is Portuguese-based.

however, Chitiri and Willows (1994) believed that English's opacity requires more visual processing for word recognition than more transparent alphabetic languages.<sup>47</sup> Since low-level processing strategies (e.g., word recognition) are dependent on orthography, L1 to L2 literacy transfer can be positive, negative, or absent (Birch, 2002; Ellis, 2003, Chapter 6). For example, novice Mandarin second language learners from alphabetic first languages have a harder time than learners from a non-alphabetic orthography (C. H. Lee & Kalyuga, 2011). The second characteristic of first language was whether the language's orthography is alphabetic or non-alphabetic.<sup>48</sup> Some alphabets are non-letter-based; however, for purposes of this study, alphabetic languages was defined as languages using the Roman, Cyrillic, or Greek alphabets because, like English, their orthographies are letter-based.<sup>49</sup> Linguistic texts were consulted to categorize the orthographies of non-Latinate first languages, and first languages were recoded as alphabetic = 1 and non-alphabetic = 2. First language characteristics are outlined in Appendix F.

***Age-related variables.*** As described more fully below, the date of birth was used to calculate the student's age at the time of the June 2012 Biology MCAS. Using age and years in Massachusetts schools, the age of arrival was calculated to disaggregate the late-entry ELL subgroup.

---

<sup>47</sup> Opacity refers to sound-letter correspondence. A transparent language has clear, unambiguous sound-letter correspondence.

<sup>48</sup> Non-alphabetic orthographies include syllabic and logographic writing systems; Chinese languages use sinograms, which are logograms with phonetic complements (Birch, 2002).

<sup>49</sup> An alphabet matches sounds to written symbols, such as letters; syllabic orthographies (such as Arabic and Hebrew) are consonantal scripts rather than full alphabets (Finegan, 2004)

Table 3.3  
*Age-Related Variable Summary*

<u>Variable name</u>	<u>Type</u>	<u>Values</u>	<u>Data source</u>
Date of birth	Numeric		MEPA
Age	Interval	13-22 years	Calculated
Years in MA schools	Interval	0-11	MEPA
Age of arrival	Interval		Calculated
Grade	Categorical	09 = Grade 9 10 = Grade 10 11 = Grade 11 12 = Grade 12	SIMS

***Item attribute variables.*** Test-takers respond to item features such as particular words and phrases, context, layout, and format in creating the mental representation that leads to relevant knowledge and skill retrieval from long-term memory (Leighton & Gokiert, 2005). The cognitive effects of test item features (such as vocabulary) represent an emerging area of interest in item generation (Leighton & Gokiert, 2005). This study examined performance on three MCAS item attributes: (1) content domain, (2) cognitive skill level, and (3) linguistic complexity. Domain was reported with the MCAS results, and the cognitive skill level was provided by MA DESE. Linguistic elements resulted from a textual analysis of the June 2012 Biology MCAS. Item attribute variables are summarized in Table 3.4. Descriptive analyses of frequencies and measures of central tendency were conducted for performance on multiple-choice items at the three cognitive skill levels (foundational, conceptual, and application) for the entire sample.

Table 3.4  
*Item Attribute Variable Summary*

<u>Variable name</u>	<u>Type</u>	<u>Values</u>	<u>Data source</u>
Stem lexical density (SLD)	Interval	9-102	Textual analysis
Total answer lexical density (TALD)	Interval	4-88	Textual analysis
Answer lexical density (ALD)	Interval	1-22	Calculated
Total lexical density (TLD)	Interval	21-141	Calculated
Stem syntax (SS)	Interval	1-10	Textual analysis
Stem syntactic density (SSD)	Interval	5.75-25	Calculated
Reading complexity score (RCS)	Interval	540L – 1520L	Lexile® Analyzer
Composite linguistic complexity (CLC)	Ordinal	5-25	Calculated
Cognitive skill level	Categorical	1-5	MA DESE
Domain	Categorical	BC, CB, EC, EV, GE, AP	MCAS instrument

***Linguistic complexity.*** Textual analysis of each multiple-choice item yielded four variables that operationalized linguistic complexity, and three more linguistic variables were calculated.<sup>50</sup> An eighth linguistic complexity variable, composite linguistic complexity (CLC), was calculated.<sup>51</sup> Descriptions of the linguistic complexity variables are summarized in Table 3.5.

<sup>50</sup> The operationalization of linguistic complexity is described more fully in the Phase I—Data Analysis section.

<sup>51</sup> The normalization of linguistic complexity values is described more fully in the Phase I—Data Transformation section.

Table 3.5

*Descriptions of Linguistic Complexity Variables*

<u>Variable name</u>	<u>Description</u>
Stem lexical density (SLD)	Number of word in the item's stem.
Total answer lexical density (TALD)	Total number of words in all four answer options
Answer lexical density (ALD)	TALD/4
Total lexical density (TLD)	SLD + TALD
Stem syntax (SS)	Number of sentences in the stem
Stem syntactic density (SSD)	Mean number of words per sentence in the stem. SLD/SS
Reading complexity score (RCS)	Reading level complexity
Composite linguistic complexity (CLC)	Composite of scaled values: $TLD_N + SSD_N + RCS_N^{52}$

**Domain.** “Individuals think and reason in relation to a content domain” (Barnett & Ceci, 2005, as cited in Leighton, Gokierto, & Cui, 2007, p. 143). “In conceptual domains, there are many interacting knowledge structures that must be processed simultaneously in working memory in order to be understood” (Van Merriënboer & Sweller, 2005, p. 156). The June 2012 Biology MCAS assessed content knowledge in six biology domains: (1) anatomy and physiology, (2) biochemistry, (3) cell biology, (4) ecology, (5) evolution and biodiversity, and (6) genetics. These domains align with the Massachusetts curriculum standards for high school biology, which are included in Appendix C (MA DESE, 2006). State data and the publically released instrument report the content domain of each test item as AP (anatomy and physiology), BC/CB (biochemistry and cell biology), EC (ecology), EV (evolution), or GE (genetics);

<sup>52</sup> The subscript N denotes the normalized values. These are discussed in the Phase I—Data Transformations section.



however, the MCAS performance data report biochemistry (BC) and cell biology (CB) as two separate content domains.

***Cognitive skill level.*** Leighton, Gokierto, and Ying (2007) believe that cognitive skills are more indicative of mastery than content because the former manipulates the latter; however, even though cognitive skills require content knowledge, the reverse is not always true – i.e., it is possible to answer an item correctly with some content knowledge, but no reasoning skill. During the Biology MCAS test development process, potential items are labeled with the cognitive skill level. These cognitive skills levels do not refer to the difficulty of an item, but rather to the “complexity of the mental processing a student must use to answer an item correctly” (MA DESE, 2011b). Prior to 2013 test item development, the cognitive skill levels assigned to test items were: foundational, conceptual, application, constructive, or quantitative (Appendix D); however, the cognitive skill level of the item was not reported with test data results. The item cognitive skill levels were obtained from the MA DESE.

### **Analysis Strategy—Phase I**

Phase I consisted of a textual analysis of the 40 multiple-choice questions on the June 2012 Biology MCAS instrument. Each item was assigned a value for each linguistic complexity element; these values, along with the item domain and cognitive skill level, created an item attribute dataset in SPSS.

**Content domain.** “[S]ubtle nuances in content area skill [drive] item difficulty” (Schneider, et al., 2013, p. 112). The content domain was reported with MCAS data file from MA DESE. Each item’s content domain was entered in the item attribute dataset and recoded as: AP = 1, BC = 2, CB = 3, EC = 4, EV = 5, and GE = 6.

**Cognitive skill level.** The cognitive skill level for each multiple-choice item was from MA DESE. The cognitive skill level was entered in the item attribute dataset and recoded as: foundational = 1, conceptual = 2, and application = 3.

**Operationalization of item linguistic complexity.** Second language acquisition research treats linguistic complexity as either an independent variable that influences performance or a dependent variable that describes performance (Kuiken & Vedder, 2012). With respect to influencing performance, studies have shown that item developers and test-takers do not always interpret an item's meaning in the same way (Leighton & Gokiert, 2005); linguistic complexity impacts comprehension, which in turn impacts performance. Studies have also shown that cognitive load impacts the linguistic complexity of oral output (Skehan & Foster, 1999). Over the past 20 years, linguistic complexity has been quantified in various ways, including frequencies, ratios, and formulae (Housen & Kuiken, 2009). "[A] review of the literature shows that there is no consistency in terms of how complexity is defined, operationalized and measured in L2 research, which at least partly explains the inconsistency of complexity findings both across and within studies" (Kuiken & Vedder, 2012, p. 277).

This study explored ELL performance across various levels of linguistic complexity of input (test items). Phase I included the operationalization and analysis of five linguistic complexity elements of the instrument's multiple-choice items: (1) stem lexical density (SLD);<sup>53</sup> (2) answer lexical density (ALD); (3) total lexical density (TLD); (4) stem syntactic density (SSD); and (5) reading complexity score (RCS) (see Table

---

<sup>53</sup> A test item consists of a lead, a stem, and answer options. The lead is contextual information that precedes the stem, which is the question; not all Biology MCAS items have both a lead and a stem. For purposes of this study, stem will include both the lead and the stem—i.e., the stem will refer to the text that precedes the answer options.

3.5). The SLD, ALD, and RCS values resulted from the textual analysis and were input into the SPSS item attribute dataset. The TLD and SSD variables were calculated in SPSS using the textual analysis data.

***Stem lexical density.*** Words can increase or decrease uncertainty (cognitive load) for comprehension, and more words can increase the cognitive load, especially for second language learners (Hale, 2003; Sweller & Chandler, 1991). Redundant elements, such as information (words) not intrinsic to the task, increase cognitive load because redundant information must be processed to make the determination that it is extraneous to the task (Sweller & Chandler, 1991). “[E]lement interactivity can be determined only by counting the number of interacting elements that people deal with a *particular* level of expertise” (Van Merriënboer & Sweller, 2005, p. 150). The words in an item’s stem were counted, including articles (e.g., *a*), demonstratives (e.g., *this*), and conjunctions (e.g., *and*). Numbers were counted as words (e.g., 1%, 100, etc.), as were images, labels, and table headings. This yielded an SLD score that was input into the item attribute dataset. For example, each of the four images and labels in the stem of question 15 counted as a word, yielding an SLD of 35 (Figure 3.2).

***Stem syntax (SS).*** “[I]n interpreting sentences, speakers utilize different cues in accordance with the syntactic characteristics of their respective languages” (MacWhinney et al., 1984, as cited in Chitiri & Willows, 1994, p. 314). Hale (2003) proposed incremental sentence processing with the maximum disambiguation at each word. It follows that the more words in a sentence, the higher the cognitive load for ambiguity reduction and comprehension. This study limited syntactic elements to mean sentence

length. A textual analysis determined the number of sentences in each item stem, and this value was input into the item attribute dataset.

***Total answer lexical density.*** Each multiple-choice question on the Biology MCAS has four answer options. The cumulative number of words (including articles, demonstratives, and conjunctions) in all four answer options was calculated and input into the item attribute dataset. Some MCAS multiple-choice answer options were diagrams or labels that referred to diagrams or images. In these instances, an image or a label counted as one word. If the answer option was a figure with words or labels differentiating the four answer options, then the words in the figures counted as words. For example, in question 1, the answer options were diagrams depicting the relationships among cellular respiration and photosynthesis, and carbon dioxide and oxygen (Figure 3.2). Since students needed to understand four components in the diagrams—cellular respiration, photosynthesis, CO<sub>2</sub>, and O<sub>2</sub>—each diagram counted as four words.<sup>54</sup> In question 15, the stem included four images of unicellular organisms, and the answer options were two image labels; each label counted as one word to yield three words for each answer option (Figure 3.2).<sup>55</sup>

***Reading complexity score.*** Assigning a reading complexity score can be difficult because of subjectivity. In their investigation of achievement level descriptors, Schneider, Huff, Egan, Gaines, and Ferrara (2013) found that coders had the lowest perfect agreement rates for reading load. This study used the Lexile Analyzer<sup>®</sup> (<https://www.lexile.com/>), a reading measurement system developed by MetaMetrics, to

---

<sup>54</sup> Since a lexeme is a unit of meaning, these four elements effectively function as diagram lexemes.

<sup>55</sup> Each of these four images was counted as a word in measuring stem lexical density.

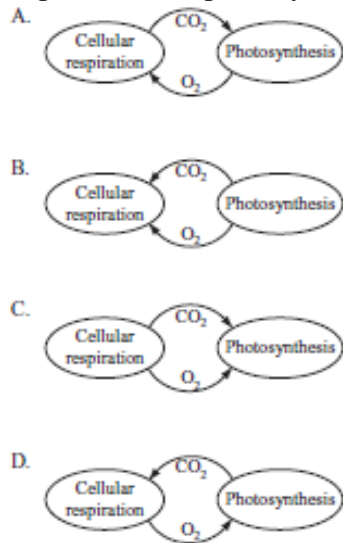
calculate the RCS score. The Lexile Analyzer<sup>®</sup> analyzes word frequency and sentence length to predict text difficulty and assigns an estimated Lexile<sup>®</sup> text measurement.<sup>56</sup> Estimated Lexile<sup>®</sup> text measurement scores for grade equivalents are reported for the interquartile range (IQR), which is the middle 50% of readers between the first and third quartiles. Estimated Lexile<sup>®</sup> text measurements range from 190L (1<sup>st</sup> quartile) to 530L (3<sup>rd</sup> quartile) for Grade 1 through 1135L (1<sup>st</sup> quartile) to 1385L (3<sup>rd</sup> quartile) for Grade 12. Reading complexity was limited to the item stem because the Lexile Analyzer<sup>®</sup> only measures conventionally punctuated, complete sentences; answer options were not complete sentences, and some contained diagrams, images, and symbols.

---

<sup>56</sup> A Lexile<sup>®</sup> text measurement is a number followed by *L* (e.g., 880L).

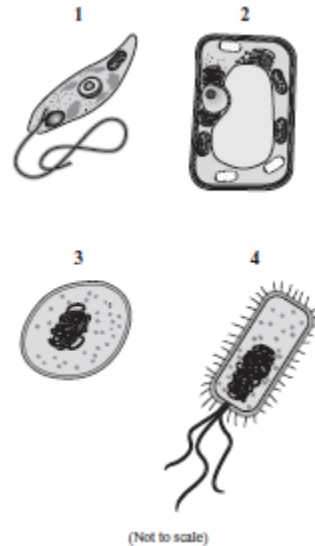
### 2012 Biology MCAS, Question 1

Which of the following diagrams accurately represents the use of gases in both cellular respiration and photosynthesis?



### 2012 Biology MCAS, Question 15

Each of the illustrations below shows either a prokaryotic cell or a eukaryotic cell. Each cell is numbered



Which two cells should be classified as prokaryotic cells?

- A. 1 and 2
- B. 1 and 3
- C. 2 and 4
- D. 3 and 4

*Figure 3.2.* 2012 Biology MCAS Questions with Diagram and Label Answer Options. Not all Biology MCAS answer options are text.

**Data preparation.** The June 2012 Biology MCAS instrument is publically available as a portable document format (PDF) file on the MA DESE website. The first step in preparing the data was to convert the PDF file into a Microsoft (MS) Word file

using Adobe® CreatePDF.<sup>57</sup> After the MS Word conversion, each item stem was saved as a separate MS Word file in US-ASCII plain text format in preparation for the Lexile Analyzer®.

***Lexile Analyzer®.*** Each plain text item file was uploaded to the Lexile Analyzer® for an estimated Lexile® reading score. The Lexile Analyzer® results could not be saved on the website or in a format that can be directly exported to SPSS. A screen capture of each item's Lexile® analysis was pasted into an MS Word file and labeled with the item number to maintain data records (see Appendix E for an example). The estimated Lexile® score for each multiple-choice item was entered in the item attribute dataset.

#### **Data transformations and calculated variables.**

***Calculated variables.*** The SPSS compute function calculated three additional linguistic variables: (1) total lexical density (TLD), (2) answer lexical density (ALD), and (3) stem syntactic density (SSD) (see Table 3.5).

***Total lexical density.*** TLD was computed in SPSS by adding the stem lexical density (SLD) and the total answer lexical density (TALD) values.

***Answer lexical density.*** Each multiple-choice question on the Biology MCAS has four answer options. An ALD variable was computed in SPSS by dividing the TALD (the cumulative number of words in all four answer options) by four.

***Stem syntactic density.*** SSD was defined as the mean number of words per sentence in the item stem. For each item stem, the SSD variable was computed in SPSS by dividing the SLD by stem syntax (SS; the number of sentences) to yield an SSD measure.

---

<sup>57</sup> <https://www.acrobat.com/createpdf/en/home.html>

**Composite linguistic complexity computation.** Three variables, TLD, SSD, and RCS, were used to calculate a composite linguistic complexity (CLC) value. Because these variables had scales of different magnitude, the first step was to normalize the values using the following equations:

$$\text{Value}_N = (\text{value} - \text{low}) / (\text{high} - \text{low})^{58}$$

$$\text{TLD}_N = (\text{TLD} - 21) / (141 - 21)$$

$$\text{SSD}_N = (\text{SSD} - 5.73) / (24 - 5.73)$$

$$\text{RCS}_N = (\text{RCS} - 540) / (1520 - 540)$$

SPSS computed a CLC variable for each multiple-choice item using the normalized values:  $\text{TLD}_N + \text{SSD}_N + \text{RCS}_N$ . The 40 multiple-choice items had CLC values that ranged from 0.06 to 2.36. The CLC values were scaled into low CLC, medium CLC, and high CLC based on tertiles: (1) low CLC = 0 to 0.7505, (2) medium CLC = 0.7506 to 1.2495, and (3) high CLC = 1.2496 to 5 (see Appendix H).

**Statistical analyses.** Using IBM SPSS software, descriptive and comparative analyses were conducted on the item attributes of content domain, cognitive skill level, and linguistic elements. Domain and cognitive skill level were categorical data, and frequency analyses were limited to percentages. Frequency analyses yielded measures of central tendency for the linguistic complexity values (continuous data). One-way ANOVA analyses and Scheffé post hoc analyses determined if there was a statistically significant difference in linguistic elements for multiple-choice items by content domain and cognitive skill level.

---

<sup>58</sup> The subscript N denotes normalized values from 0 to 1.



## **Analysis Strategy—Phase II**

Using IBM SPSS software, Phase II described and analyzed ELL performance on the June 2012 Biology MCAS by total score, performance level, and the item attributes of domain, cognitive skill level, and linguistic complexity. It further explored the impact of English proficiency levels, L1 language group, and L1 orthography on performance; disaggregation of the late-entry ELL population explored differential performance for this subgroup.

**Data management.** All data were stored on my password-protected computer, with daily electronic backups to a secure online backup service (Carbonite). In addition, all hardcopies of data and data analyses were stored in a locked cabinet at my home residence. Every effort was made to preserve the confidentiality of the data.

**Data file preparation.** The MA DESE provided an SPSS-compatible dataset for ELLs that included variables from three statewide datasets: (1) June 2012 Biology MCAS data, (2) spring 2012 MEPA data, and (3) SIMS data. The data request asked for de-identified records that were common across all three datasets. The MA DESE provided 15,295 records; these were records in which appeared a MEPA score or an ELL designation. This study used 3,315 records after data refinement and deletion of incomplete records.

***Refining the data.*** Preparation of the June 2012 Biology performance dataset began with refining the MA DESE data to meet the criteria of ELLs who took the June 2012 Biology MCAS and the spring 2012 MEPA. The first extraction was of records with a *scitry* value of 1, which indicated that the student took a Biology MCAS in

2012.<sup>59</sup> This yielded 4,914 records. The next extraction was of records with a MEPA performance level value, which yielded 4,339 records. These records represented ELLs who took a Biology MCAS in 2012 (February and June) and who had a 2012 MEPA score. The next step was to extract records with data in the *srawsc* (raw score) field, which was data for the June 2012 Biology MCAS items.<sup>60</sup> This yielded 3,345 records, representing ELLs who took the June 2012 Biology MCAS and who had a 2012 MEPA score.<sup>61</sup>

**Missing variables.** Of the 3,345 records, 3,125 had a value for *sscaleds* (scaled score); 219 records had a missing scaled score. These 219 records were extracted into a new dataset for further analysis. Seventeen records had incomplete item data: five records had no data for Session 1; 11 records had no data for Session 2; and one record only had data for the first nine items in Session 1 and no item data for Session 2.<sup>62</sup> These records were deleted and yielded 202 records with a raw score but no scaled score for the June 2012 Biology MCAS. A scaled score was calculated for these 202 records (see Data Transformation), and these records were merged back into the June 2012 Biology MCAS dataset to yield 3,327 records.<sup>63</sup> Of these 3,327 records, an additional 12 were deleted

---

<sup>59</sup> The *scitry* field had either a blank or an integer from 1 to 4. Of the records provided by MA DESE, 11,713 had a MEPA performance level, and of those, 7,380 had a value in the *scitry* field.

<sup>60</sup> The extraction was based on *srawsc* and not on *sscaleds* (scaled score) because first year ELL raw scores are not scaled.

<sup>61</sup> One record with a *srawsc* value of 0 was not included in the extracted dataset. Deletion of this record was similar to the methods used by Hambleton, Zhao, Smith, Lam, and Deng (2008) in their psychometric analyses of the four STE MCAS instruments.

<sup>62</sup> The Biology MCAS is given in two sessions on two separate days.

<sup>63</sup> Both datasets were sorted by ascending unique identifier before merging to ensure proper linkage between records.

because the SIMS first language field (FLANG) and limited English proficient (LEP) field were blank.<sup>64</sup> This resulted in the 3,315 records which were used in this study.

**Data transformations and calculated variables.** Phase II required handling missing MCAS scaled scores, transforming item scores, calculating percent correct for the item attributes, and calculating age of entry.

**Missing scaled MCAS scores.** Raw MCAS scores are scaled from 200 to 280 (in even number increments), and the scaled MCAS score determined the performance level (Fail, Needs Improvement, Proficient, or Advanced). The raw score is not scaled for ELLs who have been in Massachusetts public schools for less than one year; the performance level is reported as LEP. There were 219 records that had a raw score (total points) but not a scaled score. These records were extracted into a separate dataset, and 17 records were deleted because of incomplete item data. For the remaining 202 records, the *Transform, Recode into Same Variables* function recoded the scaled score to equal the raw score. Next, the *Transform, Recode into Same Variables* function recoded the scaled score (now equal to the raw score) using the spring 2012 MCAS Raw-to-Scaled Score Conversion on the MA DESE website (MA DESE, 2012h).<sup>65</sup> These 202 records were merged back into the June 2012 Biology MCAS dataset for further data refining.

**Item correct.** The MCAS data report a correctly answered item as “+” and incorrectly answered items as the letter option (i.e., A, B, C, or D) that the student chose.<sup>66</sup> In the June 2012 Biology MCAS dataset, the *Transform, Recode into Different*

---

<sup>64</sup> Seven of the records with a blank FLANG field from SIMS were coded in the MEPA data fields with English (00) as the first language.

<sup>65</sup> Appendix I contains the raw-to-scaled conversion chart for the June 2012 Biology MCAS.

<sup>66</sup> In addition, a blank space indicates that the item was unanswered and an “\*” indicates that more than one answer was selected and no points were given.

*Variables* function recoded item answers into a new item variable as follows: “+” = 1, all else = 0, and missing values = 0.

***Calculated item attribute performance.*** The recoded item correct variables were used to calculate the total number of points for multiple-choice items by: (1) each of the six content domains, (2) each of the three cognitive skill levels, and (3) each of the three linguistic complexity levels. The total number of points earned on an item attribute was divided by the number of items to yield a percent correct value. For example, the calculation for percent correct on genetics items where GE\_Performance is the total number of points (items correct) on the eight genetics multiple-choice items was:<sup>67</sup>

$$\% \text{ correct Genetics} = \text{GE\_Performance} / 8$$

This procedure was repeated for each content domain, each cognitive skill level, and each linguistic complexity level.

***Age of entry variable.*** Two variables were calculated and added to the dataset. The first variable was the student’s age at the time he or she took the June 2012 Biology MCAS. The first step was to recode the student’s date of birth (DOB in SIMS) from a string variable to a date variable (RC\_DOB). The next step was to recode the *adminyear* variable (2012 in MCAS data) to a string variable for June 4, 2012 (06042012; RC\_adminyear), which was then converted to a date variable (Date\_of\_MCAS).<sup>68</sup> The student’s age in June 2012 was calculated using: Age = Date of MCAS - Date of birth.<sup>69</sup> The second computed variable was the student’s age of entry (Age\_of\_Entry). The number of years in Massachusetts schools (yrsinmass\_num; MEPA variable) served as a

---

<sup>67</sup> Genetics items were items 7, 16, 17, 19, 20, 27, 37, and 41.

<sup>68</sup> The June 2012 Biology MCAS was administered on June 4 and 5.

<sup>69</sup> Age = Date\_of\_MCAS - RC\_DOB, which was truncated to an integer.

proxy for years in the United States to compute age of entry: Age of entry = Age - Years in MA schools.<sup>70</sup>

**Statistical analyses.** Phase II analyzed ELL performance on the June 2012 Biology MCAS and whether English proficiency, L1 family, L1 orthography, and late-ELL status had a statistically significant impact on total Biology MCAS score, performance level, and performance on the six content domains, the three cognitive skill levels, and the three levels of item linguistic complexity. Statistical analyses included descriptive analyses that in turn included frequencies, measures of central tendency, and graphs. “The purpose of comparative studies is to investigate the relationship of one variable to another by examining whether the value of the [variable] in one group is the same as or different from the [variable] of other groups” (McMillan, 2012, p. 179). Independent samples t-tests and one-way ANOVA analyses determined the impact of English proficiency, L1 family, L1 orthography, and late-entry ELL status on performance; Scheffé post hoc analyses were conducted as needed. Cohen’s d and univariate analyses of variance calculated effect size, and, where appropriate, simple linear regression analyses modeled factor impact on performance.

***Biology MCAS performance.*** Descriptive analyses of frequencies and measures of central tendency were conducted for total MCAS score and MCAS performance levels for the entire sample.

***Impact of English proficiency.*** To explore the impact of English proficiency on total MCAS score, the sample was disaggregated by English proficiency level (MEPA

---

<sup>70</sup> Age\_of\_Entry = Age - yrsinmass\_num. The years in MA variable was taken from the MEPA fields because it was numeric. The SIMS fields also had a yrsinmass field; however, it was a string variable and contained 5+ as a value.

level) for descriptive analyses, one-way ANOVA analyses, Scheffé post hoc analyses, and a simple linear regression analysis. The simple linear regression analysis was repeated for the subgroup at MEPA Levels 3 and above.

*Impact of L1 family.* The impact of first language family was explored by disaggregating the sample into two subgroups: ELLs with a Latinate L1 and ELLs with a non-Latinate L1. Frequencies and measures of central tendency described performance for the two subgroups. Independent samples t-tests determined whether there was a statistically significant difference in performance. Cohen's d was calculated to determine the effect size, if any, and a univariate analysis of variance yielded a partial eta-squared value to determine what percentage of the variance between groups was attributable to L1 family. To explore the impact of L1 family further, these analyses were repeated for the Latinate L1/non-Latinate L1 groups disaggregated into two English proficiency subgroups: (1) MEPA Levels 1 to 2 and (2) MEPA Levels 3 to 5.

*Impact of L1 orthography.* The impact of first language orthography was explored by disaggregating the sample into two subgroups: ELLs with an alphabetic L1 and ELLs with a non-alphabetic L1. Frequencies and measures of central tendency described performance for the two subgroups. Independent samples t-tests determined whether there was a statistically significant difference in performance. Cohen's d was calculated to determine the effect size, if any, and a univariate analysis of variance yielded a partial eta-squared value to determine what percentage of the variance between groups was attributable to L1 orthography. To explore the impact of L1 orthography further, these analyses were repeated for the alphabetic L1/non-alphabetic L1 groups disaggregated into two English proficiency subgroups: (1) MEPA Levels 1 to 2 and (2) MEPA Levels 3 to 5.

*Impact of late-entry ELL status.* The impact of late-entry ELL status (entry at the age of 12 years or later) was explored by disaggregating the sample into two subgroups: late-entry ELLs and not-late-entry ELLs. Frequencies and measures of central tendency described performance for the two subgroups. Independent samples t-tests determined whether there was a statistically significant difference in performance. Cohen's d was calculated to determine the effect size, if any, and a univariate analysis of variance yielded a partial eta-squared value to determine what percentage of the variance between groups was attributable to late-entry ELL status. To explore the impact of late-entry ELL status further, these analyses were repeated for the late-entry/not-late-entry groups disaggregated into two English proficiency subgroups: (1) MEPA Levels 1 to 2 and (2) MEPA Levels 3 to 5.

*Content domain performance.* Descriptive analyses of frequencies and measures of central tendency were conducted for performance on the multiple-choice items across the six content domains for the entire sample. For comparison, a statewide average percent correct was calculated for each domain by using the 2012 item analyses publically reported with district profiles (MA DESE, 2012c).

*Impact of English proficiency.* To explore the impact of English proficiency on content domain performance, the sample was disaggregated by English proficiency level (MEPA level) for descriptive analyses, one-way ANOVA analyses, and Scheffé post hoc analyses. A univariate analysis of variance yielded a partial eta-squared that determined what percentage of the variance among groups was attributable to English proficiency.

*Impact of L1 family.* The impact of first language family was explored by disaggregating the sample into two subgroups: ELLs with a Latinate L1 and ELLs with a

non-Latinate L1. Frequencies and measures of central tendency described performance for the two subgroups. Independent samples t-tests determined whether there was a statistically significant difference in performance. Cohen's d was calculated to determine the effect size, if any, and a univariate analysis of variance yielded a partial eta-squared value to determine what percentage of the variance between groups was attributable to L1 family.

*Impact of L1 orthography.* The impact of first language orthography was explored by disaggregating the sample into two subgroups: ELLs with an alphabetic L1 and ELLs with a non-alphabetic L1. Frequencies and measures of central tendency described performance for the two subgroups. Independent samples t-tests determined whether there was a statistically significant difference in performance. Cohen's d was calculated to determine the effect size, if any, and a univariate analysis of variance yielded a partial eta-squared value to determine what percentage of the variance between groups was attributable to L1 orthography.

*Impact of late-entry ELL status.* The impact of late-entry ELL status (entry at the age of 12 years or later) was explored by disaggregating the sample into two subgroups: late-entry ELLs and not-late-entry ELLs. Frequencies and measures of central tendency described performance for the two subgroups. Independent samples t-tests determined whether there was a statistically significant difference in performance. Cohen's d was calculated to determine the effect size, if any, and a univariate analysis of variance yielded a partial eta-squared value to determine what percentage of the variance between groups was attributable to late-entry ELL status. To explore the impact of late-entry ELL



status further, these analyses were repeated for the late-entry/not-late-entry groups at MEPA Levels 3 to 5.

*Cognitive skill level performance.* Descriptive analyses of frequencies and measures of central tendency were conducted for performance on multiple-choice items at the three cognitive skill levels (foundational, conceptual, and application) for the entire sample.

*Impact of English proficiency.* To explore the impact of English proficiency on cognitive skill level performance, the sample was disaggregated by English proficiency level (MEPA level) for descriptive analyses, one-way ANOVA analyses, and Scheffé post hoc analyses. A univariate analysis of variance yielded a partial eta-squared that determined what percentage of the variance among groups was attributable to English proficiency.

*Impact of L1 family.* The impact of first language family was explored by disaggregating the sample into two subgroups: ELLs with a Latinate L1 and ELLs with a non-Latinate L1. Frequencies and measures of central tendency described performance for the two subgroups. Independent samples t-tests determined whether there was a statistically significant difference in performance. Cohen's d was calculated to determine the effect size, if any, and a univariate analysis of variance yielded a partial eta-squared value to determine what percentage of the variance between groups was attributable to L1 family.

*Impact of L1 orthography.* The impact of first language orthography was explored by disaggregating the sample into two subgroups: ELLs with an alphabetic L1 and ELLs with a non-alphabetic L1. Frequencies and measures of central tendency described

performance for the two subgroups. Independent samples t-tests determined whether there was a statistically significant difference in performance. Cohen's d was calculated to determine the effect size, if any, and a univariate analysis of variance yielded a partial eta-squared value to determine what percentage of the variance between groups was attributable to L1 orthography.

*Impact of late-entry ELL status.* The impact of late-entry ELL status (entry at the age of 12 years or later) was explored by disaggregating the sample into two subgroups: late-entry ELLs and not-late-entry ELLs. Frequencies and measures of central tendency described performance for the two subgroups. Independent samples t-tests determined whether there was a statistically significant difference in performance. Cohen's d was calculated to determine the effect size, if any, and a univariate analysis of variance yielded a partial eta-squared value to determine what percentage of the variance between groups was attributable to late-entry ELL status.

*Linguistic complexity performance.* Descriptive analyses of frequencies and measures of central tendency were conducted for performance on multiple-choice items at three levels of item linguistic complexity (low, medium, high) for the entire sample.

*Impact of English proficiency.* To explore the impact of English proficiency on item linguistic complexity performance, the sample was disaggregated by English proficiency level (MEPA level) for descriptive analyses, one-way ANOVA analyses, and Scheffé post hoc analyses. Simple linear regression analyses modeled English proficiency impact on performance for items with: (1) low linguistic complexity, (2) medium linguistic complexity, and (3) high linguistic complexity. To further explore the impact of English proficiency on performance at three levels of item linguistic complexity, the

simple regression analyses were repeated for the subgroup of ELLs who had a MEPA score of 464 and above (MEPA Levels 3 to 5).

*Impact of L1 family.* The impact of first language family was explored by disaggregating the sample into two subgroups: ELLs with a Latinate L1 and ELLs with a non-Latinate L1. Frequencies and measures of central tendency described performance for the two subgroups. Independent samples t-tests determined whether there was a statistically significant difference in performance. Cohen's d was calculated to determine the effect size, if any, and a univariate analysis of variance yielded a partial eta-squared value to determine what percentage of the variance between groups was attributable to L1 family.

*Impact of L1 orthography.* The impact of first language orthography was explored by disaggregating the sample into two subgroups: ELLs with an alphabetic L1 and ELLs with a non-alphabetic L1. Frequencies and measures of central tendency described performance for the two subgroups. Independent samples t-tests determined whether there was a statistically significant difference in performance. Cohen's d was calculated to determine the effect size, if any, and a univariate analysis of variance yielded a partial eta-squared value to determine what percentage of the variance between groups was attributable to L1 orthography.

*Impact of late-entry ELL status.* The impact of late-entry ELL status (entry at the age of 12 years or later) was explored by disaggregating the sample into two subgroups: late-entry ELLs and not-late-entry ELLs. Frequencies and measures of central tendency described performance for the two subgroups. Independent samples t-tests determined whether there was a statistically significant difference in performance. Cohen's d was

calculated to determine the effect size, if any, and a univariate analysis of variance yielded a partial eta-squared value to determine what percentage of the variance between groups was attributable to late-entry ELL status.

## CHAPTER 4

### RESULTS

This chapter begins with a description of the ELL sample that took the June 2012 Biology MCAS as well as a description of the instrument itself. The chapter then reports the results of ELL performance analyses. Performance results begin with descriptive and comparative statistics of overall performance on the instrument (total raw score and performance level categories) for the sample as a whole and for ELL subgroups disaggregated by first language family, first language orthography, and late-entry ELL status. After reporting overall performance results, the chapter then reports whole sample and subgroup performance for three item attributes: (1) content domain, (2) cognitive skill level, and (3) linguistic complexity. Chapter 4 concludes with a summary of findings.

#### **Sample Characteristics**

As will be discussed more fully in Chapter 5, statewide MCAS results are reported for a particular Grade 10 cohort, the Class of 2014 in the case of the 2012 high school MCAS exams. The ELL sample ( $n = 3,315$ ) in this study included all ELLs

regardless of cohort year, test/retest status, or amount of time in Massachusetts schools.<sup>71</sup> ELLs who took the June 2012 Biology MCAS differed from the Class of 2014 not only in English proficiency but also in age. Sample age and grade demographics are summarized in Tables 4.1 and 4.2. The majority of the ELLs (79.5%) were in Grades 9 and 10, and the majority (51.6%,  $n = 1,713$ ) were 15 to 16 years old; however, ages ranged from 13 years ( $n = 11$ ) to 24 years ( $n = 1$ ). Although an absolute age cannot be associated with a grade, grade levels have a customary and usual age range that overlaps with adjacent grades.<sup>72</sup> The usual age range in the spring would be 14 to 15 years for Grade 9 and 15, to 16 years for Grade 10 (assuming students were 5 to 6 years old when they began Grade 1 in the fall). Of the ELLs who took the June 2012 Biology MCAS, 38.3% ( $n = 1,273$ ) were age 17 years or older. Although these ELLs were older than the typical Grade 9 or Grade 10 student (14 to 16 years old), the majority (54.8%,  $n = 697$ ) were in Grade 9 ( $n = 144$ ) and Grade 10 ( $n = 553$ ).<sup>73</sup> Under most circumstances, Massachusetts students graduate or leave high school before the age of 20 years. The ELL sample, however, included 98 students who were 20 years old or older, and these students were represented in all four grades: Grade 9 ( $n = 6$ ), Grade 10 ( $n = 20$ ), Grade 11 ( $n = 47$ ), and Grade 12 ( $n = 25$ ). Although this older age group represented only 2.9% of the ELL sample, the data suggested that there was a group of ELLs who were far older than their classroom peers;

---

<sup>71</sup> As discussed in Chapter 3, the ELL sample included all of the ELLs who had a March 2012 MEPA score and a June 2012 Biology MCAS score.

<sup>72</sup> Grade levels have overlapping age ranges because birthdays determine the beginning of school enrollment. For example, a student with an August birthday may be a year younger than a student in the same grade with a January birthday. Other factors to consider are grade retention and transfers into Massachusetts schools from places with different age requirements for schooling.

<sup>73</sup> There was one ELL who was 24 years old at the time of the June 2012 Biology MCAS; this ELL's age of entry was 23 years. It is possible that the SIMS data had an error in birthdate because students age out of Massachusetts public schools at 22 years. Thus, it is unlikely that a public school would accept a 23-year-old as a freshman in high school.

over one-quarter (26.5%, n = 26) were in Grades 9 and 10—four to six years older than customary age ranges for these grades. The grade-level placement of newly enrolled secondary ELLs is determined by district-level policies, and the data suggested that age did not appear to be the determining factor in placing ELLs in a grade.

Table 4.1

*ELL Grade and Age Demographics for the June 2012 Biology MCAS*

	<u>n</u>	<u>%</u>
Grade		
9	1146	34.6
10	1490	44.9
11	569	17.2
12	110	3.3
Age		
13-14 years	329	9.9
15-16 years	1713	51.6
17-18 years	998	30.1
19 years	177	5.3
20+ years	98	2.9
Age of Entry		
Before 12 years	1097	33.1
12 years or later	2218	66.9

When looking at the ELL sample by grade level, one in five (20.5%) were in Grades 11 and 12 (17.2% and 3.3%, respectively); most non-ELLs take and pass the Biology MCAS in Grade 9 or 10.<sup>74</sup> Recent immigration partially explains the higher grade levels for ELLs taking the Grade 10 Biology MCAS. Approximately one-third (31.5%, n = 179) of Grade 11 ELLs were in their first year in Massachusetts schools, and 36.4% of Grade 12 ELLs were in their first or second year in Massachusetts schools.

<sup>74</sup>If the high school science curriculum sequence has Biology as a Grade 9 course, these students usually take the Biology MCAS in Grade 9. If they pass, they do not have to take a science MCAS again in Grade 10.

Another partial explanation could be that these older ELLs had insufficient English proficiency and/or content knowledge to pass previous administrations of an STE MCAS and thus were taking the June 2012 Biology MCAS as a retest. Given the research suggesting that it takes five to seven years to attain L2 academic language proficiency, it is not unexpected for older ELLs to have retest status; almost all Grade 11 (91.1%) or Grade 12 (89.1%) ELLs were in Massachusetts schools for seven or fewer years. This study, however, did not have data to differentiate first-time test-takers from ELLs who were retested.

Table 4.2  
*Age Demographics by Grade Level*

	Grade 9		Grade 10		Grade 11		Grade 12	
	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>
<u>Age</u>								
13-14 years	321	28.0	8	0.5	0	0.0	0	0.0
15-16 years	681	59.4	929	62.3	100	17.6	3	2.7
17-18 years	127	11.1	480	32.2	336	59.1	55	50.0
19 years	11	1.0	53	3.6	86	15.1	27	24.5
20+ years	6	0.5	20	1.4	47	8.3	25	22.8
<u>Age of Entry</u>								
Before 12 years	539	47.0	451	30.3	89	15.6	18	16.4
12 years or later	607	53.0	1039	69.7	480	84.4	92	83.6
<u>Years in MA<sup>75</sup></u>								
1 year	236	20.6	302	20.3	179	31.5	19	17.3
2 years	204	17.8	355	23.8	125	22.0	21	19.1
3 years	181	15.8	254	17.0	87	15.3	19	17.3
4 years	129	11.3	144	9.7	54	9.5	15	13.6
5-7 years	169	14.8	230	15.4	73	12.8	24	21.8
8+ years	227	19.7	205	13.8	51	8.9	12	10.9

<sup>75</sup> Years in Massachusetts was taken from the spring 2012 MEPA data.



With respect to age of entry, the majority of ELLs at each grade level had an age of entry of 12+ years and thus were late-entry ELLs as defined in this study. The percentage of late-entry ELLs was the lowest in Grade 9 (53%) and increased through the grade levels; ELLs in Grades 11 and 12 were overwhelmingly late-entry ELLs (84.4% and 83.6%, respectively). The percentage of ELLs in the sample who entered Massachusetts schools as high school students was 41% (n = 1,358). Grade 9 had the lowest percentage (20.6%) with 236 ELLs in their first year in Massachusetts public schools. This was followed by Grade 10 ELLs, of whom 44.1% (n = 657) were in their first or second year in Massachusetts schools. Approximately two-thirds of Grade 11 and Grade 12 ELLs appeared to have entered Massachusetts public schools at the secondary level. There were 67.3% of Grade 12 ELLs (n = 74) in their first through fourth year in Massachusetts public schools, and 68.8% of Grade 11 ELLs (n = 391) were in their first through third years (Table 4.2).

Sheltered English instruction (SEI) is the most common ELL pedagogy in Massachusetts, and 92.4% of the ELLs were in SEI programs. ELL program statistics are summarized in Table 4.3. This study also explored MCAS performance for subgroups disaggregated by: (1) English proficiency, (2) Latinate or non-Latinate L1, (3) alphabetic or non-alphabetic L1, and (4) late-entry ELL status. The subgroup demographics are summarized in Tables 4.4, 4.8, 4.9, and 4.10.

Table 4.3  
*ELL Programs*

ELL Program	<u>n</u>	<u>%</u>
Sheltered English immersion (SEI)	3062	92.4
Other bilingual	186	5.6
Opted out	44	1.3
Not enrolled in a program	22	0.7
Two-way bilingual	1	0.0

**English proficiency level.** English proficiency was assessed by the MEPA instrument in March 2012, a few months before the June 2012 Biology MCAS administration. The English proficiency of ELLs who took the June 2012 Biology MCAS included all five MEPA levels, and Table 4.4 summarizes measures of central tendency. English proficiency, as measured by the MEPA scaled score, exhibited a normal distribution with a skewness value of -.28 and a kurtosis value of -.02. The MEPA scaled score mean was 480.09 (SD = 20.73), which was at the higher end of the scaled score range (464 to 488) for a MEPA Level 3 proficiency level. The greatest number of ELLs (43.5%, n = 1,441) were at MEPA Level 3, where a student “communicates using basic English at school, although errors sometimes interfere with communication and understanding” (MA DESE, 2012f, p. 4). An additional 34.9% of the ELLs were above MEPA Level 3, and 21.6% were below MEPA Level 3.

Table 4.4  
*English Proficiency Level*

	<u>M</u>	<u>SD</u>
MEPA Scaled Score (400 to 550)	480.09	20.73
English proficiency	<u>n</u>	<u>%</u>
MEPA Level 1	276	8.3
MEPA Level 2	441	13.3
MEPA Level 3	1,441	43.5
MEPA Level 4	603	18.2
MEPA Level 5	554	16.7

This study explored performance for ELL subgroups based on first language family (Linate/non-Linate), orthography (alphabetic/non-alphabetic), and late-entry ELL status. Table 4.5 summarizes the measures of central tendency for English proficiency levels of these subgroups. The difference between the mean MEPA scaled score for the Linate/non-Linate subgroups was minor, and the mean for both groups fell into MEPA Level 3 proficiency range. Likewise, the mean MEPA scaled score for alphabetic/non-alphabetic subgroups was nearly identical, and both subgroups had a mean MEPA scaled score that fell into the MEPA Level 3 proficiency range. When looking at English proficiency for the late-entry/not-late-entry subgroups, the mean MEPA scaled score for not-late-entry ELLs was approximately 16 points higher. This difference was enough to change the mean MEPA proficiency levels between the two groups. The mean MEPA scaled score of 474.68 for late-entry ELLs corresponded to MEPA Level 3 proficiency, and the mean MEPA scaled score of 491.02 for not-late-entry ELLs corresponded to MEPA Level 4 proficiency.

Table 4.5  
*English Proficiency Means by Subgroups*

	Lاتinate L1 n = 2,420		Non-Lاتinate L1 n = 895		<u>t</u>
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	
MEPA Scaled Score (400 to 550)	479.03	20.89	482.96	20.00	-4.86***
	Alphabetic L1 n = 2,724		Non-alphabetic L1 n = 591		<u>t</u>
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	
MEPA Scaled Score (400 to 550)	479.93	20.88	480.80	20.02	-.92
	Late-entry n = 2,218		Not-late-entry n = 1,097		<u>t</u>
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	
MEPA Scaled Score (400 to 550)	474.68	20.98	491.02	15.17	-25.58***

Note: \*p < .05 \*\*p < .01 \*\*\*p < .000

<sup>a</sup>Levene's test indicated that equal variance was not assumed.

**First language characteristics.** The ELLs in the sample were linguistically diverse with over 70 first languages (Appendix F), and Tables 4.6, 4.7, 4.8, and 4.9 summarize linguistic demographics. Slightly more than half (51.2%) of the ELLs who took the June 2012 Biology MCAS spoke Spanish as their first language. The ten most frequent first languages accounted for 86% of the participants. There were 48 first languages with 10 or fewer speakers, and these represented 161 test-takers. In addition, there were 35 ELLs with a first language categorized as *Other* by MA DESE.

Table 4.6

*Ten Most Common First Languages for ELLs Who Took the June 2012 Biology MCAS*

	<u>n</u>	<u>%</u>		<u>n</u>	<u>%</u>
Spanish	1,697	51.2	Chinese (not Mandarin or Cantonese)	103	3.1
Haitian Creole	285	8.6	Vietnamese	82	2.5
Cape Verdean	227	6.8	Khmer/Khmai	78	2.4
Portuguese	160	4.8	Mandarin	60	1.8
Arabic	107	3.2	Nepali	52	1.6

The first language codes in SIMS differentiate between Chinese languages:

Chinese, not Mandarin or Cantonese (code 002); Canton (code 200); Hakka Dialect (code 350); and Mandarin (code 520).<sup>76</sup> If these languages are combined into a Chinese language family category, then Chinese languages becomes the fourth most common first language (n = 208, 6.1%), and French (n = 44, 1.3%) becomes the tenth most common first language. These ten most common first languages accounted for 88.5% (n = 2,935) of the ELL test-takers. That five of the ten most common first languages were non-Latinate languages reflects the diversity of the Massachusetts ELL population.

Table 4.7

*Ten Most Common First Languages with Combined Chinese Languages*

	<u>n</u>	<u>%</u>		<u>n</u>	<u>%</u>
Spanish	1,697	51.2	Arabic	107	3.2
Haitian Creole	285	8.6	Vietnamese	82	2.5
Cape Verdean	227	6.8	Khmer/Khmai	78	2.4
Chinese languages	203	6.1	Nepali	52	1.6
Portuguese	160	4.8	French	44	1.3

<sup>76</sup> Other SIMS codes for Chinese languages include Fukien (code 315) and Shanghai (code 695); however, no ELLs with these first languages took the June 2012 Biology MCAS.

***First language family.*** The three most common first languages (Spanish, Haitian Creole, and Cape Verdean) were Latinate and represented 66.6% ( $n = 2,209$ ) of ELL first languages in this study (Table 4.6). Thus, it was not unexpected that nearly three-quarters (73%,  $n = 2,420$ ) of the ELLs had a Latinate first language. Table 4.8 summarizes the demographics for the Latinate L1 and non-Latinate L1 groups. Like the whole sample, both groups had the greatest number of ELLs at MEPA Level 3 (43.6% and 43.1%, respectively). Both groups also had similar percentages for Levels 2, 3, and 4; however, the non-Latinate L1 group had a slightly lower percentage at Level 1 and a slightly higher percentage at Level 5 compared to the Latinate L1 group. An independent samples t-test on the mean MEPA scaled score explored whether the small differences at the extreme ends of English proficiency were statistically significant. Table 4.5 summarizes these results. Although the results indicated a statistically significant difference in English proficiency,  $t(3313) = -4.86$ ,  $p < .001$ , Cohen's  $d (.19)$  determined that the effect size favoring non-Latinate L1 ELLs was negligible. Thus, any performance difference between these groups was not attributable to differences in English proficiency.

Since all Latinate languages use the Roman alphabet, all Latinate L1 ELLs had an alphabetic L1. For the non-Latinate L1 ELLs ( $n = 895$ ), approximately one-third (34%,  $n = 304$ ) had an alphabetic L1 and approximately two-thirds (66%,  $n = 591$ ) had a non-alphabetic L1 (Appendix F). Another difference between the Latinate/non-Latinate L1 groups was that a greater percentage of the non-Latinate L1 ELLs (74.5%) were late-entry compared to the Latinate L1 ELLs (64.1%). First language orthography and late-entry ELL status demographics are discussed more fully below.

Table 4.8  
*Demographics by L1 Language Family*

	Linate L1 (n = 2,420)		Non-Linate L1 (n = 895)	
English Proficiency	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>
MEPA Level 1	233	9.6	43	4.8
MEPA Level 2	329	13.6	112	12.5
MEPA Level 3	1,055	43.6	386	43.1
MEPA Level 4	433	17.9	170	19.0
MEPA Level 5	370	15.3	184	20.6
L1 Orthography				
Alphabetic L1	2,420	100.0	304	34.0
Non-alphabetic L1			591	66.0
Age of Entry				
Late-entry ELL	1,551	64.1	667	74.5
Not-late-entry ELL	869	35.9	228	25.5

***First language orthography.*** First languages were classified for orthography, and the researcher encountered some difficulty in the classification of orthographies as alphabetic or non-alphabetic.<sup>77</sup> The *Compendium of the World's Languages* (Campbell, 1991a, 1991b) was the primary resource, and it contained information for most of the 70 first languages represented by the test-takers. There were, however, some languages that were not included in this reference. In these cases, the researcher consulted Internet resources for orthographic samples and made a visual determination whether it essentially used the Roman alphabet.<sup>78</sup> Table 4.9 summarizes the demographics for the alphabetic L1 and non-alphabetic L1 groups. The majority (82.2%, n = 2,724) of ELLs had a first language with one of the following alphabetic orthographies: (1) Roman (n =

<sup>77</sup> First languages classified as *Other* by MA DESE were categorized as non-Linate and non-alphabetic.

<sup>78</sup> Several languages used the Roman alphabet with some additional letters and/or diacritics to represent sounds specific to that language.

2,703), (2) Cyrillic (n = 18), or Greek (n = 3).<sup>79</sup> This was not unexpected since 73% of the ELLs had a Latinate L1.

Table 4.9  
*Demographics by L1 Orthography*

	Alphabetic L1 (n = 2,724)		Non-alphabetic L1 (n = 591)	
	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>
English proficiency				
MEPA Level 1	242	8.9	34	5.8
MEPA Level 2	358	13.1	83	14.0
MEPA Level 3	1,177	43.2	264	44.7
MEPA Level 4	491	18.0	112	19.0
MEPA Level 5	456	16.7	98	16.6
L1 language family				
Latinate	2,420	88.8		
Non-Latinate	304	11.2	591	100.0
Age of Entry				
Late-entry ELL	1,763	64.7	455	77.0
Not-late-entry ELL	961	35.3	136	23.0

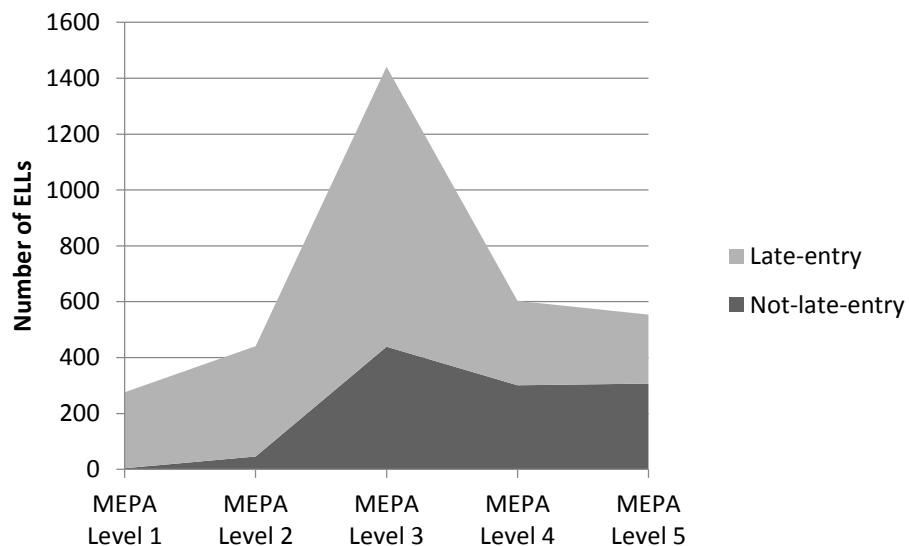
Like the whole sample, both groups had the greatest number of ELLs at MEPA Level 3: Alphabetic L1 (43.2%, n = 1,177) and non-alphabetic L1 (44.7%, n = 264). With the exception of MEPA Level 1, the frequencies at each MEPA level were similar. For MEPA Level 1, there were relatively fewer ELLs with a non-alphabetic L1 (5.8%) than ELLs with an alphabetic L1 (8.9%); however, this difference was minor. An independent samples t-test on the mean MEPA scaled score confirmed that the minor difference between the groups for Level 1 was not statistically significant at the  $p < .05$  level (Table 4.5). Thus, any performance difference between these groups was not

<sup>79</sup> Although this study defined alphabetic orthography as the Roman, Cyrillic, or Greek alphabets, the Cyrillic and Greek alphabets only represented 21 ELLs, less than 1% of the sample.



attributable to differences in English proficiency. As expected, the majority of ELLs with an alphabetic L1 also had a Latinate L1 (88.8%,  $n = 2,420$ ); 11.2% ( $n = 304$ ) had a non-Latinate L1. The majority of both groups were late-entry ELLs; however, a greater percentage of the non-alphabetic L1 group (77%) were late-entry ELLs compared to the alphabetic L1 group (64.7%).

**Late-entry ELLs.** Age of entry ranged from 4 to 23 years, and the majority of ELLs (66.9%) were late-entry ELLs as defined herein (Tables 4.1 and 4.2). Table 4.10 summarizes sample demographics for the late-entry ELL and not-late-entry ELL subgroups. The greatest number of late-entry ELLs (45.2 %) was at MEPA Level 3, with 30.1% below Level 3 and 24.7% above. In comparison, although the greatest number (40%) of not-late-entry ELLs was at MEPA Level 3, English proficiency for this group was negatively skewed, with the majority (55.4%) above Level 3 and only 4.5% below



*Figure 4.1.* MEPA Performance Levels by Late-Entry ELL Status. ELLs who entered before the age of 12 years are negatively skewed.

The mean MEPA scaled score for late-entry ELLs ( $n = 2,218$ ) was 474.68 ( $SD = 20.98$ ), which fell in the Level 3 proficiency range (464 to 488). In comparison, not-late-entry ELLs ( $n = 1,097$ ) had a mean of 491.02 ( $SD = 15.17$ ), which fell in the Level 4 proficiency range (489 to 499). Independent samples  $t$ -tests confirmed that English proficiency means differed significantly between late-entry ELLs and not-late-entry ELLs,  $t(2877.11) = -25.58$ ,  $p < .001$ . Table 4.5 summarizes the  $t$ -test results. This was not unexpected since not-late-entry ELLs would have been in the United States longer, had more exposure to English, and likely have progressed beyond the lowest levels of English proficiency. The demographic data suggested that not-late-entry ELLs might perform better on the Biology MCAS because of their higher English proficiency levels.

Table 4.10

*ELL Demographics by Late-Entry Status*

	Late-Entry ELL ( $n = 2,218$ )		Not-Late-Entry ELL ( $n = 1,097$ )	
	<u><math>n</math></u>	<u>%</u>	<u><math>n</math></u>	<u>%</u>
English Proficiency				
MEPA Level 1	272	12.3	4	0.4
MEPA Level 2	395	17.8	46	4.2
MEPA Level 3	1,002	45.2	439	40.0
MEPA Level 4	302	13.6	301	27.4
MEPA Level 5	247	11.1	307	28.0
L1 Language Family				
Latinate L1	1,551	69.9	869	79.2
Non-Latinate L1	667	30.1	228	20.8
L1 Orthography				
Alphabetic L1	1,763	79.5	961	87.6
Non-alphabetic L1	455	20.5	136	12.4

For the sample as a whole, 73% had a Latinate L1 (Table 4.8). When the ELL sample was disaggregated by late-entry ELL status, the majority of both groups still had a Latinate L1, but the proportion of Latinate L1 ELLs was greater in the not-late-entry group (79.2% compared to 69.9% of the late-entry group). There was a similar pattern for L1 orthography. For the sample as a whole, 82.2% had an alphabetic L1 (Table 4.9). When the sample was disaggregated by late-entry ELL status, the percentage of alphabetic L1 was higher for not-late-entry ELLs (87.6%) compared to late-entry ELLs (79.5%). The data suggested that for the ELLs in this sample, those who had a non-Latinate or non-alphabetic L1 generally entered later.

**Summary of sample characteristics.** The ELLs in the sample who took the June 2012 Biology MCAS were generally older than their native-English-speaking peers in Grades 9 and 10. The demographic data suggested that 41% (n = 1,358) of the ELLs in the sample entered Massachusetts schools as high school students. These students have at most four years to acquire enough English to demonstrate proficiency in English language arts, mathematics, and biology (or another science content area) for high school graduation and for post-secondary studies. Given the research on second language acquisition, especially with respect to academic language, these ELLs face a seemingly insurmountable task.

Most of the ELLs (78.4%) in the sample were at MEPA Level 3 or above, where they are acquiring academic English, but only 34.9% reached MEPA Level 4, where they understand “basic grade-level academic vocabulary” (MA DESE, 2012f, p. 6). The ELLs in the sample were linguistically diverse, with over 70 first languages, but the majority (73%) had a Latinate L1. The majority (82.2%) also had an alphabetic L1; however, four

of the ten most common first languages were non-alphabetic (Chinese languages, Arabic, Khmer/Khmer, and Nepali). Any difference in English proficiency among the Latinate/non-Latinate and alphabetic/non-alphabetic subgroups was either not statistically significant or negligible.

The majority of ELLs (66.9%) were late-entry ELLs as defined by this study (i.e., age of entry was 12+ years). The majority of ELLs in both the Latinate L1 and non-Latinate L1 groups were late-entry, with a slightly higher percentage in the latter group (64.1% and 74.5%, respectively). Likewise, a majority of ELLs in both the alphabetic L1 and non-alphabetic L1 groups were late-entry ELLs with a slightly higher percentage for non-alphabetic L1 ELLs (64.7% and 77%, respectively). The secondary data for this study did not include any data for L1 education or L1 science knowledge; however, some of these older ELLs undoubtedly studied biology and other sciences in their home country. The greatest difference between the late-entry and not-late-entry ELLs was in their English proficiency. Although both groups had the greatest number at MEPA Level 3, the English proficiency for the not-late-entry ELLs was negatively skewed, with the majority (55.4%) above at MEPA Levels 4 and 5 and only 4.5% below at MEPA Levels 1 and 2. For the ELLs in this sample, the demographic data suggested that ELLs who entered the U.S. before the age of 12 years had higher levels of English proficiency.

### **Analysis of the MCAS Instrument**

The June 2012 Biology MCAS instrument consisted of 40 multiple-choice items and five constructed-response items. In addition to ELL performance on the whole instrument, this study also explored ELL performance on three attributes of the multiple-choice items: (1) content domain, (2) cognitive skill level, and (3) linguistic complexity.

Tables 4.11 and 4.12 summarize the content domain and cognitive skill level attributes of the 40 multiple-choice items, and Tables 4.13 and 4.16 summarize the linguistic elements.

**Content domains.** The instrument followed the test blueprint for the six Massachusetts high school biology standards: anatomy and physiology (n = 5), biochemistry (n = 5), cell biology (n = 6), ecology (n = 8), evolution (n = 8), and genetics (n = 8) (MA DESE, 2013a).<sup>80</sup> Table 4.11 summarizes the distribution of multiple-choice items across the content domains. The instrument had slightly more emphasis on three domains (ecology, evolution, and genetics), which represented 60% of the multiple-choice items; the remaining 40% were anatomy and physiology, biochemistry, and cell biology items. The distribution of multiple-choice items across the six standards allowed test-takers multiple entry points. If a test-taker was particularly weak in one of the six content domains, it would not preclude passing the Biology MCAS since no content domain represented more than 20% of the multiple-choice items.<sup>81</sup>

---

<sup>80</sup> The six Massachusetts high school biology standards are: (1) the Chemistry of Life (biochemistry), (2) Cell Biology, (3) Genetics, (4) Anatomy and Physiology, (5) Evolution and Biodiversity (evolution), and (6) Ecology. The content domains are reported in alphabetical order in the tables.

<sup>81</sup> The five constructed-response items covered the following domains: (1) anatomy and physiology (item 32), (2) cell biology (item 23), ecology (item 45), evolution (item 12), and genetics (item 44).

Table 4.11

*Domain and Cognitive Skill Level Item Attributes of the Multiple-Choice Items on the June 2012 Biology MCAS*

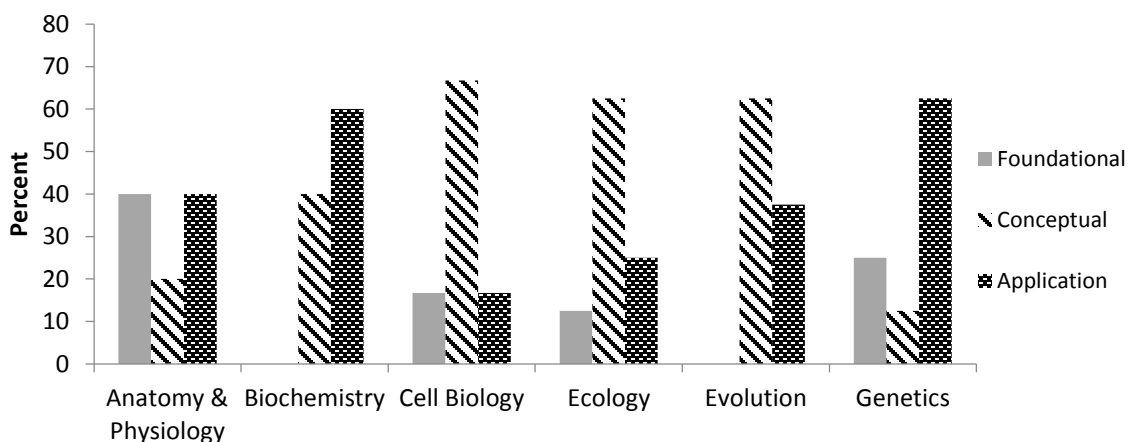
<u>Domain</u>	<u>n</u>	<u>%</u>	<u>Cognitive Skill Level</u>	<u>n</u>	<u>%</u>
Anatomy & Physiology	5	12.5	Foundational	6	15.0
Biochemistry	5	12.5	Conceptual	18	45.0
Cell biology	6	15.0	Application	16	40.0
Ecology	8	20.0			
Evolution	8	20.0			
Genetics	8	20.0			

**Cognitive skill level.** The 40 multiple-choice items represented three cognitive skill levels: foundational (n = 6), conceptual (n = 18), and application (n = 16). Table 4.11 summarizes the distribution of the multiple-choice items across the cognitive skill levels, and Table 4.12 and Figure 4.2 summarize the distribution across the six domains. Although the foundational cognitive skill level only represented 15% of the multiple-choice items, it represented 40% of the anatomy and physiology items, and there were no foundational items for biochemistry or evolution. Three domains—cell biology, ecology, and evolution—had predominately conceptual items. Biochemistry and genetics items were predominantly at the application skill level (60% and 62.5%, respectively).

Table 4.12  
*Multiple-Choice Cognitive Skill Level across Content Domains on the June 2012 Biology MCAS*

<u>Domain</u>	Cognitive Skill Level					
	Foundational		Conceptual		Application	
	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>
Anatomy & Physiology	2	40.0	1	20.0	2	40.0
Biochemistry	0	0.0	2	40.0	3	60.0
Cell Biology	1	16.7	4	66.7	1	16.7
Ecology	1	12.5	5	62.5	2	25.0
Evolution	0	0.0	5	62.5	3	37.5
Genetics	2	25.0	1	12.5	5	62.5

As discussed in Chapter 2, item difficulty and cognitive skill level cannot be equated. There can be items of varying difficulty within a cognitive skill level such as an easy application item or a difficult conceptual item. This study explored performance at cognitive skill levels from the perspective of cognitive load and second language acquisition. For ELLs who are still working toward English proficiency, working memory resources are needed for comprehension of the item's language, which reduces the working memory available for the item's content. It would be expected that as working memory resources are needed for higher cognitive skill demands, combined with the working memory demands for the language, performance would decrease as cognitive skill level increased. On the surface, this would suggest that ELLs would perform best on anatomy and physiology items and have lower performance on biochemistry and genetics items, a majority of which are at the application level.



*Figure 4.2.* Percent Distribution of Item Cognitive Skills across Domains on the June 2012 Biology MCAS. Biochemistry and genetics had more than 50% of the items at the application cognitive skill level.

**Linguistic complexity.** The textual analysis in Phase I operationalized item linguistic complexity. Frequency analyses yielded measures of central tendency for the linguistic variables, which are summarized in Table 4.13. The data show a wide range of values for the lexical, syntactic, and discourse elements. Item stems ranged from nine to 103 words and from one to 15 sentences; the average number of words in a stem's sentences ranged from six to 24. Although this is a wide range, the majority of items had one or two sentences in the stem; seven items (17.5%) had one sentence, and an additional 16 items (40%) had two sentences (Appendix G). The majority of items (52.5%,  $n = 21$ ) had 21 to 40 words in the stem with approximately an equal number of items with less than 20 words (25%,  $n = 10$ ) and more than 40 words (22.5%,  $n = 9$ ). It appeared that although some item stems had a high linguistic demand in terms of word number, the majority of items provided entry points for understanding the question. The same was true for the average number of words in the answer options. The average number of words in an answer option ranged from one to 22 words; however, nearly half



(47.5%) of the multiple-choice items had an average of one to three words in the answer options. Likewise, the estimated Lexile<sup>®</sup> score for items ranged widely from 540 to 1520. The data indicated that some items appeared to have high linguistic demands; however, the wide ranges indicated that, from a language perspective, there appeared to be entry points for all students.

Table 4.13

*Measures of Central Tendency of Linguistic Elements of the Multiple-Choice Items on the June 2012 Biology MCAS*

n = 40	<u>Mean</u>	<u>SD</u>	<u>Range</u>
<u>Lexical Variables</u>			
SLD: Stem Lexical Density	33.25	18.94	9 - 103
TALD: Total Words in 4 Answer Options	22.42	15.50	4 - 87
ALD: Average words in answer option	5.61	4.94	1 – 21.75
TLD: Total words in stem and answer options	55.68	28.50	21 - 141
<u>Syntactic Variables</u>			
SS: Sentences in stem	3.15	2.78	1 - 15
SSD: Average words in stem sentence	12.47	4.35	5.73 - 24
<u>Discourse Variable</u>			
RCS: Reading complexity score (Lexile Analyzer <sup>®</sup> )	946	250.95	540 - 1520

***Content domain linguistic complexity.*** The linguistic element data were explored further for patterns in the content domains. Table 4.16 summarizes measures of central tendency for the linguistic variables for each domain. The data suggested a difference in total answer lexical density (TALD) among the domains. Evolution items had a mean TALD of 46.62 (SD = 23.87), approximately four times greater than biochemistry items (M = 12.40, SD = 15.52) or cell biology items (M = 10.33, SD = 5.85), and at least twice

greater than the other domains. One-way ANOVA analyses confirmed that there was a statistically significant difference in TALD among domains,  $F(5,34) = 4.91$ ,  $p < .01$ . Scheffé post hoc analyses determined that these differences were statistically significant at the  $p < .05$  level between: (1) evolution and biochemistry items, and (2) evolution and cell biology items. Table 4.14 summarizes the results of the post hoc analyses. The higher TALD suggested that evolution items might be problematic for ELLs, even if they understood the content and the stem because differentiating between the answer options required parsing more words for comprehension.

Table 4.14  
*Summary of Scheffé Post Hoc Analysis on Total Answer Lexical Density (TALD) across Content Domains*

Domain	Mean Difference (I-J)				
	<u>I1</u>	<u>I2</u>	<u>I3</u>	<u>I4</u>	<u>I5</u>
J2	9.00				
J3	11.07	2.07			
J4	1.4	-7.60	-9.67		
J5	-25.22	-34.22*	-36.29*	-26.62	
J6	4.78	-4.22	-6.29	3.38	30.00*

Note: 1 = anatomy and physiology, 2 = biochemistry, 3 = cell biology, 4 = ecology, 5 = evolution, and 6 = genetics

\* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$

***Cognitive skill linguistic complexity.*** The linguistic element data were also explored for patterns in the content domains. Table 4.16 summarizes measures of central tendency for the linguistic variables for each cognitive skill level. The data indicated a pattern for stem lexical density (SLD) and stem syntax (SS) where the means for these linguistic elements increased as the cognitive skill levels increased. Table 4.16 summarizes the measures of central tendency. Foundational items had the smallest mean

for the number of words in the stem ( $M = 20.50$ ,  $SD = 9.16$ ), followed by conceptual items ( $M = 27.89$ ,  $SD = 9.62$ ), and then by application items ( $M = 44.06$ ,  $SD = 24.00$ ), which was more than twice the mean for foundational items. Likewise, foundational items had the smallest mean for number of sentences in the stem ( $M = 1.67$ ,  $SD = 1.21$ ), followed by conceptual items ( $M = 2.33$ ,  $SD = 0.84$ ), and then by application items ( $M = 4.62$ ,  $SD = 3.84$ ), which was two times greater than foundational or conceptual items. One-way ANOVA analyses confirmed that there were statistically significant differences in both the number of words and sentences among the cognitive skill levels,  $F(2, 37) = 5.85$ ,  $p < .01$ , and,  $F(2, 37) = 4.60$ ,  $p < .05$ , respectively. Table 4.15 summarizes the Scheffé post hoc analyses. The Scheffé post hoc analyses determined that for the number of words in the stem (SLD), these differences were statistically significant at the  $p < .05$  level between: (1) foundational and application items, and (2) conceptual and application items. Scheffé post hoc analyses also determined that for the number of sentences in the stem (SS), these differences were statistically significant at the  $p < .05$  level between conceptual and application items. The data suggested that from a linguistic perspective, application items might pose difficulties for ELLs.

Table 4.15

*Summary of Scheffé Post Hoc Analysis on Linguistic Elements across Cognitive Skill Levels*

	(I) Cognitive Skill	(J) Cognitive Skill	Mean Difference (I-J)
Stem Lexical Density (SLD)	Foundational	Conceptual	-7.39
		Application	-23.56*
	Conceptual	Foundational	7.39
		Application	-16.17*
	Application	Foundational	23.56*
		Conceptual	16.17*
Stem Syntax (SS)	Foundational	Conceptual	-.67
		Application	-2.96
	Conceptual	Foundational	.67
		Application	-2.29*
	Application	Foundational	2.96
		Conceptual	2.29*

\*p ≤ .05, \*\*p ≤ .01, \*\*\*p ≤ .001

Table 4.16

*Measures of Central Tendency for Linguistic Complexity Variables by Domain and Cognitive Skill Level*

<b><u>Domain</u></b>	<b><u>n</u></b>	<b><u>SLD</u></b>		<b><u>TALD</u></b>		<b><u>TLD</u></b>		<b><u>SS</u></b>		<b><u>SSD</u></b>		<b><u>RCS</u></b>	
		<b><u>M</u></b>	<b><u>SD</u></b>	<b><u>M</u></b>	<b><u>SD</u></b>	<b><u>M</u></b>	<b><u>SD</u></b>	<b><u>M</u></b>	<b><u>SD</u></b>	<b><u>M</u></b>	<b><u>SD</u></b>	<b><u>M</u></b>	<b><u>SD</u></b>
Anatomy & physiology	5	25.80	12.03	21.40	14.76	47.20	15.09	2.40	2.07	14.07	7.06	902.00	359.26
Biochemistry	5	35.60	16.91	12.40	15.52	48.00	23.32	3.60	3.05	11.61	2.80	972.00	118.62
Cell biology	6	28.50	9.31	10.33	5.85	38.83	8.04	2.17	0.98	14.21	4.53	1010.00	208.13
Ecology	8	39.00	27.58	20.00	15.30	59.00	39.91	3.50	2.78	12.02	3.68	950.00	285.96
Evolution	8	32.00	15.24	46.62	23.87	78.62	33.07	2.25	4.60	14.19	3.54	982.50	282.32
Genetics	8	35.50	24.28	16.62	13.12	52.13	24.80	4.62	2.78	9.44	3.60	868.75	250.80
<b><u>Cognitive Skill</u></b>	<b><u>n</u></b>	<b><u>SLD</u></b>		<b><u>TALD</u></b>		<b><u>TLD</u></b>		<b><u>SS</u></b>		<b><u>SSD</u></b>		<b><u>RCS</u></b>	
		<b><u>M</u></b>	<b><u>SD</u></b>	<b><u>M</u></b>	<b><u>SD</u></b>	<b><u>M</u></b>	<b><u>SD</u></b>	<b><u>M</u></b>	<b><u>SD</u></b>	<b><u>M</u></b>	<b><u>SD</u></b>	<b><u>M</u></b>	<b><u>SD</u></b>
Foundational	6	20.50	9.16	11.00	8.46	31.50	8.34	1.67	1.21	14.71	6.76	986.67	313.86
Conceptual	18	27.89	9.62	23.44	18.29	51.33	21.73	2.33	0.84	12.57	3.65	922.22	219.43
Application	16	44.06	24.00	25.56	23.41	69.62	33.00	4.62	3.84	11.52	4.00	957.50	273.58

***Summary of linguistic elements.*** For the multiple-choice items on the June 2012 Biology MCAS, the data suggested that the statistically significant linguistic difference among the domains was in the answer options (TALD). In comparison, the statistically significant linguistic difference among cognitive skill levels was in the item stem (SLD and SS). When the linguistic element differences among domains and cognitive skill levels are taken together, evolution items at the application level could be the most difficult for ELLs since both the stem and answer options had a statistically significant higher number of words. This would be followed by application items, which had a statistically significant higher number of words and sentences in the stem. Thus, one might expect ELL performance to be lower for evolution items at the application level as well as for biochemistry and genetics items where 60% and 62.5%, respectively, were at the application level. These potential areas of difficulty for ELLs represented 11, or 27.5%, of the multiple-choice items.<sup>82</sup>

***Composite linguistic complexity.*** As discussed in Chapter 3, the linguistic variables of total lexical density (TLD), stem syntactic density (SSD), and reading complexity score (RCS) were normalized to compute a composite linguistic complexity (CLC) variable that had potential values ranging from 0 to 3. Table 4.17 summarizes measures of central tendency for the normalized variables and CLC.<sup>83</sup> The normalized values for TLD were positively skewed, and normalized SSD had a relatively flat distribution with an exaggerated peak below the mean; normalized RCS values were multimodal. The CLC values exhibited a leptokurtic distribution with a peak just below

---

<sup>82</sup> Three items were evolution items at the application level, three items were biochemistry items at the application level, and five were genetics items at the application level.

<sup>83</sup> See Appendix H for the CLC values for the 40 multiple-choice items.

the mean of 1.07. As with the underlying variables discussed previously, it appeared that most multiple-choice items fell in the lower end of the linguistic complexity range and would have entry points for ELLs.

Table 4.17

*Measures of Central Tendency for Normalized Linguistic Variables and Composite Linguistic Complexity of the Multiple-Choice Items on the June 2012 Biology MCAS*

n = 40	<u>Mean</u>	<u>SD</u>
TLD <sub>N</sub> : Total words in stem and answer options	.29	.24
SSD <sub>N</sub> : Average words in stem sentence	.37	.24
RCS <sub>N</sub> : Reading complexity score (Lexile Analyzer <sup>®</sup> )	.41	.26
CLC: TLD <sub>N</sub> + SSD <sub>N</sub> + RCS <sub>N</sub>	1.07	.54

Table 4.18 and Figure 4.3 summarize the distribution of low, medium, and high CLC across the six content domains. Two domains had the majority of items at low linguistic complexity as measured by the CLC variable, anatomy and physiology (60%) and genetics (50%). Two domains had the majority of items at high linguistic complexity, cell biology (50%) and evolution (50%).

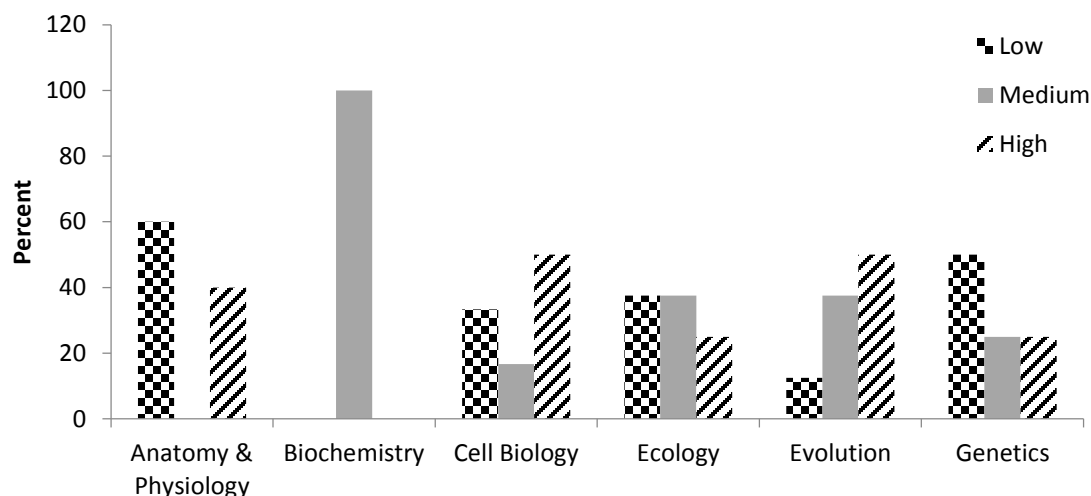
Table 4.18

*Percent Composite Linguistic Complexity Items across Domains*

<u>Content Domain</u>	Linguistic Complexity					
	Low		Medium		High	
	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>
Anatomy & Physiology	3	60.0	0	0.0	2	40.0
Biochemistry	0	0.0	5	100.0	0	0.0
Cell Biology	2	33.3	1	16.7	3	50.0
Ecology	3	37.5	3	37.5	2	25.0
Evolution	1	12.5	3	37.5	4	50.0
Genetics	4	50.0	2	25.0	2	25.0

Biochemistry items were unusual in that they were all medium CLC. Since at least half of the multiple-choice items were at low linguistic complexity, ELLs might have higher performance for the anatomy and physiology and genetics content domains. From a perspective of second language acquisition and cognitive load, ELLs might have more difficulty understanding cell biology and evolution items, followed by biochemistry items irrespective of domain knowledge.





*Figure 4.3.* Percent Composite Linguistic Complexity Items across Domains. The majority of anatomy and physiology and genetics were at low CLC, and cell biology and evolution had the majority of items at high CLC.

Table 4.19 summarizes the distribution of low, medium, and high CLC across the cognitive skill levels. Although a pattern did not emerge, over half (50.1%) of foundational items had high CLC. Conceptual items had the most equal distribution of CLC, with 38.9% low CLC, 33.3% medium CLC, and 27.8% high CLC. The greatest number of application items had medium CLC (43.8%), followed by 31.3% high CLC and 25% low CLC.

Table 4.19

*Item Composite Linguistic Complexity across Cognitive Skill Levels*

<u>Cognitive Skill Level</u>	Linguistic Complexity					
	Low CLC		Medium CLC		High CLC	
	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>
Foundational	2	33.3	1	16.7	3	50.1
Conceptual	7	38.9	6	33.3	5	27.8
Application	4	25.0	7	43.8	5	31.3

**Summary of June 2012 MCAS instrument.** The June 2012 Biology MCAS followed its technical blueprint. Multiple-choice items across six domains, three cognitive skill levels, and a range of linguistic complexity ensured that all students had entry points to the assessment. The Biology MCAS, however, is written for and normed on English-proficient test-takers. The Phase I textual analyses highlighted domains that might be more challenging for ELLs through a combination of cognitive skill and linguistic complexity. As cognitive skill level increases, the cognitive load increases; however, it must be remembered that an increased cognitive skill does not translate to increased item difficulty. As linguistic complexity increases, cognitive load increases. Increased cognitive load from linguistic elements reduces the working memory available for content, and ELLs might have difficulty demonstrating domain knowledge. The domains that might increase cognitive load for ELLs appeared to be biochemistry, evolution, cell biology, and genetics. Biochemistry items might have been challenging to ELLs since there were no items at the foundational cognitive skill level or with low composite linguistic complexity. Evolution items might have been challenging since there were no items at the foundational cognitive skill level, and the majority of items had high composite linguistic complexity. Cell biology had the majority of items with high composite linguistic, and genetics had the majority of items at the application cognitive skill level.

## **ELL Performance on the June 2012 Biology MCAS**

The maximum possible score on the June 2012 Biology MCAS was 60 points.<sup>84</sup>

The score-cut points for performance levels on the June 2012 Biology MCAS were: (1) Fail: 0 to 21 points, (2) Needs Improvement: 22 to 32 points, (3) Proficient: 33 to 47 points, and (4) Advanced: 48 to 60 points. Performance levels were explored with descriptive statistics, and this study followed the MA DESE recommendation to use unscaled MCAS scores for analyses (MA DESE, 2012k).<sup>85</sup> Table 4.20 summarizes ELL performance on the June 2012 Biology MCAS. English language learner scores ranged from four to 57 points ( $M = 22.12$ ,  $SD = 9.53$ ), and approximately half (52.8%,  $n = 1748$ ) passed. Their MCAS scores clustered just below the mean with an exaggerated peak that corresponded to the high end of the Fail performance range; however, the kurtosis value of 0.5 was within the acceptable range to treat the data as a normal distribution.

---

<sup>84</sup> As discussed in Chapter 3, the scores are scaled and reported as even numbers from 220 to 280. A scaled score of 218 and below was an MCAS Performance Level of Fail. Scaled scores for Needs Improvement were 220 to 238, for Proficient 240 to 258, and for Advanced 260 to 280.

<sup>85</sup> See Chapter 3 for a discussion on the limitations of not having unscaled MEPA scores.

Table 4.20  
*ELL Performance on the June 2012 Biology MCAS*

	<u>n</u>	<u>M</u>	<u>SD</u>
MCAS score	3,315	22.12	9.53
MCAS performance level	<u>n</u>	<u>%</u>	
Fail	1,567	47.3	
Needs improvement	1,256	37.9	
Proficient	440	13.3	
Advanced	52	1.6	

The greatest number of ELLs (47.3%) scored at the Fail performance level, and an additional 37.9% scored at the Needs Improvement level; only 14.9% of ELLs scored Proficient or higher. The data clearly indicated an achievement gap when compared to the statewide STE results reported for the Class of 2014, of which 91% passed an STE exam and 69% scored Proficient or higher (MA DESE, 2012i).<sup>86</sup> This was not unexpected since it is well-established in the literature that ELLs do not perform as well as native English speakers on standardized assessments (Abedi & Dietel, 2004; Cook et al., 2011; Duran, 2008; Xu & Drame, 2008).

English language learners in our public schools, however, are a diverse population with wide ranges of English proficiency, linguistic background, background knowledge, and academic experiences. Looking at ELL performance as a single statistic homogenizes their heterogeneity. This study explored ELL performance beyond the single statistic to gain a better understanding of the ELL achievement gap. This study disaggregated the

---

<sup>86</sup>The statewide statistics were for all STE exams for the Class of 2014 and included results from both the February 2012 Biology MCAS and the June 2012 Biology MCAS, as well as the physics, chemistry, and technology/engineering MCAS exams. See Chapter 5 for a discussion of statewide reporting versus this study's sample.

ELL sample by English proficiency, first language characteristics, and late-entry ELL status to explore the persistence and nature of the achievement gap for these subgroups.

**English proficiency impact.** Standardized assessments written for and normed on native English speakers become de facto dual assessments for content knowledge and academic English proficiency (Abedi, 2002, 2008b; Abedi & Gándara, 2006; Solorzano, 2008). The sample was disaggregated by English proficiency (MEPA level) to explore the impact of English proficiency on MCAS performance and whether it was a source of construct-irrelevant variance. Table 4.21 and Figure 4.4 summarize the measures of central tendency Biology MCAS performance by English proficiency levels. The data showed differential performance: As English proficiency levels increased, the mean score and percentage of ELLs who passed the Biology MCAS also increased. This was not unexpected since research has shown that English proficiency impacts ELL achievement on standardized tests (Solorzano, 2008) and is a source of differential item functioning (see Martiniello, 2008, p. 363).

Table 4.21

*Summary of ELL June 2012 Biology MCAS Performance by English Proficiency*

<u>English Proficiency</u>	Mean (SD)	Fail (%)	Needs Improvement (%)	Proficient (%)	Advanced (%)
Level 1 (n = 276)	15.47 <sub>a</sub> (6.75)	81.2	14.9	4.0	0.0
Level 2 (n = 441)	16.32 <sub>a</sub> (6.21)	76.2	21.5	2.3	0.0
Level 3 (n = 1,441)	20.21 <sub>b</sub> (7.37)	53.4	39.4	6.9	0.3
Level 4 (n = 603)	25.62 <sub>c</sub> (9.30)	28.2	50.7	18.4	2.7
Level 5 (n = 554)	31.21 <sub>d</sub> (10.02)	12.3	44.4	37.5	5.8

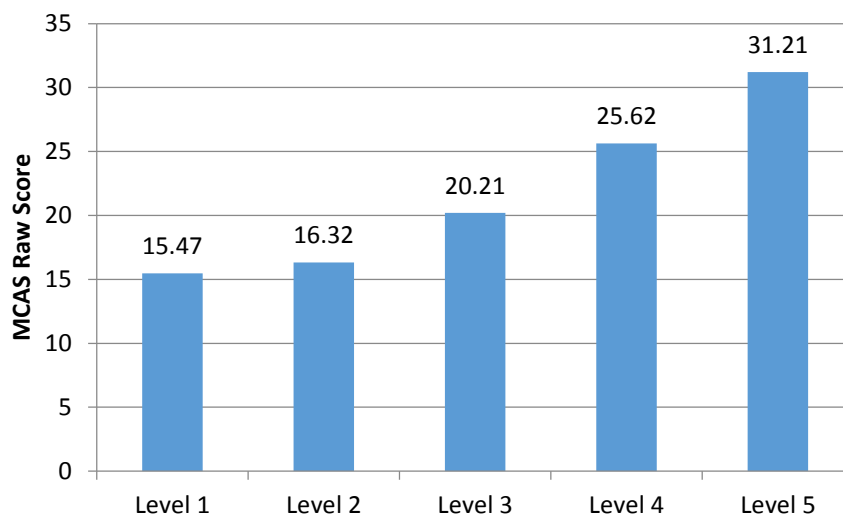
One-way ANOVA:  $F(4, 3310) = 328.12, p < .001$

Results of Scheffé post hoc analyses using paired comparisons are shown using subscripts (a, b, c, d).

Means with the same subscript are not significantly different while means with different subscripts are significantly different from one another at the  $p < .05$  level.

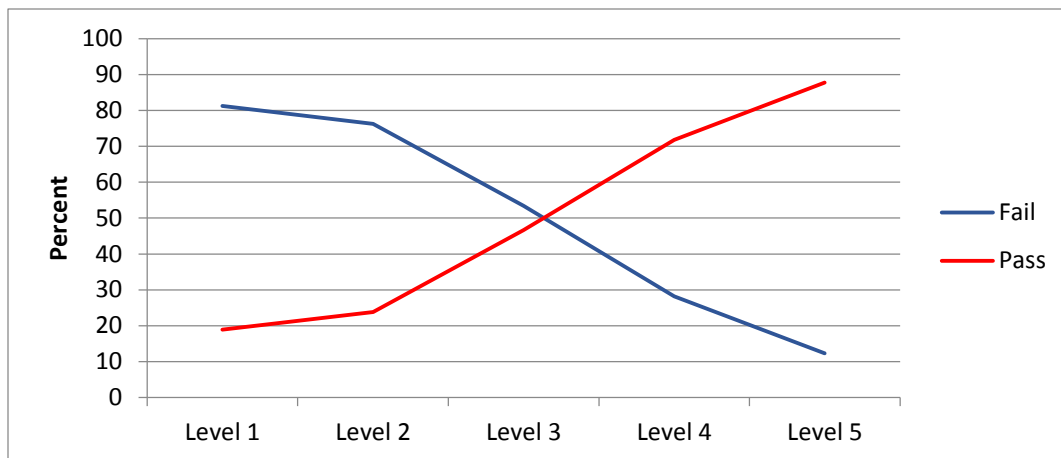
This study explored at what point in the English acquisition process ELLs started to demonstrate biology content knowledge. The mean scores for ELLs at MEPA Levels 1, 2, and 3 were below the passing threshold score of 22, and the majority of ELLs at these English proficiency levels failed: Level 1 (81.2%), Level 2 (76.2%), and Level 3 (53.4%).<sup>87</sup>

<sup>87</sup> Even though the majority of ELLs at MEPA Levels 1 to 3 failed, a small percent scored Proficient, and 0.3% of ELLs at MEPA Level 3 (n = 4) scored at the Advanced performance level. The threshold score for the Proficient performance level was 33, and the threshold score for the Advanced performance level was 48.



*Figure 4.4. Mean MCAS Score across English Proficiency Levels. As English proficiency increased, the mean MCAS score increased. A score of 22 points was the threshold for passing at the Needs Improvement level.*

Moving from MEPA Level 3 to MEPA Level 4 appeared to be a turning point for ELL performance (see Figure 4.5). In contrast to MEPA Levels 1 to 3, MEPA Levels 4 and 5 had a mean score above passing, and the majority (71.8% and 87.7%, respectively) passed. The passing rate gap almost closed for MEPA Level 5 ELLs when compared to the 91% passing rate for the 2014 Cohort (MA DESE, 2012i). Although the majority of ELLs at these higher English proficiency levels passed, the mean MCAS score for both Level 4 and Level 5 still fell in the Needs Improvement performance level ( $M = 25.62$  and  $M = 31.21$ , respectively); however, the mean score for MEPA Level 5 approached the Proficient threshold score of 33 points.

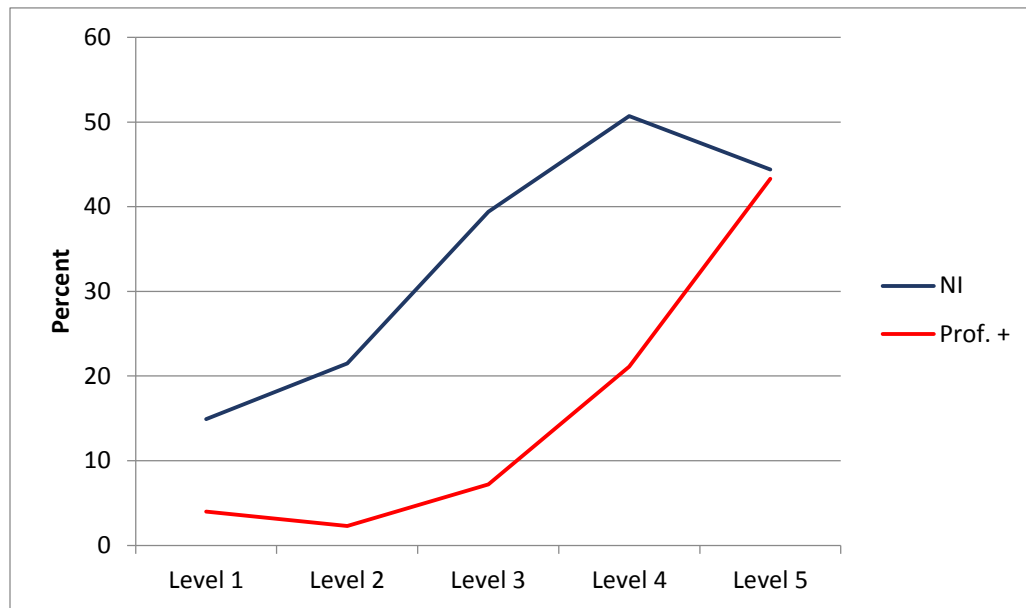


*Figure 4.5.* Percent of ELLs Failing or Passing the June 2012 Biology MCAS by English Proficiency Levels. MEPA Level 3 appeared to be the turning point for ELLs to pass the June 2012 Biology MCAS.

The mean scores indicated that the performance gap persisted at these higher English proficiency levels; however, performance level frequencies indicated that the gap appeared to narrow as English proficiency increased. Although only 7.2% of MEPA Level 3 scored Proficient or higher, this nearly tripled to 21.1% for MEPA Level 4 and was over seven times greater at 43.3% for MEPA Level 5. The data indicated that attaining MEPA Level 4 proficiency, where an ELL “reads and understands most grade-level texts, including academic vocabulary and most grade-level features of written English” (MA DESE, 2012f, p. 6), was not only where the majority began to pass the Biology MCAS but also where ELLs began to demonstrate content proficiency on an instrument written and normed for English proficient test-takers (see Figures 4.5 and 4.6). Level 4 ELLs are generally proficient with basic grade-level academic language, but they have not yet attained native-like fluency (see Appendix B). Therefore, it was not surprising that even though 71.8% passed, the achievement gap persisted, though it had



narrowed. At MEPA Level 5, where ELLs approach reclassification as English proficient, the achievement gap narrowed further, and nearly half (43.3%,  $n = 240$ ) of the passing ELLs scored Proficient or higher (Figure 4.6).



*Figure 4.6.* Percent at Performance Levels by English Proficiency for ELLs Who Passed the June 2012 Biology. MEPA Level 5 appeared to be point where ELL passing at Proficient or above approached Needs Improvement.

One-way ANOVA and Scheffé post hoc analyses confirmed that, except for between MEPA Level 1 and Level 2, there was a statistically significant increase in mean MCAS score between adjacent MEPA levels. These results are summarized in Tables 4.21 and 4.22. The results indicated that the Biology MCAS was equally inaccessible for ELLs who had low levels of English and lacked academic language (MEPA Levels 1 and 2). The mean MCAS score increase was approximately four points between Level 2 and Level 3, and approximately five points between Level 3 and 4 and between Level 4 and 5. The mean MCAS score increase from MEPA Levels 3 to 5 was 11.0 points. Although

this may not seem large, it was enough to move a performance level from Needs Improvement to Proficient.<sup>88</sup> This raises the question of whether the impact of English proficiency, especially academic language, obfuscated content knowledge for some Level 3 and Level 4 ELLs who may have been proficient in biology. The data confirmed earlier studies that English proficiency level appeared to be a source of construct-irrelevant variance (Abedi, 2002; Abedi & Gándara, 2006; Leighton & Gokierto, 2005).

Table 4.22  
*Summary of Scheffé Post Hoc Analysis on MCAS Score across English Proficiency Levels*

(I) MEPA Level	(J) MEPA Level	Mean Difference (I-J)	p
Level 1	Level 2	-.85	> .05
	Level 3	-4.75	< .001
	Level 4	-10.15	< .001
	Level 5	-15.74	< .001
Level 2	Level 1	.85	> .05
	Level 3	-3.90	< .001
	Level 4	-9.30	< .001
	Level 5	-14.89	< .001
Level 3	Level 1	4.48	< .001
	Level 2	3.90	< .001
	Level 4	-5.40	< .001
	Level 5	-11.00	< .001
Level 4	Level 1	10.15	< .001
	Level 2	9.30	< .001
	Level 3	5.40	< .001
	Level 5	-5.60	< .001
Level 5	Level 1	15.74	< .001
	Level 2	14.89	< .001
	Level 3	11.00	< .001
	Level 4	5.60	< .001

<sup>88</sup> Needs Improvement was 22 to 32 points, and the threshold score for proficient was 33 points.

This study went beyond previous studies and quantified the ELL performance variance attributed to differences in English proficiency levels. This was possible because all the ELLs in the sample had been assessed by the same English proficiency instrument (the MEPA) less than three months prior to the June 2012 Biology MCAS.<sup>89</sup> A scatterplot of MCAS score against MEPA score showed the emergence of a strong, positive linear relationship around MEPA Level 3, the level where ELLs begin to use academic language (Appendix B).<sup>90, 91</sup> A simple linear regression analysis tested to what extent the independent variable English proficiency significantly predicted Biology MCAS performance. The results of the linear regression analysis are summarized in Table 4.23. The model emerged as statistically significant,  $F(1, 3313) = 1326.8, p < .001$ , with an  $R^2$  value of .286, accounting for approximately 29% of the variation in Biology MCAS performance among ELLs. English proficiency was a strong predictor of Biology MCAS score with  $B = .25$ , which indicated that for each additional 4-point increase in MEPA score, the Biology MCAS score would increase by one point:  $\text{Biology MCAS Score} = .25 (\text{MEPA score}) - 95.9$ .<sup>92</sup>

To further explore the relationship between English proficiency and Biology MCAS performance, the simple regression analysis was repeated for the subgroup of ELLs who had a MEPA score of 464 and above (MEPA Levels 3 to 5). The results of the regression analysis are summarized in Table 4.29. The model emerged as statistically

---

<sup>89</sup> The 2012 testing window for the paper and pencil administration of the R/W MEPA sections was March 5 to March, 16, 2012.

<sup>90</sup> Appendix I contains the scatterplot of Biology MCAS raw score against MEPA scaled score.

<sup>91</sup> Appendix B contains the MEPA performance descriptors. MEPA Level 3 began at a MEPA scaled score of 464.

<sup>92</sup> The point intervals between MEPA levels were not equidistant. MEPA Level 1 ranged from a scaled score of 400 to 449; Level 2 from 450 to 463; Level 3 from 464 to 488; Level 4 from 489 to 499; and Level 5 from 500 to 550.

significant,  $F(1, 1596) = 881.0$ ,  $p < .001$ , with an  $R^2$  value of .253, accounting for approximately 25% of the variation in Biology MCAS performance among MEPA Levels 3 to 5 ELLs. English proficiency remained a strong predictor of Biology MCAS score with  $B = .33$ , which indicated that for each additional 3-point increase in MEPA score, the Biology MCAS score would increase by one point: Biology MCAS Score =  $.33(\text{MEPA score}) - 137.73$ . Although English proficiency explained approximately 4% less of the variance among MEPA Levels 3 to 5 than for the entire sample, the B value indicated that each point increase in MEPA score had a slightly stronger impact for ELLs who were developing academic language proficiency.

Table 4.23

*Summary of Linear Regression between English Proficiency and Biology MCAS Score*

	<u>N</u>	<u>Mean</u>	<u>SD</u>	<u><math>\beta</math></u>	<u>t</u>
MCAS Score	3,315	22.12	9.53	.54***	-29.57***
English proficiency	3,315	480.10	20.73		
$R^2$	.286				
F	1326.8***				
Biology MCAS Score = $.25(\text{MEPA score}) - 95.9$					
<u>MEPA Levels 3 to 5</u>					
MCAS Score	2,598	23.81	9.55	.50***	-25.30***
English proficiency	2,598	488.20	14.53		
$R^2$	.253				
F	881.00***				
Biology MCAS Score = $.33(\text{MEPA score}) - 137.73$					

Note: \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

The ELL achievement gap is a gap between test-takers who are English proficient and those who are not, and this gap has been well-identified in the literature. Thus, it was not surprising that that English proficiency emerged as a strong predictor of Biology MCAS score. What did emerge from this study was the quantification of the impact

within the ELL sample, and both second language acquisition and cognitive load theories support the findings.

The ability to read and understand the item is a factor in performance. If the content item is incomprehensible input, then language proficiency will interfere with the content construct under assessment. As English proficiency increased, the items became more comprehensible, and ELL performance increased. The exception was that there was no statistically significant increase in performance between MEPA Levels 1 and 2; however, this is consistent with the constructs of BICS and CALP. English language learners at ELL Levels 1 and 2 communicate using BICS, but it is only at MEPA Level 3 that ELLs start using academic language (CALP), which continues to develop through MEPA Levels 4 and 5. The linearity for English proficiency impact that emerged at MEPA Level 3 is consistent with earlier studies on the impact of academic vocabulary on standardized assessment performance (see Abedi & Lord, 2001; Lawrence et al., 2010; Martiniello, 2008).

The findings are also consistent with cognitive load theory, which predicts that cognitive load decreases as English proficiency increases. As ELLs gain automaticity in reading, there is reduced demand on working memory resources, which reduces the cognitive load of the assessment's language (see C. H. Lee & Kalyuga, 2011). Accordingly, this frees up working memory resources for the construct under assessment (i.e., biology content knowledge). Increased academic language acquisition, especially low-frequency words, would increase reading automaticity as ELLs progressed from MEPA Levels 3 to 5. As expected, the data indicated statistically and practically significant differences in Biology MCAS performance among MEPA Levels 3 to 5.

The American Educational Research Association (2000) recommends that an assessment should not be used with a student who does not understand “the language of the test” (as cited in Solorzano, 2008, p. 262), and Abedi (2008b) argued that ELLs should participate in standardized content assessments only when English proficiency assessments show that language proficiency matches the language of the content assessment. This study explored at what point this happens for ELLs in Massachusetts. The findings indicated that MEPA Level 3 was the cusp of language proficiency where the failing and passing percentages were approximately equal. At MEPA Level 3, however, academic language acquisition was nascent, and the findings suggested that although ELLs were beginning to pass, few could demonstrate proficiency in biology content. The findings further suggested that the achievement gap narrowed as English proficiency increased. By MEPA Level 4, the overwhelming majority of ELLs passed the Biology MCAS with approximately one-fifth attaining at least a Proficient performance level. MEPA Level 5 ELLs were closing the gap with non-ELLs for passing rate, and although there was still a gap for Proficient and above, almost half of the Level 5 ELLs were at least at the Proficient performance level. This study suggested that although the June 2012 Biology MCAS started to become accessible to ELLs at MEPA Level 3, it was at MEPA Level 4 where ELLs could be expected to pass the Biology MCAS, and MEPA Level 5 where ELLs started to demonstrate content proficiency.

**First language family impact.** “Culture and language influence cognitive processes. Consequently, they may affect a person’s performance in cognitive tests” (Schaap & Vermeulen, 2008, as cited in Schaap, 2011, p. 138). The common underlying proficiency model postulates that L1 linguistic skills and knowledge can transfer to the

L2 (Cummins, 2008), and as discussed in Chapter 2, previous studies have shown that the first language can impact performance. Syntactic characteristics of the first language can impact sentence interpretation, which impacts reading comprehension and cognitive load (McWhinney et al., 1984, as cited in Chitiri & Willows, 1994, p. 314). Schaap (2011) found that differential item functioning was not the same across first language groups on a non-verbal instrument, and Martiniello (2008) explored differential item functioning for the 2003 Grade 4 Mathematics MCAS for Spanish-speaking ELLs. Martiniello (2008) also explored the impact of Spanish-English cognates and called for further research on other languages.

This study disaggregated the sample into two L1 language family groups to explore the impact of a Latinate or non-Latinate L1 on Biology MCAS performance. Frequencies for MCAS performance levels for the two subgroups are summarized in Table 4.24 and Figure 4.7. The majority (51%) of ELLs with a Latinate L1 failed the Biology MCAS. In contrast, the majority (62.6%) of ELLs with a non-Latinate L1 passed the Biology MCAS. The Needs Improvement level had similar percentages for both subgroups, but differences emerged at the Proficient and Advanced performance levels. Only 10.7% of Latinate L1 ELLs scored Proficient or higher compared to 26.6% of non-Latinate L1 ELLs scoring Proficient or higher, including 4% who scored Advanced (only 0.7% of Latinate L1 ELLs scored Advanced).

Table 4.24  
*ELL Performance Levels by First Language Family*

MCAS Performance Level	Llatinate L1 (n = 2,420)		Non-Llatinate L1 (n = 895)	
	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>
Fail	1,233	51.0	334	37.3
Needs Improvement	929	38.4	327	36.5
Proficient	242	10.0	198	22.1
Advanced	16	0.7	36	4.0

Table 4.24 summarizes measures of central tendency. The data showed that as a subgroup, ELLs with a non-Llatinate L1 performed better than ELLs with Llatinate L1. Llatinate L1 ELLs (n = 2,420) had a mean score of 20.96 (SD = 8.54), which was below the threshold passing score of 22. In comparison, non-Llatinate L1 ELLs (n = 895) had a mean score of 25.26 (SD = 11.20), which was passing at the Needs Improvement performance level. The data suggested a performance gap between ELLs based on whether their L1 was Llatinate or non-Llatinate, with non-Llatinate L1 ELLs performing better.

Table 4.25  
*ELL Scores on the June 2012 Biology MCAS Score by Language Family*

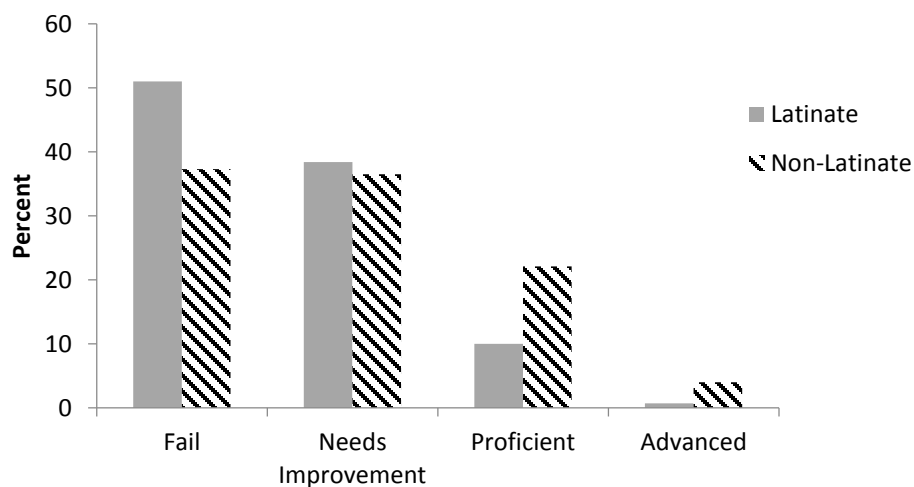
<u>L1</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>t</u>	Cohen's <u>d</u>	Partial <u>eta</u> <sup>2</sup>
Llatinate	2,420	20.96	8.54	-10.45***	.43	.04
Non-Llatinate	895	25.26	11.20			

\*p < .05 \*\*p < .01 \*\*\*p < .001.

An independent samples t-test confirmed that L1 language family (Llatinate/non-Llatinate) had a statistically significant impact on Biology MCAS score that favored ELLs



with a non-Latinate L1,  $t(1298) = -10.4$ ,  $p < .001$ . Tables 4.25 and 4.34 summarize the results of this analysis. A Cohen's  $d$  value of 0.43 indicated a small to medium effect size, and a univariate analysis of variance determined that L1 language family accounted for 4% of the difference between subgroups. Although 4% of the variance appears to have minor practical significance, in the context of this high-stakes assessment, the 4-point difference in mean score was the difference between failing and passing at the Needs Improvement performance level—or, in other words, meeting one of the requirements for a high school diploma.



*Figure 4.7. MCAS Performance Level Percentages by First Language Family.* Latinate L1 ELLs had a greater percentage at the Fail level, and non-Latinate L1 ELLs had greater percentages at the Proficient and Advanced levels.

The finding that ELLs with a non-Latinate L1 scored higher than ELLs with a Latinate L1 was unexpected. The common underlying proficiency model suggests that an ELL with a Latinate L1 would be able to draw on L1 cognates as a reading comprehension strategy (see Martiniello, 2008) and thus have a linguistic advantage over

ELLs with a non-Latinate L1.<sup>93</sup> For example, item 3 contained the phrase “produces a toxin” (Appendix A). If an ELL did not know “produce” or “toxin,” both of which are not content-specific, Latinate L1 cognates could aid item comprehension: (1) Spanish: *producir una toxina*, (2) Portuguese: *produzir uma toxina*, (3) French: *produire une toxine*, and (4) Haitian Creole: *pwodwi yon toksin*. Many scientific terms have Latin (or Greek) roots and affixes, and they are often similar across Latinate languages. For example, the Tier III content-specific words “photosynthesis” (item 1) and “amino acid” (item 7) have the following Latinate L1 cognates: (1) Spanish: *la fotosíntesis* and *aminoácidos*, (2) Portuguese: *fotossíntese* and *aminoácidos*, (3) French: *photosynthèse* and *acides aminés*, and (4) Haitian Creole: *fotosentèz* and *asid amine*.<sup>94,95</sup> The linguistic similarities in scientific terms, however, would only be an advantage if there was L1 content knowledge.

One explanation is that there was no underlying common proficiency for Latinate L1 ELLs. That is, the Latinate L1 ELLs did not know the L1 cognates for Tier II words (non-content-specific academic language), Tier III words (content-specific academic language), or both. This could be due to interruptions in or lack of formal schooling in the L1. With no L1 academic discourse or prior L1 content knowledge, having a Latinate L1 would not confer an advantage. Even if there was knowledge of L1 cognates, reading comprehension strategies need explicit instruction and opportunities to practice (Echevarria, Vogt, & Short, 2013). As discussed in Chapter 3, a limitation of the

---

<sup>93</sup> Using L1 cognates as a reading comprehension strategy presupposes that the L1 cognate is known.

<sup>94</sup> Other examples of scientific terms with cognates in Latinate languages include “gamete” and “mitochondria”: (1) Spanish: *gameto* and *mitocondrias*, (2) Portuguese: *gameta* and *mitocôndrias*, (3) French: *gamète* and *mitochondrie*, and (4) Haitian Creole: *gamèt* and *mitokondri*.

<sup>95</sup> These scientific terms were translated using [www.translate.google.com](http://www.translate.google.com).

secondary data was that they did not include data on whether ELLs had access to ESL-certified and Biology-certified teachers who sheltered the content as well as developed scientific discourse.

The converse finding that non-Latinate L1 ELLs scored higher and had more representation at the Proficient and Advanced performance levels was also unexpected; however, it is not necessarily in opposition to the common underlying proficiency model. The sample demographics (Table 4.8) showed that approximately 10% more non-Latinate L1 ELLs (74.5%) were late-entry than Latinate L1 ELLs (64.1%). A possible explanation is that the late-entry ELLs may have studied the biology constructs under assessment in their L1, and with the accommodations of word-to-word dictionaries and content glossaries, the underlying L1 content knowledge was an advantage that mitigated the lack of L1-to-English cognates. It also raises the question of whether lack of cognates impacted non-Latinate L1 ELL performance by depressing their demonstration of biology content knowledge. As English proficiency increases and nears native-speaker equivalencies, the need for cognates as a reading comprehension strategy diminishes. First language demographics (Table 4.8) showed small differences for the subgroups at MEPA Level 1 and MEPA Level 5. There were twice as many Latinate L1 ELLs at MEPA Level 1 (9.6%) than non-Latinate ELLs (4.8%), and 5.3% fewer Latinate L1 ELLs (15.3%) at MEPA Level 5 than non-Latinate L1 ELLs (20.6.3%). Another possibility is that some non-Latinate languages may adopt or adapt slightly the English version of scientific terms; this would level the cognate advantage with respect to Tier III words but not Tier II words. As discussed in Chapter 5, this is an area for further study.

Based on the finding discussed previously that a linear relationship between English proficiency and Biology MCAS performance emerged at MEPA Level 3, the Latinate L1 and non-Latinate L1 subgroups were further disaggregated into two English proficiency subgroups to explore further the impact of L1 language family: (1) MEPA Levels 1 and 2, and (2) MEPA Levels 3 to 5. MCAS performance levels for the disaggregated subgroups are summarized in Table 4.26 and Figure 4.8. The patterns were the same as for disaggregation by English proficiency level (without regard to L1 language family) discussed previously. Performance at all performance levels was indistinguishable for MEPA Levels 1 and 2 regardless of L1 language family.

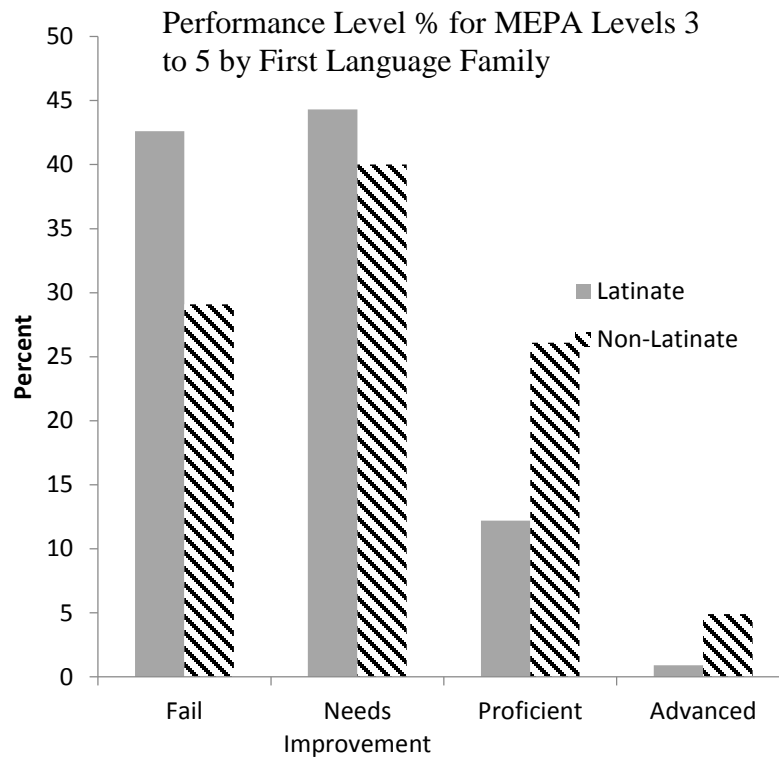
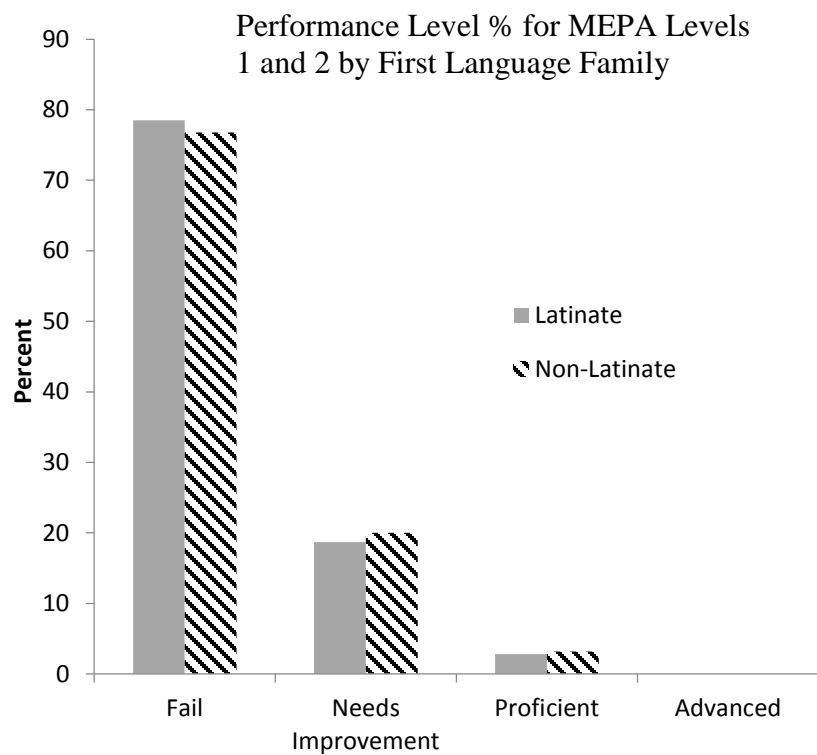
Differences emerged for the subgroups at MEPA Levels 3 to 5 for all performance levels except Needs Improvement. Fewer non-Latinate L1 ELLs failed the Biology MCAS (29.1% non-Latinate compared to 42.6% Latinate). The non-Latinate L1 ELLs at the Proficient performance level were more than twice that of Latinate L1 ELLs (26.1% and 12.2%, respectively). The greatest difference occurred at the Advanced performance level, which had over five times as many non-Latinate ELLs (4.9%) than Latinate L1 ELLs (0.9%). The data indicated that having reached MEPA Levels 3 and above, non-Latinate L1 ELLs not only had greater passing percentages (69.9% compared to 57.4% Latinate L1) but also greater percentages at the Proficient or higher performance levels (31% non-Latinate L1 compared to 13.1% Latinate L1).

Table 4.26

*Summary of MCAS Performance Level by L1 Language Family and English Proficiency Levels*

	MEPA Levels 1 -2		MEPA Levels 3-5	
	Latinate % (n = 562)	Non-Latinate % (n = 155)	Latinate % (n = 1858)	Non-Latinate % (n = 740)
Fail	78.5	76.8	42.6	29.1
Needs Improvement	18.7	20	44.3	40.0
Proficient	2.8	3.2	12.2	26.1
Advanced	0	0	0.9	4.9

Independent samples t-tests confirmed that L1 family had no statistical significance for ELLs at MEPA Levels 1 and 2,  $t(715) = -.47$ ,  $p < .05$ , but it had a statistically significant impact on mean Biology MCAS score for ELLs at MEPA Levels 3 and above,  $t(1112) = -10.4$ ,  $p < .001$ . Table 4.27 summarizes the results. A Cohen's d value of 0.48 indicated that the effect size approached medium for ELLs at MEPA Levels 3 and above, and a univariate analysis of variance determined that L1 language family contributed 4.9% of the variance in performance between Latinate L1 and non-Latinate L1 ELLs. The data suggested that once ELLs acquired enough English proficiency to begin using academic language, ELLs with a non-Latinate L1 performed better on the Biology MCAS. This finding was unexpected, and as discussed in Chapter 5, further investigation is needed into what appears to be an achievement gap between ELLs with a Latinate L1 (Spanish, Haitian Creole, Portuguese, French, and Italian) and ELLs with a non-Latinate L1 after they reach English proficiency levels where there is a positive linear relationship with Biology MCAS performance.



*Figure 4.8.* MCAS Performance Level Percentages by English Proficiency and L1 Language Family. At MEPA Levels 1 and 2, ELLs with a Latinate L1 and ELLs with a non-Latinate L1 have similar performance. At MEPA Levels 3 and above, ELLs with a non-Latinate L1 have greater percentages at the Proficient and Advanced levels than ELLs with a Latinate L1.

Table 4.27

*Impact of First Language Characteristics and Late-Entry ELL Status on Biology MCAS Score*

	Linate L1			Non-Linate L1			t	Cohen's d	Partial eta <sup>2</sup>
	n	M	SD	n	M	SD			
MCAS score	2,420	20.96	8.54	895	25.26	11.20	-10.4***	.43 <sup>a</sup>	.04
MEPA Level 3+ (n = 2,598)	1,858	22.48	8.55	740	27.16	11.00	-10.4***	.48 <sup>a</sup>	.049
MEPA Levels 1-2 (n = 717)	562	15.93	6.31	155	16.21	6.86	-.47		
	Alphabetic L1			Non-alphabetic L1			t	Cohen's d	Partial eta <sup>2</sup>
	n	M	SD	n	M	SD			
MCAS score	2,724	21.45	8.94	591	25.20	11.39	-7.5***	.37 <sup>b</sup>	.02
MEPA Level 3+ (n = 2598)	2,124	22.98	8.95	474	27.53	11.13	-8.3***	.45 <sup>b</sup>	.034
MEPA Levels 1-2 (n = 717)	600	16.04	6.42	117	15.77	6.48	.40		
	Late-entry ELL			Not Late-entry ELL			t	Cohen's d	Partial eta <sup>2</sup>
	n	M	SD	n	M	SD			
MCAS score	2,218	22.15	9.98	1,097	22.06	8.54	.26		
MEPA Level 3+ (n = 2598)	1,551	24.73	10.11	1,047	22.46	8.48	5.98***	0.24 <sup>c</sup>	.014
MEPA Levels 1-2 (n = 717)	667	16.16	6.52	50	13.78	4.68	2.53*	0.42 <sup>c</sup>	.009

\*p &lt; .05 \*\*p &lt; .01 \*\*\*p &lt; .001.

<sup>a</sup> Favoring non-Linate L1; <sup>b</sup> favoring non-alphabetic L1; <sup>c</sup> favoring late-entry ELLs.<sup>d</sup> Levene's test indicated equal variances not assumed.

**First language orthography impact.** As discussed in Chapter 2, the distance between L1 and L2 orthographies can be a factor in the second language acquisition process. For ELLs who are already familiar with the Roman alphabet, language input would be more comprehensible relative to ELLs who are still acquiring automaticity in sound-letter correspondence. For alphabetic L1 ELLs who have L1 literacy, there would be a positive L1 to L2 transfer of a common underlying proficiency (see Birch, 2002; Ellis, 2003, Chapter 6). Even for alphabetic L1 ELLs who have minimal or emerging L1 literacy, their L1 linguistic environment created exposure to the alphabet.<sup>96</sup> An alphabetic L1 also impacts cognitive processes. “Underlying cognitive resources are tapped differentially, to the degree demanded by the orthographic or linguistic characteristics of L1 and L2” (Geva, 1999, as cited in Birch, 2002, p. 38). In theory, ELLs with an alphabetic L1 should have a lower linguistic cognitive load compared to ELLs with a non-alphabetic L1, especially at the lower levels of English proficiency; this presupposes L1 literacy. A lower cognitive load would increase the availability of working memory for content and attention to other item attributes.

Also as discussed earlier, studies have indicated that first language can impact performance, and orthography is a linguistic characteristic. This study disaggregated the ELL sample by L1 orthography to explore the impact of an alphabetic L1 on MCAS performance. Frequencies for MCAS performance levels for the two subgroups are summarized in Table 4.28 and Figure 4.9. Both subgroups had the greatest number at the Fail performance level: 49.2% of alphabetic L1 ELLs and 38.4% of non-alphabetic L1 ELLs. Both subgroups had similar percentages at the Needs Improvement level, but

---

<sup>96</sup> For example, an alphabetic L1 environment creates exposure through product labels and signage.



differences emerged at the Proficient and Advanced performance levels. The percent at the Proficient level was nearly double for ELLs with a non-alphabetic L1 (22%) compared to ELLs with an alphabetic L1 (11.4%). The gap widened at the Advanced level where 4.4% of the non-alphabetic L1 ELLs scored compared to 1% of the alphabetic L1 ELLs.

Table 4.28  
*ELL Performance Levels by L1 Orthography*

MCAS Performance Level	Alphabetic L1 (n = 2,724)		Non-Alphabetic L1 (n = 591)	
	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>
Fail	1,340	49.2	227	38.4
Needs Improvement	1,048	38.5	208	35.2
Proficient	310	11.4	130	22.0
Advanced	26	1.0	26	4.4

Tables 4.27 and 4.29 summarize measures of central tendency. Alphabetic L1 ELLs (n = 2,724) had a mean score of 21.45 (SD = 8.94), which was below the threshold passing score of 22. In comparison, ELLs with a non-alphabetic L1 (n = 591) had a mean score of 25.20 (SD = 11.29), which was passing at the Needs Improvement performance level. The data indicated a performance gap between ELLs with an alphabetic L1 and ELLs with a non-alphabetic L1.

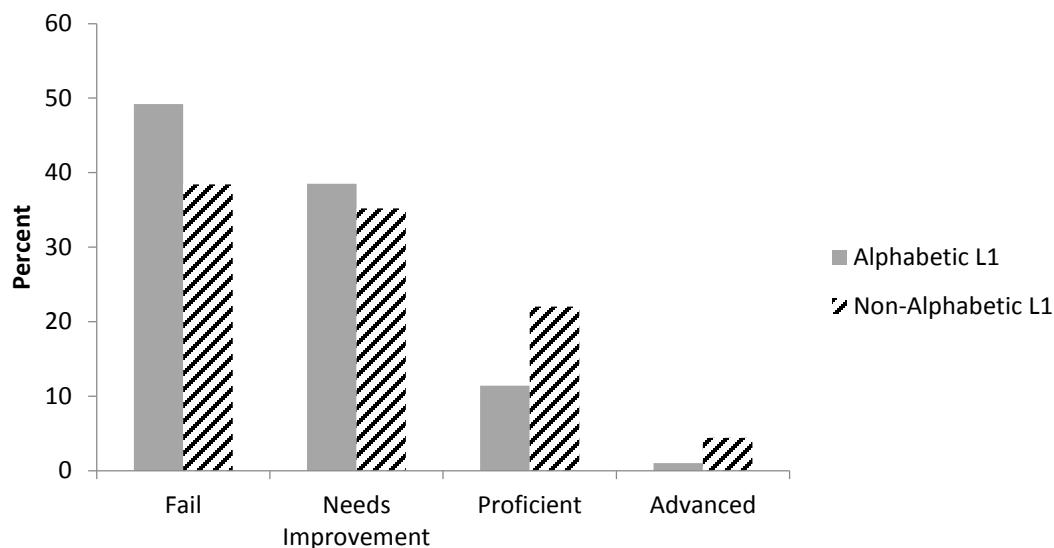
Table 4.29

*ELL Scores on the June 2012 Biology MCAS by L1 Orthography*

<u>L1 Orthography</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>t</u>	Cohen's <u>d</u>	Partial <u>eta</u> <sup>2</sup>
Alphabetic	2,724	21.45	8.94	-7.5***	.37	.02
Non-alphabetic	591	25.20	11.39			

\*p &lt; .05 \*\*p &lt; .01 \*\*\*p &lt; .001.

An independent samples t-test confirmed that L1 orthography (alphabetic/non-alphabetic) had a statistically significant impact on mean Biology MCAS score that favored ELLs with a non-alphabetic L1 ( $t(755) = -7.52, p < .001$ ). Tables 4.27 and 4.29 summarize the results. A Cohen's d value of 0.37 indicated a small effect size, and a univariate analysis determined that 2% of the variance between the subgroups was attributed to L1 orthography. Although the impact of L1 orthography was statistically significant, its practical significance was minor. Notwithstanding the small practical impact, this finding was unexpected. The common underlying proficiency model (Cummins, 2000), the comprehensible input hypothesis (see Finegan, 2004), and cognitive load (see Birch, 2002) suggest that ELLs with an alphabetic L1 might perform better than ELLs with a non-alphabetic L1.



*Figure 4.9. MCAS Performance Level Percentages by First Language Orthography.* ELLs with an alphabetic L1 had a greater percentage at the Fail level, and non-alphabetic L1 ELLs had greater percentages at the Proficient and Advanced levels.

Based on the findings that a linear relationship between English proficiency and Biology MCAS performance emerged at MEPA Level 3, the L1 orthography subgroups were further disaggregated into two English proficiency subgroups: (1) MEPA Levels 1 and 2, and (2) MEPA Levels 3 to 5. Table 4.27 summarizes measures of central tendency, and Table 4.30 and Figure 4.10 summarize MCAS performance levels. At MEPA Levels 1 and 2, there was little difference in the passing rate between ELLs with an alphabetic L1 (22%) and non-alphabetic L1 (21.4%), and the two subgroups also had similar percentages across the performance levels. The data, however, indicated differential performance for ELLs at MEPA Levels 3 and above. Although the majority of ELLs in both subgroups passed, the percentage was higher for ELLs with a non-alphabetic L1 (72.5%) than for those with an alphabetic L1 (58.1%). Almost one-third (32.3%) of the non-alphabetic L1 ELLs scored Proficient or higher compared to 14.9% of the alphabetic

L1 ELLs, with nearly five times as many non-Latinate L1 ELLs scoring Advanced. The data suggested that at MEPA Levels 3 and above, there was an achievement gap where ELLs with a non-alphabetic L1 performed better than ELLs with an alphabetic L1.

Table 4.30  
*Summary of MCAS Performance Level by L1 Orthography and English Proficiency Levels*

MCAS Performance Level	MEPA Levels 1 -2		MEPA Levels 3-5	
	Alphabetic % (n = 600)	Non-Alphabetic % (n = 117)	Alphabetic % (n = 2124)	Non-Alphabetic % (n = 474)
Fail	78.0	78.6	41.1	28.5
Needs Improvement	19.0	18.8	44.0	39.2
Proficient	3.0	2.6	13.7	26.8
Advanced	0.0	0.0	1.2	5.5

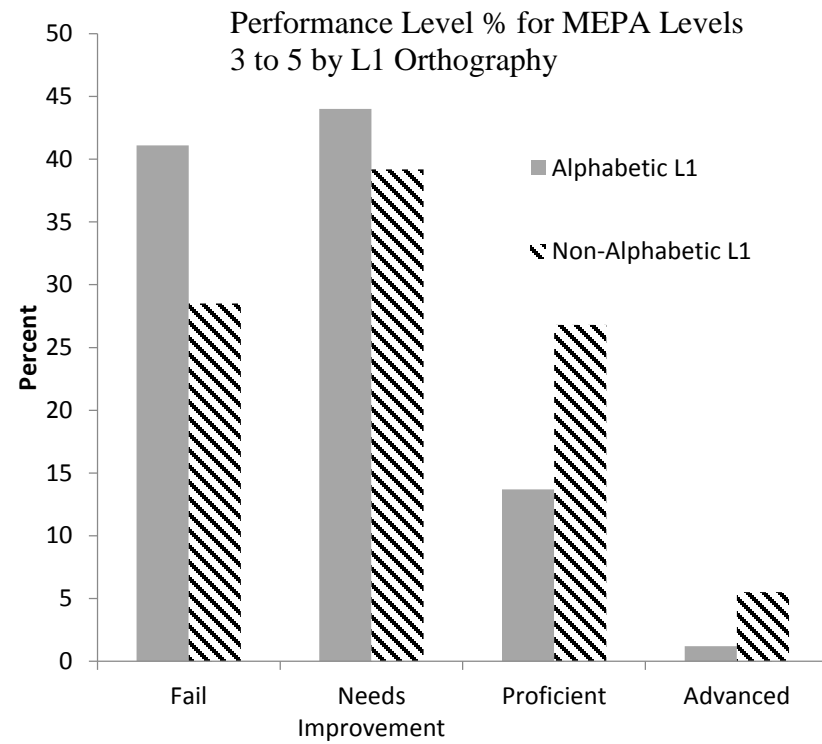
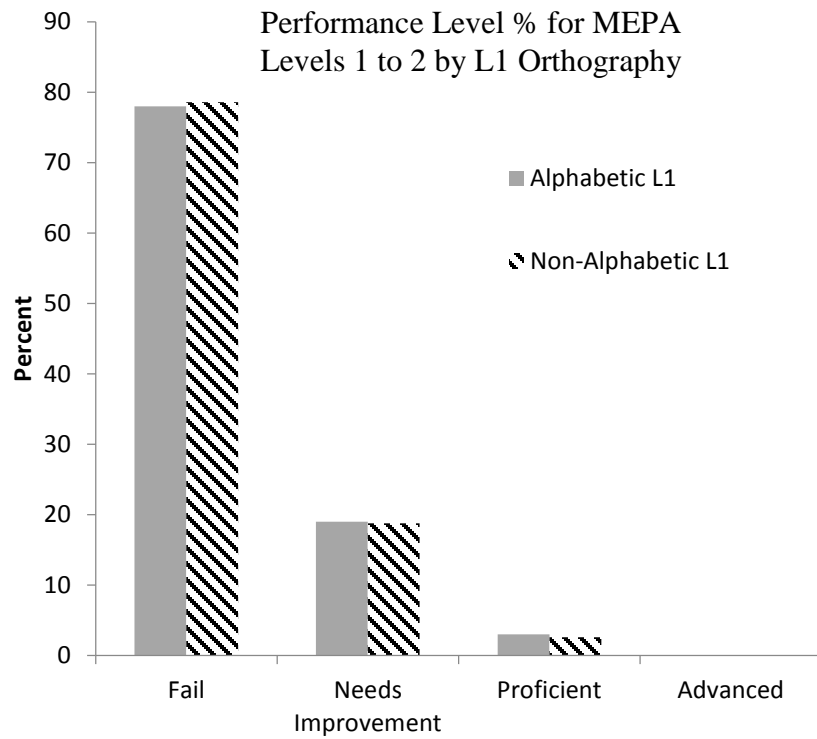
Independent samples t-tests confirmed that L1 orthography had no statistically significant impact on Biology MCAS score for ELLs at MEPA Levels 1 and 2,  $t(715) = .41$ ,  $p > .05$ , but there was a statistically significant impact for ELLs at MEPA Levels 3 and above,  $t(616) = -8.32$ ,  $p < .001$ . Table 4.27 summarizes the results. For ELLs at MEPA Levels 3 to 5, a Cohen's d value of 0.45 indicated a small to medium effect size that favored ELLs with a non-alphabetic L1, and partial eta-squared determined that L1 orthography contributed 3.4% of the variance in Biology MCAS score between the groups.

As discussed previously, this study found no statistically significant difference in overall performance between ELLs at MEPA Levels 1 and 2, as well as no statistically significant impact for L1 family for MEPA Levels 1 and 2 ELLs. Therefore, it was not

unexpected that L1 orthography had no statistically significant impact at these lower levels of English proficiency. It appeared that the low level of English proficiency and lack of academic language below MEPA Level 3 created sufficient barriers to the assessment to obscure the impact of L1 characteristics such as orthography. When English proficiency reached MEPA Level 3, the instrument became more accessible, and L1 orthography had a small impact on performance. It was unexpected, however, that the non-alphabetic L1 subgroup performed better. One hypothesis is that upon reaching intermediate and higher English proficiency levels, non-alphabetic L1 ELLs have acquired a level of alphabetic automaticity that releases cognitive resources for reading comprehension. The freed cognitive resources are now available for content demands. This study classified Chinese languages, the fourth most common first language of the sample, as non-alphabetic; however, ELLs with a Chinese first language likely had familiarity with pīnyīn, the Romanized script for Mandarin and other Chinese languages. One hypothesis is that familiarity with pīnyīn lessened the cognitive load from the L1-L2 orthographic distance between non-alphabetic and alphabetic languages.

For MEPA Levels 3 to 5, the impact of an alphabetic/non-alphabetic L1 was 3.4%, which was less than the 4.9% impact found for Latinate/non-Latinate L1. All Latinate languages are alphabetic, but non-Latinate languages can be alphabetic or non-alphabetic. Thus, ELLs with an alphabetic, non-Latinate L1 were the only difference between the Latinate/non-Latinate subgroups and alphabetic/non-alphabetic subgroups. One explanation for the lesser impact of L1 orthography is that alphabetic, non-Latinate L1 ELLs, who were in the higher performing non-Latinate L1 subgroup, were now in the same subgroup as the lower performing Latinate L1 ELLs when disaggregated by

orthography. As discussed in Chapter 5, the finding that ELLs with a non-alphabetic L1 performed better warrants future research that compares the performance of three L1 subgroups: (1) Latinate; (2) non-Latinate, alphabetic; and (3) non-Latinate, non-alphabetic.



*Figure 4.10.* MCAS Performance Level Percentages by English Proficiency and L1 Orthography. At MEPA Levels 3 and above, ELLs with a non-alphabetic L1 have greater percentages at the Proficient and Advanced levels than ELLs with an alphabetic L1.

**Late-entry ELL status impact.** As discussed in Chapter 2, there is not uniform agreement on a critical period for second language acquisition (see Long, 2007, Chapter 3). This notwithstanding, studies have shown that it takes significantly longer to acquire the register requisite for academic success (see Hakuta et al., 2000). It is generally accepted that it takes four to seven years to acquire academic language, and late-entry secondary ELLs, by definition, do not have sufficient time to do so and are “vulnerable to academic failure” (Boyson & Short, 2012, p. 5). Collier (1987a) found that among ELLs who arrived between the ages of 5 and 15 years, the 12- to 15-year-old group exhibited the most difficulty and took the longest (six to eight years) to achieve academic language parity with native speakers. This study defined late-entry ELLs as entering the United States at 12 years of age or later, and it explored whether late-entry ELLs had differential performance on the June 2012 Biology MCAS.

The ELL sample was disaggregated into late-entry and not-late-entry subgroups to explore the impact of age of entry on Biology MCAS performance. Tables 4.27 and 4.31 summarize measures of central tendency. Late-entry ELLs and not-late-entry ELLs had similar passing rates (52.2% and 53.9%, respectively) and a mean score at the passing threshold of 22 points. Both subgroups also had similar percentages at each of the performance levels, except for the Advanced level where 2.1% of late-entry ELLs and 0.5% of not-late-entry ELLs scored. An independent samples t-test confirmed that late-entry ELL status did not have a statistically significant impact on MCAS score,  $t(2510) = .26$ ,  $p > .05$ . Tables 4.27 and 4.31 summarize the results. Given that not-late-entry ELLs arrived at an earlier age, their exposure to English, both in general and in academic



settings, would be greater than that of late-entry ELLs. Thus, it was surprising that there was no statistically significant difference in performance between the groups.

Table 4.31  
*ELL MCAS Scores and Performance Levels by Late-Entry ELL Status*

	Late-entry ELL (n = 2,218)		Not-late-entry ELL (n = 1,097)		t
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	
MCAS Score	22.15	9.98	22.06	8.54	.26
MCAS Performance Level	<u>n</u>	<u>%</u>	<u>n</u>	<u>%</u>	
Fail	1,061	47.8	506	46.1	
Needs Improvement	805	36.3	451	41.1	
Proficient	305	13.8	135	12.3	
Advanced	47	2.1	5	0.5	

\*p < .05 \*\*p < .01 \*\*\*p < .001.

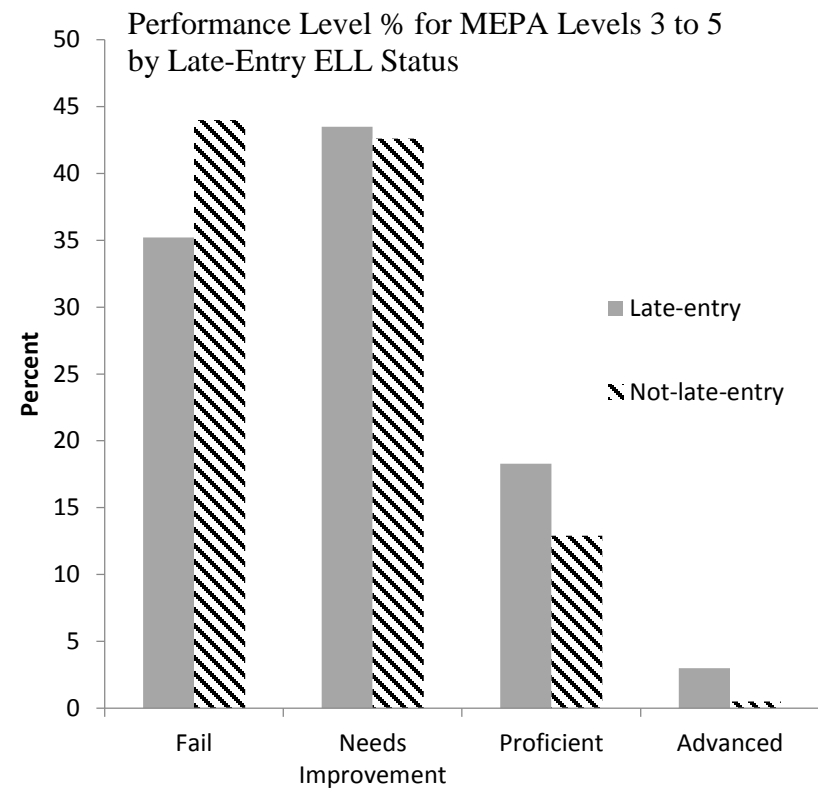
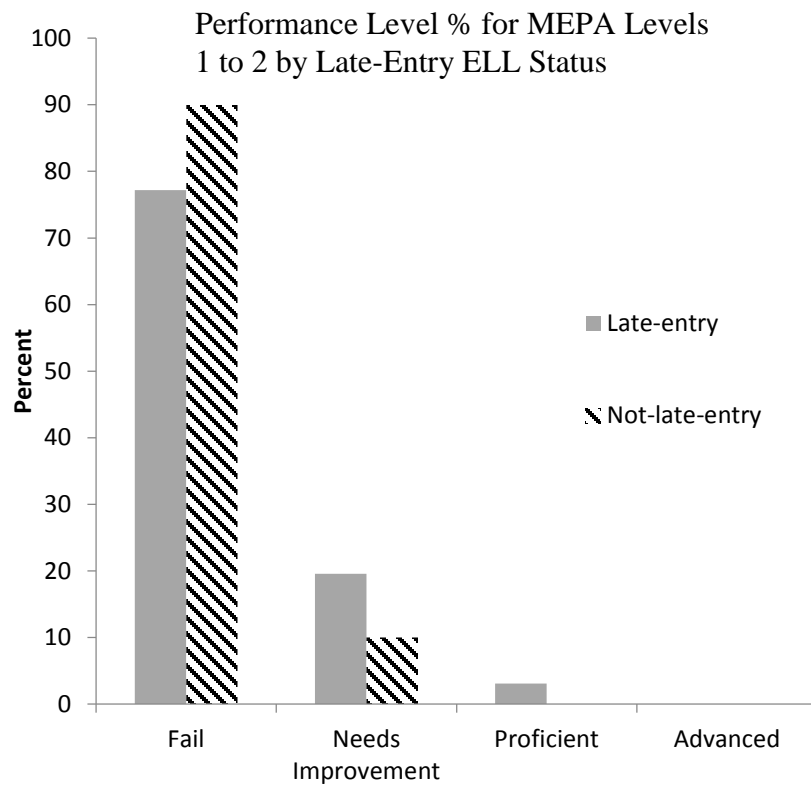
Based on the findings that English proficiency contributed 29% of the variance in Biology MCAS performance among MEPA levels, the late-entry/not-late-entry subgroups were further disaggregated into two English proficiency subgroups: (1) MEPA Levels 1 to 2, and (2) MEPA Levels 3 to 5. Table 4.27 summarizes measures of central tendency, and Table 4.32 and Figure 4.11 summarize the MCAS performance levels. Independent samples t-tests determined that late-entry ELL status had a statistically significant impact for ELLs at the low end of English proficiency (MEPA Levels 1 and 2) ( $t(715) = 2.53, p < .05$ ), as well as at the intermediate and higher levels (MEPA Levels 3 to 5) ( $t(2476) = 6.18, p < .001$ ). Table 4.27 summarizes these results. Cohen's d values and partial eta-squared for both subgroups, however, indicated that late-entry ELL status had negligible impact on performance; it contributed 1% of the variance

at MEPA Levels 1 and 2, and 1.4% of the variance at MEPA Levels 3 to 5. Although 77.2% of the MEPA Levels 1 and 2 late-entry ELLs failed, less time in the United States is the likely reason for low English proficiency, and as English proficiency increases, Biology MCAS performance should increase. The not-late-entry ELLs at MEPA Levels 1 and 2 were relatively few (n = 50); however, 90% of them failed the Biology MCAS. This subgroup appears to be at high risk for academic failure and should be studied further.

Table 4.32

*Summary of MCAS Performance Level by Late-Entry ELL Status and English Proficiency Levels*

MCAS Performance Level	MEPA Levels 1 -2		MEPA Levels 3-5	
	Late-entry % (n = 667)	Not-late-entry % (n = 50)	Late-entry % (n = 1551)	Not-late-entry % (n = 1047)
Fail	77.2	90.0	35.2	44.0
Needs Improvement	19.6	10.0	43.5	42.6
Proficient	3.1	0.0	18.3	12.9
Advanced	0.0	0.0	3.0	0.5



*Figure 4.11.* MCAS Performance Level Percentages by English Proficiency and Late-Entry ELL Status. At MEPA Levels 3 and above, late-entry ELLs have greater percentages at the Proficient and Advanced levels than not-late-entry ELLs.

**Summary of ELL MCAS performance.** The majority of ELLs (52.7%) who took the June 2012 Biology MCAS passed; however, only 14.9% passed with a performance level of Proficient or higher. For ELLs at MEPA Levels 4 and 5, the overwhelming majority (71.8% and 87.7%, respectively) passed. The data indicated that at lower levels of English proficiency (MEPA Levels 1 and 2), there was no statistically significant impact of English proficiency on Biology MCAS score. A linear relationship between English proficiency, as measured by the MEPA score, and Biology MCAS score emerged at MEPA Level 3, which is when ELLs begin to use academic language. Simple regression analyses led to two statistically significant models, which showed a stronger impact for ELLs at MEPA Levels 3 to 5:

$$\text{All MEPA Levels: Biology MCAS Score} = .25(\text{MEPA score}) - 95.9$$

$$\text{MEPA Level 3+: Biology MCAS Score} = .33(\text{MEPA score}) - 137.73$$

For Massachusetts ELLs, MEPA Level 3, where academic language use begins, appeared to be the turning point where approximately half of the ELLs in the sample passed the Biology MCAS. Progressing from Level 3 to 4 increased the passing rate to nearly three-fourths and appeared to be the point where approximately one-fifth demonstrated content proficiency. At MEPA Level 5, the achievement gap almost closed for passing the Biology MCAS, but an achievement gap still remained for scoring proficient or higher.

With respect to first language characteristics, the data indicated that ELLs with a non-Latinate first language had a statistically significant higher mean MCAS score and that the performance gap between the groups widened at the Proficient and Advanced MCAS performance levels. The data indicated a small to medium effect size that favored

ELLs with a non-Latinate L1 and contributed 4% of the difference between the groups. When the subgroups were disaggregated into MEPA Levels 1 to 2 and MEPA Levels 3 to 5, differences emerged. First language family impact was not statistically significant for ELLs at low levels of English proficiency (MEPA Levels 1 to 2). In contrast, first language family impact was statistically significant for ELLs at MEPA Levels 3 and above, and 4.9% of the variance was attributed to the L1 language family favoring ELLs with a non-Latinate L1.

Second language acquisition and cognitive load theories supported a prediction that ELLs with an alphabetic L1 would perform better than ELLs with a non-alphabetic L1. Although the data indicated that L1 orthography (alphabetic/non-alphabetic) had a statistically significant impact on Biology MCAS score, it was unexpected that ELLs with a non-alphabetic L1 performed better. Though the impact of L1 orthography was statistically significant, its practical significance was minor, explaining 2% of the variance between groups. As with L1 language family, further disaggregation into subgroups based on English proficiency showed that the impact of L1 orthography was: (1) not statistically significant for ELLs at MEPA Levels 1 to 2 and (2) greater for ELLs at MEPA Levels 3 to 5 than for the whole sample. The finding that ELLs with a non-alphabetic L1 performed better, especially at MEPA Levels 3 and above where L1 orthography explained 3.4% of the variance between groups, was unexpected. The data suggested that at MEPA Level 3 alphabetic automaticity reduces the cognitive demands, and resources are available for content and attention to other item attributes.

Although some studies have shown that ELLs who arrive after 12 years of age encounter more difficulty in achieving academic parity with their native-speaking peers,

the data in this study indicated that late-entry ELL status did not have a statistically significant impact on Biology MCAS score. When the sample was further disaggregated into MEPA Levels 1 to 2 and MEPA Levels 3 to 5, there appeared to be a statistically significant impact that favored late-entry ELLs, though the practical significance was negligible.

In summary, ELLs at the higher levels of English proficiency overwhelmingly passed the Biology MCAS. English proficiency appeared to be the linguistic factor with the most impact on Biology MCAS score, and as English proficiency increased, Biology MCAS scores increased. The data indicated that English proficiency was a strong predictor of Biology MCAS performance and contributed 29% of the variance for ELLs as a whole and 25% of the variance for ELLs at MEPA Levels 3 to 5. First language characteristics favored ELLs with a non-Latinate L1 and ELLs with a non-alphabetic L1. The data suggested that the impact of the L1 language family and L1 orthography was greater for ELLs at MEPA Levels 3 to 5 than for the whole sample and not statistically significant for ELLs at MEPA Levels 1 and 2. Late-entry ELL status had no statistically significant impact for the whole sample and no practical significance for the MEPA Levels 1 to 2 and MEPA Levels 3 to 5 subgroups.

### **Performance on Content Domains**

“Individuals think and reason in relation to a content domain” (Barnett & Ceci, 2005, as cited in Leighton, Gokiart, & Cui, 2007, p. 143). The Biology MCAS assesses six content domains: anatomy and physiology, biochemistry, cell biology, ecology, evolution and biodiversity, and genetics. Each domain contains multiple content strands, which are specified in the High School Biology section of the 2006 Massachusetts

Science and Technology/Engineering Curriculum Framework.<sup>97</sup> As discussed earlier, no domain represented more than 20% of the multiple-choice items on the June 2012 Biology MCAS, and since there are only five constructed-response items, there was no biochemistry constructed-response item.

“[S]ubtle nuances in content area skill [drive] item difficulty” (Schneider et al., 2013, p. 112), and this presents challenges for evaluating content domain performance across multiple MCAS instruments because the domain standards under assessment may differ. Even within a content domain on a single instrument, some items may be easier than others depending on the nature and number of interacting knowledge elements (Van Merriënboer & Sweller, 2005). Items 5 and 18 were both ecology items on the June 2012 Biology MCAS (Appendix A). Item 5 assessed Standard 6.3, energy transfer in a food web, and item 18 assessed Standard 6.1, population dynamics. In item 5, a student needed to evaluate a food web to determine which organism competed with the mouse for the same food source. In item 18, a student had to evaluate data given in a table and determine which answer option would most improve the chances of a budworm surviving to adulthood.<sup>98</sup>

Studies indicate that wide-scale science assessments measure multiple dimensions, which supports the benefit of reporting sub-scores (Leighton et al., 2007). In addition to overall performance on the June 2012 Biology MCAS (40 multiple-choice and five constructed-response items), this study further explored ELL performance on the multiple-choice items across the six content domains. Table 4.33 and Figure 4.12

---

<sup>97</sup> The High School Biology standards are in Appendix C.

<sup>98</sup> Statewide, 97% of the test-takers correctly answered item 5, and 76% correctly answered item 18; both items were in the same content domain.

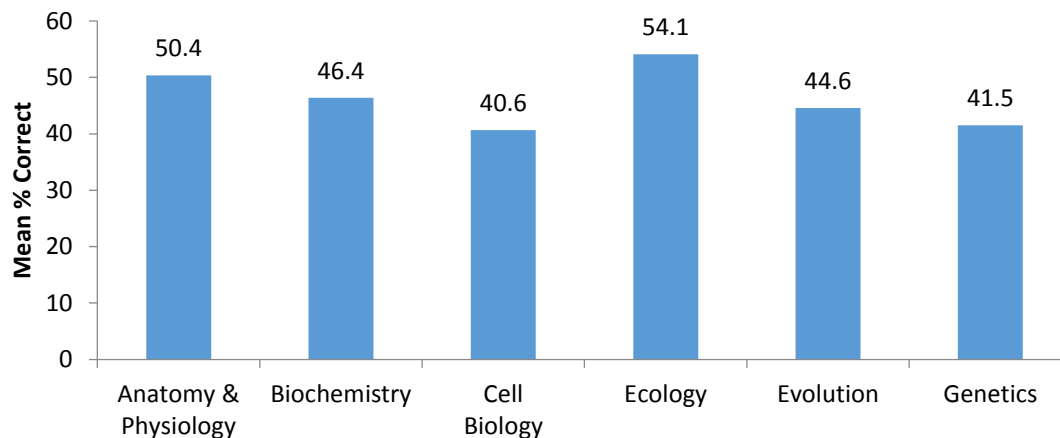
summarize the measures of central tendency for the content domains. ELLs performed better on two domains, ecology and anatomy and physiology, both of which had a mean percent correct greater than 50%: Ecology had a mean percent correct of 54.1% (SD = 21.66) and anatomy and physiology had a mean percent correct of 50.4% (SD = 26.06). On the remaining four content domains, the mean percent correct was less than 50%: biochemistry (M = 46.38, SD = 26.04), evolution (M = 44.56, SD = 22.91), genetics (M = 41.52, SD = 25.28), and cell biology (M = 40.63, SD = 23.38).

Table 4.33  
*ELL Multiple-Choice Percent Correct by Content Domain and Statewide Average Percent Correct*

n = 3,315		ELLs		Statewide	Difference
<u>Content domain</u>	<u>n</u>	<u>M</u>	<u>SD</u>	<u>Avg. %</u>	<u>(%)</u>
Anatomy & physiology	5	50.38	26.06	77.2	26.8
Biochemistry	5	46.38	26.04	74.8	28.4
Cell Biology	6	40.63	23.38	62.3	21.7
Ecology	8	54.11	21.66	79.8	25.6
Evolution	8	44.56	22.91	76.8	32.2
Genetics	8	41.52	25.28	71.7	30.0

The data suggested that ecology items were the easiest for ELLs and that cell biology items were the most challenging. As discussed previously, ecology and cell biology both had the majority of items at the conceptual cognitive skill (Table 4.12), and ecology items had higher mean linguistic complexity values than cell biology (Table 4.16). This suggested that the highest performance on ecology items and the lowest performance on cell biology items were due to the domain content.

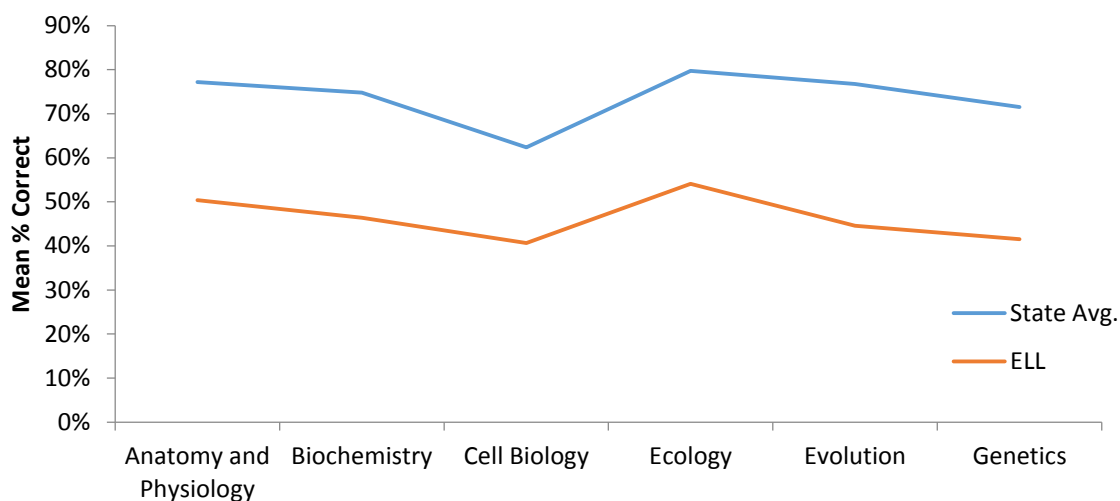




*Figure 4.12.* ELL Mean Percent Correct by Content Domain Multiple-Choice Items. ELLs had a mean percent correct greater than 50% for two content domains: Ecology and anatomy and physiology. The least mean percent correct was on cell biology items.

For comparison, a statewide mean percent correct was calculated for each domain.<sup>99</sup> Table 4.33 and Figure 4.13 summarize the data. The ELL sample and the reported statewide results had similar patterns for content domain performance. This appeared to confirm that differential performance across domains is due to intrinsic content elements. The data further confirmed that the ELL achievement gap seen for overall Biology MCAS performance persisted in each of the six content domains. The gap was smallest for cell biology, but this narrowing could be due to poor statewide performance on this domain.

<sup>99</sup> Statewide reported percent correct for items in a content domain were averaged.



*Figure 4.13.* ELL Domain Mean Percent Correct Compared to State Average. Although an achievement gap is evident, ELLs mirrored the statewide performance across content domains.

The higher performance on ecology and anatomy and physiology items was not surprising since these domains are also part of the middle school curriculum, though these domains are more developed at the high school level. Nevertheless, previous exposure to some domain concepts may explain the higher performance levels.<sup>100</sup> Although some cell biology content is taught in middle school, it is introductory, with three strands that cover unicellular and multicellular organisms and the structures and functions of animal and plant cells. At the high school level, cell biology has eight strands that include (but are not limited to) mitosis, meiosis, the role of ATP in metabolism, and using cellular evidence to classify organisms into the six taxonomic kingdoms; it also introduces Tier III words such as *prokaryote* and *eukaryote* (Appendix C). Many ecology and anatomy and physiology concepts are concrete (such as predation,

<sup>100</sup> These two domains are taught at earlier grades in Massachusetts.

competition, digestion,<sup>101</sup> and the water cycle), whereas cell biology concepts can be abstract (such as passive and active transport,<sup>102</sup> the ADP/ATP cycle, and the composition of the cell wall—cellulose or chitin—which distinguishes between multicellular organisms in the kingdoms Plantae and Fungi).<sup>103</sup>

It was, however, unexpected that ELL performance across domains mirrored the highest to lowest performance ranking of statewide averages, except for a reversal of biochemistry and evolution (Table 4.33). The data suggested that both English proficient and ELL test-takers found the ecology items the easiest and cell biology items the most difficult. The data further suggested an achievement gap in mean percent correct that ranged from approximately 21.7% (cell biology) to 32.2% (evolution). The existence of an achievement gap is clear; however, caution should be used in interpretation. Since this study only had access to statewide ELL data, it used district-level results, which included statewide percent correct for each item, to calculate an average percent correct for the content domain (MA DESE, 2012c). The ELLs in this study were also included in the statewide item percent correct figure, so the magnitude of the achievement gap is only an approximation. As discussed in Chapter 5, further study of performance among the ELL, former ELL, and never ELL subgroups is needed.<sup>104</sup>

**English proficiency impact.** As discussed previously, this study found that English proficiency level contributed 29% of the variance among ELLs on overall Biology MCAS score. The ELL sample was disaggregated by MEPA levels to explore

---

<sup>101</sup> See item 25 in Appendix A.

<sup>102</sup> See item 10 in Appendix A.

<sup>103</sup> Mode of nutrition (autotrophic or heterotrophic) also distinguishes between Plantae and Fungi.

<sup>104</sup> Former ELLs are ELLs who have been reclassified as English proficient, and never ELL are students who were never classified as limited English proficient (i.e., upon entry to Massachusetts schools, they were English proficient).

whether English proficiency also impacted domain performance. Table 4.34 and Figure 4.14 summarize measures of central tendency. At the lower end of English proficiency, MEPA Levels 1 and 2 had indistinguishable performance across content domains, and the mean percent correct for all domains was less than 50%. This was not unexpected since this was the same finding discussed previously for overall Biology MCAS performance at the lower end of English proficiency. As with the finding on overall Biology MCAS performance, the data indicated that domain performance increased as English proficiency increased. At MEPA Level 3, one domain, ecology, had a mean percent correct greater than 50% ( $M = 50.74$ ,  $SD = 19.76$ ). At MEPA Level 4, four domains had a mean percent correct greater than 50%: anatomy and physiology ( $M = 55.59$ ,  $SD = 26.36$ ), biochemistry ( $M = 50.28$ ,  $SD = 27.79$ ), ecology ( $M = 60.92$ ,  $SD = 19.63$ ), and evolution ( $M = 51.14$ ,  $SD = 22.65$ ). At MEPA Level 5, all domains had a mean percent correct greater than 50%. The highest mean percent correct for MEPA Level 5 was 69.49% ( $SD = 19.14$ ) on ecology items, and the most challenging domain appeared to be cell biology with a mean percent correct of 51.29% ( $SD = 25.03$ ). At MEPA Level 5, the achievement gap narrowed and ranged from an approximately 10.3% difference in mean percent correct for ecology to 15.1% for evolution; the difference between Level 5 ELLs and the statewide percentages was approximately half of what it was for the entire ELL sample. This suggested that although the achievement gap persisted, it narrowed with the gains in English proficiency by MEPA Level 5.

Table 4.34

*Domain Mean Percent Correct by English Proficiency (MEPA) Levels*

	Level 1 n = 276	Level 2 n = 441	Level 3 n = 1,441	Level 4 n = 603	Level 5 n = 554
<u>Content Domain</u>	<u>M</u>	<u>M</u>	<u>M</u>	<u>M</u>	<u>M</u>
Anatomy & physiology	40.51 (24.72)	39.14 (24.05)	47.81 (23.98)	55.59 (26.46)	65.23 (25.04)
Biochemistry	38.62 (23.31)	38.73 (22.57)	43.15 (24.37)	50.28 (27.79)	60.47 (26.08)
Cell biology	35.27 (21.84)	34.28 (21.02)	38.21 (21.78)	43.73 (24.22)	51.29 (25.03)
Ecology	42.57 (21.62)	43.74 (19.68)	50.74 (19.76)	60.92 (19.63)	69.49 (19.14)
Evolution	33.42 (20.24)	33.65 (18.29)	40.68 (20.39)	51.14 (22.65)	61.69 (22.31)
Genetics	29.71 (20.65)	32.28 (20.58)	37.38 (22.79)	47.55 (25.40)	58.96 (26.44)

Note: SD appears below the mean in parentheses.

The multiple-choice items represented only 40 of the 60 possible points on the June 2012 Biology MCAS. Caution should be used in interpreting the mean percent correct for each domain as passing or failing; however, the number of items in a content domain multiplied by the mean percent correct might predict the mean number of items answered correctly. Table 4.35 summarizes the predictions of domain items correct by English proficiency level. Level 5 ELLs are predicted to have 24.6 multiple-choice items correct, which would be passing at the Needs Improvement performance level. Level 4 ELLs are predicted to have 20.7 multiple-choice items correct, and with two points out of

the possible 20 points for the constructed-response items, they would also pass at the Needs Improvement performance level.

An additional 20 points were possible from the five constructed-response items, which have a maximum score of 4 points each. On the June 2012 Biology MCAS, the statewide average points for the constructed-response items were: (1) 2.00 for item 32, anatomy and physiology; (2) 1.35 for item 23, cell biology; (3) 2.17 for item 45, ecology; (4) 2.19 for item 12, evolution; and (5) 1.75 for item 44, genetics (MA DESE, 2012c). For ELLs, writing a constructed-response is more demanding than recognizing the correct answer and penciling in the answer option; the former is productive and the latter is receptive. It is difficult to predict how ELLs might score on the constructed-response items for each domain; however, multiplying the multiple-choice mean percent correct for the domain by the statewide average score on the domain's constructed-response yielded an approximation. Table 4.35 summarizes these calculations.

Table 4.35

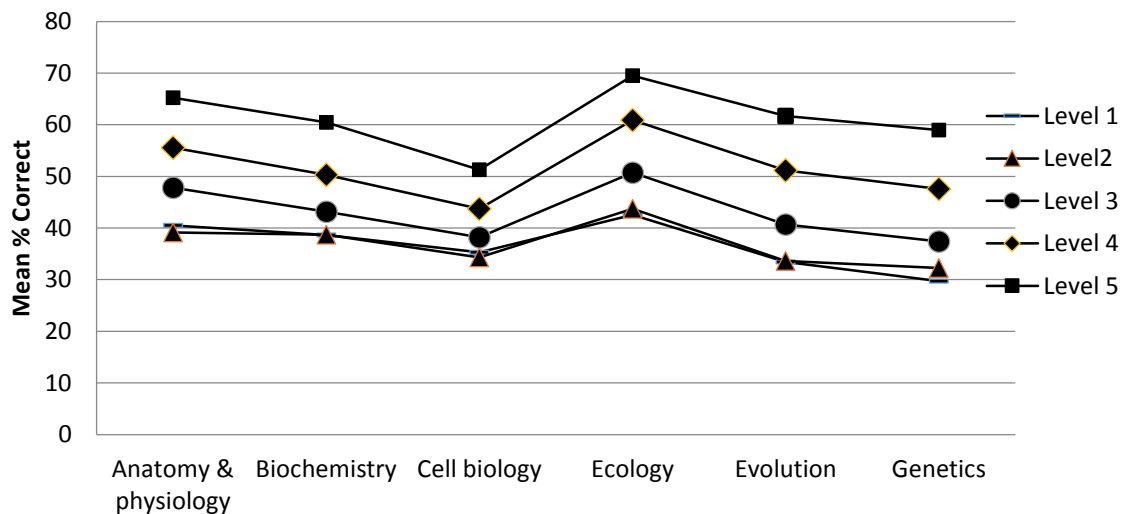
*Expected Domain Performance by English Proficiency Levels*

<u>Multiple-Choice</u>	<u>n</u>	<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Level 5</u>
Anatomy & physiology	5	2.0	2.0	2.4	2.8	3.3
Biochemistry	5	1.9	1.9	2.2	2.5	3.0
Cell biology	6	2.1	2.1	2.3	2.6	3.1
Ecology	8	3.4	3.5	4.1	4.9	5.6
Evolution	8	2.7	2.7	3.3	4.1	4.9
Genetics	8	2.4	2.6	3.0	3.8	4.7
Additional Points from Constructed-Response						
Anatomy & physiology		0.8	0.8	1.0	1.1	1.3
Cell biology		0.5	0.5	0.6	0.7	0.8
Ecology		0.8	0.7	0.8	0.9	1.1
Evolution		0.9	1.0	1.1	1.3	1.5
Genetics		0.6	0.6	0.7	0.9	1.1
Predicted Mean Score		18.1	18.3	21.3	25.7	30.4
Actual Mean Score		15.4	16.3	20.2	25.6	31.2

The actual mean scores for the entire instrument are less than the predicted score for Levels 1 through 3, and Levels 4 and 5 have similar actual and predicted scores (see Table 4.24). Performance on the content domains is consistent with the passing rates on the whole instrument as discussed previously. The majority of Levels 1 through 3 were below passing, but Level 3 approached the passing threshold of 22 points. In comparison, the majority of Levels 4 and 5 passed the Biology MCAS, with the mean of Level 5 ELLs approaching the Proficient threshold.

Interestingly, when the mean percent correct data were graphed, similar patterns emerged across the five MEPA levels (Figure 4.14), which still mirrored the statewide average percent correct (Figure 4.13). This suggested that content is the prevailing factor

in domain performance and that the higher scores as English proficiency increases result from increased item accessibility (comprehensibility). Comprehensible input would release cognitive resources for domain knowledge and skills, and result in an increased likelihood of demonstrating domain proficiency. As with overall MCAS score, linearity emerged at MEPA Level 3 (see Appendix K).



*Figure 4.14.* Mean Percent Correct by Domain by English Proficiency. Levels 1 and 2 had indistinguishable performance across domains. Level 5 had greater than 50% mean percent correct for all domains.

When the domains were ranked from highest to lowest percent correct by English proficiency levels, patterns emerged (see Table 4.36). MEPA Levels 1 and 2 had the same ordinal ranking, which was not unexpected since performance at these lower levels was indistinguishable. Likewise, MEPA Levels 4 and 5 had the same ordinal ranking. MEPA Level 3 appeared to fall between these two groups, with four of the six domains (ecology, anatomy and physiology, biochemistry, and genetics) having the same ranking as MEPA Levels 1 and 2, and cell biology and evolution moving towards the rankings for



MEPA Levels 4 and 5. Although the most difficult domain (defined as the least mean percent correct) differed among English proficiency levels, ecology items appeared the easiest for ELLs at all levels of English proficiency, followed by anatomy and physiology items.

Table 4.36

*Ranking of Domains by Mean Percent Correct for Each English Proficiency Level*

<u>Level 1</u>	<u>Level 2</u>	<u>Level 3</u>	<u>Level 4</u>	<u>Level 5</u>
Ecology	Ecology	Ecology	Ecology	Ecology
Anatomy	Anatomy	Anatomy	Anatomy	Anatomy
Biochemistry	Biochemistry	Biochemistry	Evolution	Evolution
Cell biology	Cell biology	Evolution	Biochemistry	Biochemistry
Evolution	Evolution	Cell biology	Genetics	Genetics
Genetics	Genetics	Genetics	Cell biology	Cell biology

One-way ANOVA analyses estimated the degree of difference in mean percent correct for each content domain across English proficiency levels. The independent variable was English proficiency (five MEPA levels), and the dependent variable was the percent correct for each domain. The data indicated that English proficiency had a statistically significant impact on performance across all six content domains: anatomy and physiology,  $F(4, 3310) = 94.54$ ,  $p < .001$ ; biochemistry,  $F(4, 3310) = 70.60$ ,  $p < .001$ ; cell biology,  $F(4, 3310) = 49.86$ ,  $p < .001$ ; ecology,  $F(4, 3310) = 165.88$ ,  $p < .001$ ; evolution,  $F(4, 3310) = 170.41$ ,  $p < .001$ ; and genetics,  $F(4, 3310) = 131.95$ ,  $p < .001$ . Univariate analyses of variance determined that English proficiency had practical significance with a medium or large effect sizes on domain performance. English

proficiency had the greatest impact on ecology and evolution items, where it contributed 17% of the variance, followed by genetics, where it contributed 14% of the variance.

This finding was not surprising since evolution had the highest mean for total answer lexical density (TALD) and the majority of genetics items were at the application cognitive skill level (Tables 4.16 and 4.12). The impact on ecology items is more difficult to explain. Although ecology items had the greatest stem lexical density (SLD), this was not statistically significant. The majority of ecology items were at the conceptual level, and as discussed previously, ELLs at all levels found these items the easiest. One hypothesis is that ecology items, especially food web items, have several organisms whose English names are unfamiliar (e.g., items 5 and 29; Appendix A). Parsing the unfamiliar names when reading the item and answer options could increase cognitive load. It may be that as English proficiency increases, ELLs are able to discriminate that these unfamiliar words were referential and cognitive resources focused on domain knowledge and skills, similar to native speakers, many of whom may not have known the names for these organisms. The impact of English proficiency on ecology performance is an area for future research. English proficiency had the least impact on cell biology items, contributing 6% of the variance. As discussed previously, cell biology items were the most challenging for all test-takers, including native speakers. One explanation is that English proficiency had the least impact for this domain because content knowledge and skills, not language, made these items more difficult.

Scheffé post hoc analyses confirmed that except for between MEPA Levels 1 and Level 2, all six domains had a statistically significant increase in mean percent correct between adjacent MEPA levels. These results are summarized in Table 4.37. The

increase in mean percent correct between MEPA Level 3 and Level 5 ranged from 13.8% (cell biology) to 21.58% (genetics), with five of the six domains having increases in mean percent correct greater than 17%. This was consistent with the findings discussed previously that when ELLs reached MEPA Level 3, English proficiency had both statistical and practical impact on performance.

Table 4.37

*Summary of Scheffé Post Hoc Analysis on Domain Percent Correct across Adjacent English Proficiency Levels*

	(I) MEPA Level	(J) MEPA Level	Mean Difference* (I-J)
Anatomy and Physiology	Level 1	Level 2	1.37
	Level 2	Level 3	-.8.68***
	Level 3	Level 4	-7.78***
	Level 4	Level 5	-9.65***
Biochemistry	Level 1	Level 2	-0.12
	Level 2	Level 3	-4.42*
	Level 3	Level 4	-7.13***
	Level 4	Level 5	-10.19***
Cell Biology	Level 1	Level 2	0.99
	Level 2	Level 3	-3.94*
	Level 3	Level 4	-5.51***
	Level 4	Level 5	-7.57***
Ecology	Level 1	Level 2	-1.16
	Level 2	Level 3	-7.00***
	Level 3	Level 4	-10.19***
	Level 4	Level 5	-8.57***
Evolution	Level 1	Level 2	-0.00
	Level 2	Level 3	-7.04***
	Level 3	Level 4	-10.46***
	Level 4	Level 5	-10.55***
Genetics	Level 1	Level 2	-2.57
	Level 2	Level 3	-5.09**
	Level 3	Level 4	-10.18***
	Level 4	Level 5	-11.40***

\*p < .05 \*\*p < .01 \*\*\*p < .001.

This study further determined that the impact of English proficiency on performance is observed in domain sub-scores. Domain performance data were consistent with the previously discussed finding that English proficiency impacts overall MCAS performance. MEPA Level 3 was the transition point, partway between the lower levels

(MEPA Levels 1 and 2) where performance was indistinguishable and the higher levels of English proficiency (MEPA Levels 4 and 5) where the achievement gap narrowed.<sup>105</sup> The domain performance data were consistent with the findings discussed previously that MEPA Level 3 is the cusp of where ELLs can begin to pass the Biology MCAS but still cannot demonstrate proficiency, and MEPA Levels 4 and 5 represent the point where ELLs start to demonstrate content proficiency.

**First language family impact.** Disaggregation by first language family explored the impact of a Latinate or non-Latinate L1 on performance across domains. Table 4.38 and Figure 4.15 summarize measures of central tendency. The data indicated that ELLs with a non-Latinate L1 (n = 895) performed better on every content domain than ELLs with a Latinate L1 (n = 2,420); however, the difference appeared minor for ecology and anatomy and physiology, the two domains that appeared easiest for the whole sample. The greatest difference in performance appeared to be on genetics items (11.7%), followed by cell biology (9.7%), which was the most difficult domain for both groups, and biochemistry (8.6%).

---

<sup>105</sup> The ordinal ranking of MEPA Levels 4 and 5 was the same as for the statewide average percent correct.

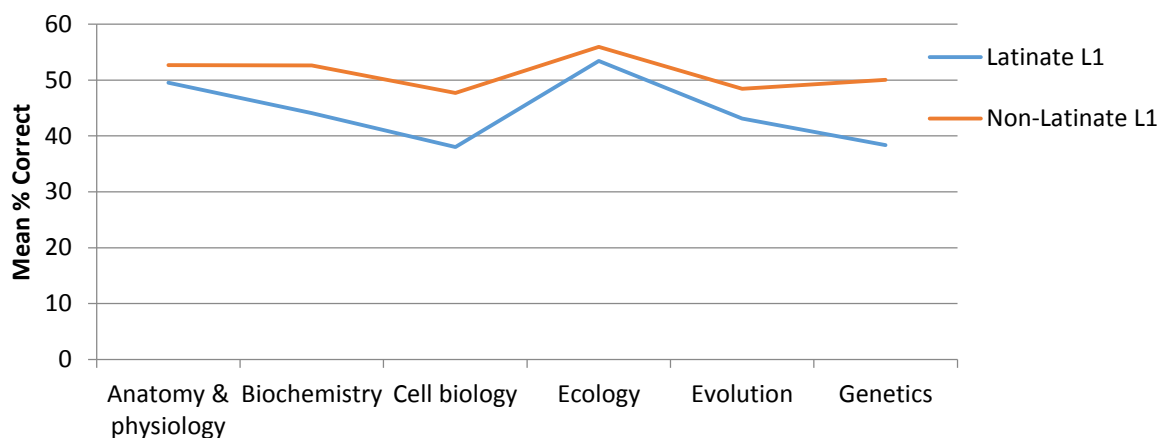
Table 4.38  
*Domain Mean Percent Correct by Language Family*

<u>Domain</u>	Látinate (n = 2,420)		Non-Látinate (n = 895)		<u>t</u>	Cohen's <u>d</u>	Partial <u>eta</u> <sup>2</sup>
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>			
Anatomy & physiology	49.52	27.20	52.69	25.58	-3.03**	.12	.003
Biochemistry	44.07	25.09	52.63	27.50	-8.14***	.32	.021
Cell biology	38.02	22.00	47.69	25.45	-10.06***	.41	.034
Ecology	53.42	21.30	55.98	22.53	-2.94**	.12	.003
Evolution	43.12	21.99	48.44	24.82	-5.64***	.23	.011
Genetics	38.35	23.72	50.08	27.30	-11.36***	.46	.042

\*p < .05, \*\* p < .01, \*\*\*p < .001

Independent samples t-tests confirmed that L1 language family (Látinate/non-Látinate) had a statistically significant impact on performance for all content domains with non-Látinate L1 ELLs scoring higher: (1) anatomy and physiology,  $t(1514) = -3.03$ ,  $p < .01$ ; (2) biochemistry,  $t(1477) = -8.14$ ,  $p < .001$ ; (3) cell biology,  $t(1416) = -10.06$ ,  $p < .001$ ; (4) ecology,  $t(1520) = -2.94$ ,  $p < .01$ ; (5) evolution,  $t(1443) = -5.64$ ,  $p < .001$ ; and (6) genetics,  $t(1422) = -11.36$ ,  $p < .001$ . Table 4.38 summarizes the results of these analyses. Cohen's d values determined that L1 language family had a negligible effect size for ecology (.12) and anatomy and physiology (.12), and a small effect size for the remaining four content domains: evolution (.23), biochemistry (.32), cell biology (.41), and genetics (.46).<sup>106</sup>

<sup>106</sup> Cohen's d for the cell biology and genetics domains indicated that the effect sizes approached medium.



*Figure 4.15.* Domain Mean Percent Correct by L1 Language Family. ELLs with a non-Latinate L1 had a greater percent correct than ELLs with a Latinate L1 for all content domains, but the gap almost closed for anatomy and physiology and ecology.

Although the data indicated a positive impact of a non-Latinate L1, univariate analyses of variance determined that the practical significance was negligible except for genetics and cell biology, where it contributed 4.2% and 3.4%, respectively, of the variance between the groups. As discussed previously with the similar finding that non-Latinate L1 ELLs performed better on overall score, this finding was unexpected. The performance gap between ELLs with a Latinate L1 and a non-Latinate L1 persisted in domain sub-scores, but the gap was minimal for ecology and anatomy and physiology, the two domains where ELLs performed best, regardless of English proficiency level. Many anatomical terms derive from Latin, and this may have helped ELLs with a Latinate L1. Although Latinate L1-to-English cognates did not appear to confer a consistent advantage, further research is needed to determine whether cognates helped close the performance gap for anatomy and physiology and ecology.

### First language orthography impact. Disaggregation by first language

orthography explored the impact of an alphabetic or non-alphabetic L1 on content domain performance. Table 4.39 and Figures 4.16 summarize measures of central tendency.

Table 4.39

*Domain Percent Correct by First Language Orthography*

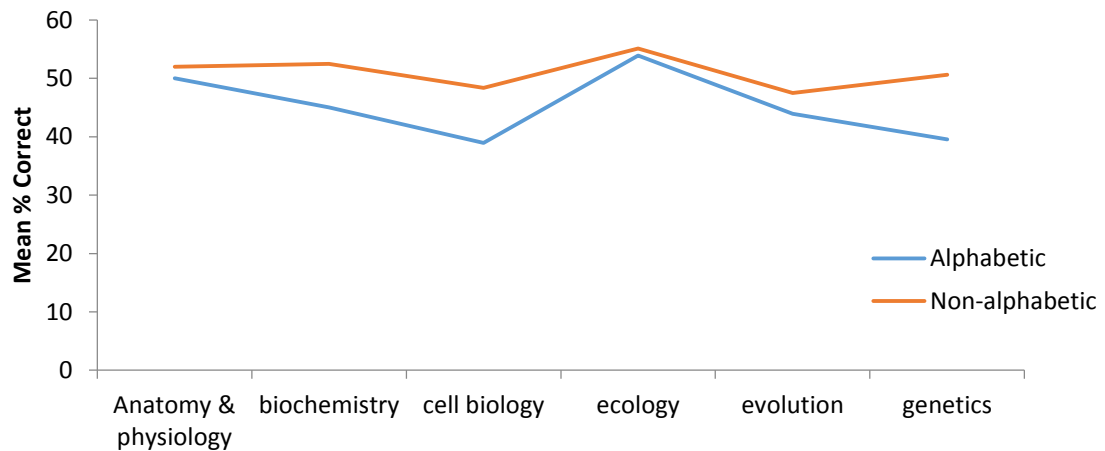
<u>Domain</u>	Alphabetic (n = 2,724)		Non-alphabetic (n = 591)		<u>t</u>	Cohen's <u>d</u>	Partial <u>eta</u> <sup>2</sup>
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>			
Anatomy & physiology	50.03	25.78	51.98	27.28	-1.65		
Biochemistry	45.05	25.54	52.49	27.39	-6.05***	.28	.012
Cell biology	38.96	22.41	48.36	26.04	-8.15***	.39	.024
Ecology	53.90	21.38	55.12	22.94	-1.19		
Evolution	43.92	22.48	47.50	24.59	-3.26**	.15	.004
Genetics	39.54	24.18	50.63	28.10	-8.91***	.43	.028

\*p < .05, \*\* p < .01, \*\*\*p < .001

The pattern of mean percent correct across domains for L1 orthography was similar to the pattern for L1 language family. Alphabetic L1 ELLs (n = 2,724) and non-alphabetic L1 ELLs (n = 591) had similar performance on ecology and anatomy and physiology, and the difference was minor on evolution (3.6%), with non-alphabetic L1 ELLs favored. The data suggested that non-alphabetic L1 ELLs performed better on three domains: biochemistry (7.4% difference), cell biology (9.4% difference), and genetics (11.1% difference). As with the entire sample and the L1 language family subgroups, ecology appeared the easiest for both L1 orthography subgroups. Both subgroups also



found cell biology difficult, but the data suggested that non-alphabetic L1 ELLs struggled more with evolution.



*Figure 4.16.* Domain Mean Percent Correct by L1 Orthography. ELLs with a non-alphabetic L1 had a greater percent correct than ELLs with an alphabetic L1 on four content domains, but the gap almost closed for anatomy and physiology and ecology.

Independent samples t-tests confirmed that there was no statistically significant difference in performance for anatomy and physiology,  $t(2213) = -1.65$ ,  $p > .05$ , and ecology,  $t(827) = -1.19$ ,  $p > .05$ . Table 4.39 summarizes these results. The analyses further indicated that L1 orthography had a statistically significant impact on performance on the other four domains: biochemistry,  $t(827) = -6.05$ ,  $p < .001$ ; cell biology,  $t(790) = -8.15$ ,  $p < .001$ ; evolution,  $t(818) = -3.26$ ,  $p < .01$ ; and genetics,  $t(790) = -8.91$ ,  $p < .001$ . Cohen's  $d$  values and univariate analyses of variance, however, determined that the effect size was negligible for evolution and small for biochemistry, cell biology, and genetics, with L1 orthography contributing only 1.2% to 2.8% of the

variance between subgroups. The data suggested that L1 orthography had minor practical significance for three domains and overall did not appear to be a good predictor of Biology MCAS performance.

**Late-entry ELL status impact.** The sample was disaggregated into late-entry and not-late-entry ELLs to explore the impact of age of entry on performance across domains. Table 4.40 summarizes measures of central tendency. Late-entry ELLs and not-late-entry ELLs generally had a similar mean percent correct across the content domains; biochemistry had the greatest difference with a mean percent correct approximately 4% higher for late-entry ELLs. For both subgroups, ecology items appeared the easiest, and cell biology and genetics presented the most difficulty. The data suggested that late-entry ELL status had minimal impact on domain performance.

Table 4.40  
*Domain Percent Correct by Late-Entry ELL Status*

<u>Domain</u>	Late-entry ELL (n = 2,218)		Not-late-entry ELL (n = 1,097)		<u>t</u>	Cohen's <u>d</u>	Partial <u>eta</u> <sup>2</sup>
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>			
Anatomy & physiology	50.85	26.43	49.43	25.28	1.50	.06	.001
Biochemistry	47.74	26.02	43.63	25.86	4.29***	.16	.006
Cell biology	41.28	24.00	39.33	22.03	2.32*	.08	.002
Ecology	53.64	22.20	55.08	20.47	-1.86	.08	.001
Evolution	44.34	23.03	44.99	22.68	-0.76	.03	.000
Genetics	42.36	25.99	39.82	23.69	2.80**	.10	.002

\*p < .05, \*\* p < .01, \*\*\*p < .001

Independent samples t-tests confirmed that there was no statistically significant impact of late-entry ELL status for three domains: anatomy and physiology,  $t(2271) = 1.50$ ,  $p > .05$ ; ecology,  $t(2351) = -1.86$ ,  $p > .05$ ; and evolution,  $t(3313) = -0.76$ ,  $p > .05$ . Table 4.40 summarizes the results of these analyses. Independent samples t-tests determined a statistically significant difference for three domains where late-entry ELLs were favored: biochemistry,  $t(3313) = 4.29$ ,  $p < .001$ ; cell biology,  $t(2358) = 2.32$ ,  $p < .05$ ; and genetics,  $t(2374) = 2.80$ ,  $p < .01$ . Cohen's  $d$  values, however, determined that the effect size was negligible, and univariate analyses of variance indicated that less than 1% of the variance was attributable to late-entry ELL status. Although this study did not predict a difference between subgroups, the finding was surprising. It is reasonable to assume that some late-entry ELLs have L1 biology knowledge, and this would transfer to the L2 once there was sufficient English proficiency. Interestingly, the three domains where late-entry ELLs had higher performance that was statistically significant but with negligible practical significance were domains that are traditionally thought of as high school content—biochemistry, cell biology, and genetics.

Based on the previously discussed findings that MEPA Levels 1 and 2 had indistinguishable performance on content domains, this study further explored the impact of late-entry ELL status on content domain performance for ELLs at MEPA Levels 3 to 5. Table 4.41 summarizes the results.

Table 4.41

*Domain Mean Percent Correct by Late-Entry ELL Status for MEPA Levels 3 to 5*

<u>Domain</u>	Late-entry ELL (n = 1,551)		Not-late-entry ELL (n = 1,047)		<u>t</u>	Cohen's <u>d</u>	Partial <u>eta</u> <sup>2</sup>
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>			
Anatomy & physiology	55.53	25.80	50.09	25.34	5.31***	.21	.011
Biochemistry	51.30	26.57	44.36	25.77	6.61***	.26	.017
Cell biology	43.93	24.62	39.84	21.92	4.43***	.18	.007
Ecology	57.87	21.49	55.96	20.30	2.30*	.09	.002
Evolution	48.77	23.09	45.85	22.66	3.19***	.13	.004
Genetics	46.98	26.64	40.44	23.76	6.55***	.26	.016

\*p &lt; .05, \*\* p &lt; .01, \*\*\*p &lt; .001

The data indicated statistically significant differences between the groups for all six content domains. Cohen's d values and univariate analyses of variance, however, determined that the effect sizes were either small or negligible, with no more than 1.7% (biochemistry) of the variance attributed to late-entry ELL status. The common underlying proficiency model would suggest that late-entry ELLs with prior L1 content knowledge would perform better. Statistical significance for all six domains at MEPA Levels 3 to 5 but not for MEPA Levels 1 and 2 suggest that further research is needed into the impact of English language proficiency and possible interactions with L1 content knowledge. As discussed in Chapter 5, a limitation of using secondary data was the lack of information on prior schooling in the L1.

**Summary of content domain performance.** The ELL sample had a mean percent correct of 50% or greater only on two of the six content domains, ecology and anatomy and physiology. It appeared that ELLs generally performed best on ecology items (54.1%) and worst on cell biology items (40.6%). When the ELL sample was disaggregated by English proficiency, ecology items had the highest mean percent correct for all MEPA levels. The content domain with the lowest mean percent correct, however, differed among the MEPA levels. Genetics items generally appeared to be the most difficult at lower levels of English proficiency (MEPA Levels 1 to 3), and cell biology items appeared to be the most difficult at the higher levels of English proficiency (MEPA Levels 4 and 5). English proficiency was a moderate predictor of performance and contributed 6% to 10% of the variance on three content domains (anatomy and physiology, biochemistry, and cell biology) and 14% to 17% of the variance on the other three content domains (ecology, evolution, and genetics).

First language family (Latinate/non-Latinate) impact appeared to favor ELLs with a non-Latinate L1 on all six content domains. Although performance differences were statistically significant for all content domains, the practical significance was small for cell biology and genetics, where it contributed 3.4% and 4.2%, respectively, of the variance; impact was negligible for the other four domains. First language orthography impact appeared to favor ELLs with a non-alphabetic L1 on four content domains (biochemistry, cell biology, evolution, and genetics); however, the practical significance was negligible to minor, with the greatest impact on cell biology and genetics where it contributed 2% to 3% of the variance. The data suggested that L1 characteristics,

language family and orthography, only had a small impact on two domains, cell biology and genetics.

The impact of late-entry ELL status, defined as age of entry of 12 years or later, was absent or negligible. Analyses indicated that there was no statistically significant difference between the groups for performance on three of the six content domains (anatomy and physiology, ecology, and evolution). Although late-entry ELLs generally performed better on three content domains (biochemistry, cell biology, and genetics), analyses indicated there was no practical significance, with less than 1% of the variance between groups attributed to late-entry ELL status. Further disaggregation into a Level 3 to 5 English proficiency subgroup only indicated a negligible impact that contributed 1.6% and 1.7% of the variance on genetics and biochemistry, respectively. The data suggested that late-entry ELL status generally had no or negligible impact on content domain performance.

In summary, ELLs generally performed best on ecology items and worst on cell biology items. The best performance on ecology items was seen in all subgroups; however, the domain that presented the most difficulty varied among subgroups. English proficiency had the greatest impact and contributed between 6% and 17% of the variance among the subgroups. The data suggested that L1 characteristics had a small impact on cell biology and genetics performance, with L1 family having slightly more impact (3.4% and 4.2%) than L1 orthography (2.4% and 2.8%). Late-entry ELL status impact, where present, was negligible. The data further suggested that the performance gap between Latinate L1 and non-Latinate L1 ELLs on overall Biology performance existed in domain sub-scores, though to a lesser degree.

## **Performance at Cognitive Skill Levels**

Leighton, Gokiert, and Ying (2007) maintained that cognitive skills are more indicative of mastery than content because the former manipulates the latter; however, even though cognitive skills require content knowledge, the reverse is not always true—i.e., it is possible to answer an item correctly with some content knowledge but no reasoning skill. In a study of the Canadian national standardized science assessment, Leighton and Gokiert (2005) found that although most students could identify the knowledge and skills required to answer an item, they could not identify the general concept of the item. The June 2012 Biology MCAS had 45% ( $n = 18$ ) of the multiple-choice items at the conceptual cognitive skill level, raising the question of whether ELLs found conceptual items difficult. Everything else being equal, increased task complexity should be accompanied by some decrease in performance.

As discussed in Chapter 2, Robinson (2005) made predictions based on his cognition hypothesis. One prediction was that tasks with higher cognitive demand would promote heightened attention and noticing and increased interaction and negotiation, which would lead to increased self-repair and greater accuracy in L2 output. In a later study, Robinson and Gilabert (2007) found that increased task complexity resulted in second language learners making fewer errors, increased self-repair, and increased use of higher frequency words. Michel, Kuiken, and Vedder (2007) found that monologic task conditions resulted in a higher percentage of self-repair, but there was no robust interaction between task complexity and task condition (monologic or dialogic). Although the MCAS multiple-choice items do not require L2 output, the heightened attention and noticing could lead to increased self-repair in reading comprehension and

thus lead to increased performance.<sup>107</sup> Another prediction from the cognition hypothesis was that individual learner characteristics such as cognitive ability and affective factors would have increasing impact as task complexity increased (Robinson, 2005). The cognition hypothesis suggests that performance should increase as cognitive skill levels increase because of the increased noticing; however, it also suggests that as task complexity increases, individual learner characteristics would have a greater impact. It should be noted that the cognition hypothesis relates to L2 oral output, and L2 on MCAS multiple-choice items is receptive, not productive, language.

In addition to overall performance on the June 2012 Biology MCAS (40 multiple-choice and five constructed-response items) and performance on each of the six content domains, this study further explored ELL performance for the item attribute of cognitive skill level. As previously discussed, the June 2012 Biology MCAS multiple-choice items had three cognitive skill levels: foundational, conceptual, and application (Table 4.11). Measures of central tendency are summarized in Table 4.42. The mean percent correct for the item cognitive skill levels were all below 50%: foundational ( $M = 48.98$ ;  $SD = 23.93$ ), conceptual ( $M = 46.14$ ;  $SD = 19.19$ ), and application ( $M = 45.30$ ;  $SD = 19.72$ ). The descriptive statistics had two unexpected findings. The first was similar performance for conceptual and application items; the performance difference was less than 1%. The second was a tighter dispersion for the two higher skill levels. Since the cognition hypothesis posits that individual learner characteristics (including cognitive ability) would have increased impact with higher task complexity, greater dispersion would be

---

<sup>107</sup> Although the terms monologic and dialogic describe oral language output, L2 reading is more similar to the monologic condition than to the dialogic condition.



expected for application items. Since there are no publically reported results of MCAS performance by cognitive skill level, a comparison with statewide averages to determine if this was a general pattern for performance on the Biology MCAS or specific to the ELL sample was not possible. It must be noted, however, that there were approximately two to three times as many conceptual (n = 18) and application items (n = 16) as foundational items (n = 6). The similar number of conceptual and application items might be a reason for similar performance, and the disproportionately fewer foundational items may be the reason for a slight performance difference.

Table 4.42  
*Multiple-Choice Percent Correct by Cognitive Skill Level*

ELLs (n = 3315)

<u>Cognitive skill level</u>	<u>n</u>	<u>M</u>	<u>SD</u>
Foundational	6	48.98	23.96
Conceptual	18	46.14	19.19
Application	16	45.30	19.72

**English proficiency impact.** Disaggregation by MEPA level explored the impact of English proficiency on performance across item cognitive skill levels. Measures of central tendency are summarized in Table 4.43. For MEPA Levels 1, 2, and 3, the mean percent correct for all three cognitive skill levels was less than 50%. In contrast, MEPA Level 4 had mean percent correct greater than 50%, and MEPA Level 5 had mean percent correct greater than 60% for all cognitive skill levels. Except for between MEPA Levels 1 and 2, performance increased for each of the cognitive skill levels as English

proficiency increased. This was not surprising since it echoed the English proficiency pattern for overall and content domain performance.

Table 4.43

*Comparison of Cognitive Skill Mean Percent Correct by English Proficiency*

<u>Cognitive skill</u>	Level 1 n = 276	Level 2 n = 441	Level 3 n = 1,441	Level 4 n = 603	Level 5 n = 554
Foundational	37.32 <sub>a</sub> (21.98)	39.12 <sub>a</sub> (22.36)	46.62 <sub>b</sub> (22.31)	53.59 <sub>c</sub> (22.67)	63.75 <sub>d</sub> (23.31)
Conceptual	36.11 <sub>e</sub> (16.69)	36.31 <sub>e</sub> (14.57)	42.86 <sub>f</sub> (16.79)	51.97 <sub>g</sub> (18.97)	61.04 <sub>h</sub> (18.99)
Application	36.19 <sub>i</sub> (16.44)	36.39 <sub>i</sub> (15.83)	41.45 <sub>j</sub> (16.80)	50.73 <sub>k</sub> (19.89)	61.01 <sub>l</sub> (20.21)

Note: \* $p < .05$  \*\* $p < .01$  \*\*\* $p < .001$ . Standard deviations appear in parentheses below means. Results of Scheffé post hoc analyses using paired comparisons are shown using subscripts (a, b, c, ... l). Means with the same subscript are not significantly different while means with different subscripts are significantly different from one another at the  $p < .05$  level.

One-way ANOVA analyses confirmed that English proficiency had a statistically significant impact on performance for all three cognitive skill levels: foundational,  $F(4, 3310) = 109.24$ ,  $p < .001$ ; conceptual,  $F(4, 3310) = 190.61$ ,  $p < .001$ ; and application,  $F(4, 3310) = 183.36$ ,  $p < .001$ . Table 4.43 summarizes the results of the ANOVA analyses. As English proficiency increased, the mean percent correct at a cognitive skill level increased. Univariate analyses determined that English proficiency had a medium effect size on foundational items and contributed 11.7% of the variance among MEPA levels. The effect size for conceptual and application items was large, and English proficiency contributed 18.7% and 18.3%, respectively, to the variance among subgroups.

Scheffé post hoc analyses determined that except between MEPA Levels 1 and 2, there were statistically significant differences among the five MEPA levels. These results are summarized in Table 4.44. This was not unexpected since MEPA Levels 1 and 2 had

indistinguishable performance on overall score and content domains. The data indicated that English proficiency impact on cognitive skill level performance became statistically significant when English proficiency increased from MEPA Level 2 to MEPA Level 3, as well as between successively higher English proficiency levels. The greatest differences were between Level 4 and Level 5, where mean percent correct was 9% to 10% higher for each cognitive skill.

Table 4.44

*Summary of Scheffé Post Hoc Analysis on Cognitive Skill Percent Correct across Adjacent English Proficiency Levels*

	(I) MEPA Level	(J) MEPA Level	Mean Difference (I-J)
Foundational	Level 1	Level 2	-1.80
	Level 2	Level 3	-7.51***
	Level 3	Level 4	-6.97***
	Level 4	Level 5	-10.16***
Conceptual	Level 1	Level 2	-0.30
	Level 2	Level 3	-6.46***
	Level 3	Level 4	-9.12***
	Level 4	Level 5	-9.07***
Application	Level 1	Level 2	-0.21
	Level 2	Level 3	-5.06***
	Level 3	Level 4	-9.27***
	Level 4	Level 5	-10.28***

\*p < .05, \*\* p < .01, \*\*\*p < .001

The finding that ELL performance was similar for conceptual and application items was unexpected; however, it might be explained by one of the cognition hypothesis predictions discussed previously. At higher levels of English proficiency, content becomes increasingly more dominant than language in allocation of cognitive resources, and students can better discriminate task complexity. After determining that an item is

complex, students can attend to subtle differences in language and content—i.e., pay closer attention to the question and answer options. It seems self-evident that, all things being equal, close reading and noticing nuances in the stem and options lead to increased performance. For ELLs, however, all things are not equal because they are still acquiring English. Thus, increasing levels of English proficiency is a factor in equalizing all other factors. MCAS performance for cognitive skill levels is not publically reported, so this study could not compare ELL performance to the entire test-taking sample. As discussed in Chapter 5, future studies might explore whether the pattern of similar performance for conceptual and application items holds for non-ELL test-takers.

**First language family impact.** Disaggregation by first language family explored the impact of a Latinate or non-Latinate L1 on performance across item cognitive skill levels. Table 4.45 and Figure 4.17 summarize measures of central tendency. Latinate L1 ELLs had a mean percent correct less than 50% for all three cognitive skill levels. In comparison, ELLs with a non-Latinate L1 had a mean percent correct greater than 50% for all three cognitive skill levels. For both groups, foundational cognitive skill items had the highest mean percent correct, and performance was similar for the conceptual and application levels.

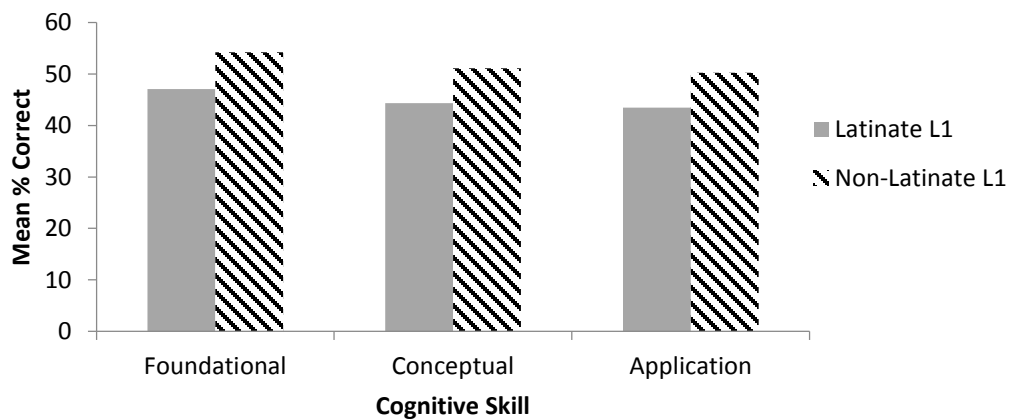
Table 4.45

*Cognitive Skill Level Percent Correct by Language Family*

	Latinate (n = 2,420)		Non-Latinate (n = 895)				
<u>Cognitive Skill</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>t</u>	Cohen's <u>d</u>	Partial <u>eta</u> <sup>2</sup>
Foundational	47.05	23.49	54.21	24.43	-7.71***	.30	.018
Conceptual	44.30	17.91	51.11	21.53	-8.45***	.34	.025
Application	43.48	18.53	50.22	21.90	-8.12***	.33	.023

\*p < .05, \*\* p < .01, \*\*\*p < .001

Independent samples t-tests confirmed that there was a statistically significant difference in mean percent correct between subgroups. The results of these analyses are summarized in Table 4.45. The results indicated that ELLs with a non-Latinate L1 performed better at all three cognitive skill levels: foundational,  $t(3313) = -7.71$ ,  $p < .001$ ; conceptual,  $t(1376) = -8.45$ ,  $p < .001$ ; and application,  $t(1394) = -8.12$ ,  $p < .001$ . Cohen's d, however, determined that L1 language family had a small effect size, and univariate analyses of variance determined that L1 language family contributed 1.8% of the difference in performance on foundational items, 2.5% on conceptual items, and 2.3% on application items. The data suggested that first language family had minor impact on performance at the different item cognitive skill levels.



*Figure 4.17.* Mean Percent Correct for Cognitive Skill by First Language Family. ELLs with a non-Latinate L1 had higher mean percent correct for each cognitive skill.

**First language orthography impact.** Disaggregation by first language orthography explored the impact of an alphabetic or non-alphabetic L1 on performance at the three item cognitive skill levels. Table 4.46 and Figure 4.18 summarize measures of central tendency. ELLs with an alphabetic L1 had mean percent correct less than 50% for all three cognitive skill levels, and ELLs with a non-alphabetic L1 had mean percent correct of 50% or greater for all cognitive skill levels.<sup>108</sup> For both groups, foundational cognitive skill items had the best performance, and performance was similar for conceptual and application items. This was the same pattern seen for the entire sample and for the subgroups discussed previously.

<sup>108</sup> Non-alphabetic L1 ELLs had a mean percent correct of 49.97%; however, this becomes 50% when rounded to one decimal point.

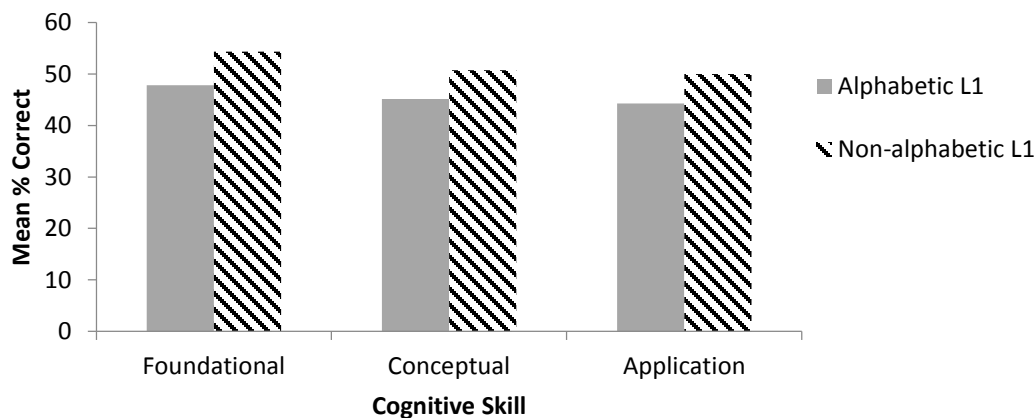
Table 4.46

*Cognitive Skill Level Percent Correct by L1 Orthography*

<u>Cognitive Skill</u>	Alphabetic (n = 2,724)		Non-alphabetic (n = 591)		<u>t</u>	Cohen's <u>d</u>	Partial <u>eta</u> <sup>2</sup>
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>			
Foundational	47.82	23.70	54.34	24.42	-6.04***	.27	.011
Conceptual	45.14	18.41	50.72	21.87	-5.78***	.28	.012
Application	44.28	19.01	49.97	22.13	-5.80***	.28	.012

\*p < .05, \*\* p < .01, \*\*\*p < .001

Independent samples t-tests determined that ELLs with a non-alphabetic L1 had statistically better performance for each cognitive skill level: foundational,  $t(3313) = -6.04$ ,  $p < .001$ ; conceptual,  $t(781) = -5.78$ ,  $p < .001$ ; and application,  $t(790) = -5.80$ ,  $p < .001$ . Table 4.46 summarizes the results. Cohen's d values, however, determined there was only a small effect size, and univariate analyses of variance determined that there was negligible impact, with 1.1% of the difference for foundational items, 1.2% for conceptual items, and 1.2% for application items attributed to L1 orthography.



*Figure 4.18.* Mean Percent Correct for Cognitive Skills by First Language Orthography. ELLs with a non-alphabetic L1 had higher mean percent correct for each cognitive skill level.

**Late-entry ELL status impact.** Disaggregation by late-entry or not-late-entry ELL status explored the impact of age of entry on performance on items by cognitive skill levels. Measures of central tendency are summarized in Table 4.47. Both groups had similar performance at all three cognitive skill levels, with the best performance on foundational items. Independent samples t-tests confirmed that there was no statistically significant difference in performance between the subgroups: foundational,  $t(3313) = 1.38, p > .05$ ; conceptual,  $t(2432) = 1.28, p > .05$ ; and application,  $t(3313) = 1.79, p > .05$ . Table 4.47 summarizes the results. The data suggested that whether an ELL entered before or after the age of 12 years had no significant impact on performance on items of different cognitive skill levels.



Table 4.47

*Cognitive Skill Level Percent Correct by Late-Entry ELL Status*

	Late-entry ELL (n = 2,218)		Not-late-entry ELL (n = 1,097)		
<u>Cognitive Skill</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>t</u>
Foundational	49.38	24.39	48.16	23.04	1.38
Conceptual	46.43	19.90	45.55	17.66	1.28
Application	45.71	19.99	44.47	19.41	1.70

\*p < .05, \*\* p < .01, \*\*\*p < .001

**Summary of cognitive skill level performance.** The June 2012 Biology MCAS multiple-choice items had three cognitive skill levels: foundational, conceptual, and application. For the ELL sample as a whole, the mean percent correct for each item cognitive skill levels was below 50%. As with overall MCAS performance and content domain performance, English proficiency had the strongest impact on performance for each cognitive skill level. Univariate analyses of variance determined that English proficiency contributed 11.7% of the performance difference among MEPA levels for foundational items, and it had a greater impact for conceptual and application items (contributing 18.7% and 18.1%, respectively). The data further suggested that L1 language family had a minor impact with between 1.8% and 2.5% of the variance attributed to whether the L1 was Latinate; non-Latinate L1 ELLs appeared to be favored. First language orthography appeared to have negligible impact, contributing approximately 1% of the variation among groups. The data also suggested that late-entry ELL status had no statistically significant impact on performance at the cognitive skill levels. The finding of similar performance for conceptual and application items for the

ELL sample as well as for subgroups was unexpected. Robinson's (2003) cognition hypothesis offers a possible expectation, as does the disproportionately fewer foundational items on the June 2012 Biology MCAS. Further research is needed.

### **Performance on Item Linguistic Complexity Levels**

As discussed in Chapter 2, assessments become de facto dual assessments for content knowledge and academic English proficiency, and item linguistic features can introduce construct-irrelevant variance (Abedi, 2002, 2008b; Abedi & Gándara, 2006; Leighton & Gokiart, 2005; Martiniello, 2008; Solorzano, 2008). Text features (such as lexical density and polysemous words) affect an item's comprehensibility and can increase cognitive load for ELLs (Hale, 2003; Leighton & Gokiart, 2005; Sweller & Chandler, 1991). Cognitive load is affected by several factors, including the content and the number of items that require attending (Shank, 2007), and if one factor, such as item linguistic complexity, is more difficult, less attention can be dedicated to other factors such as content (Dickey, 2004).

Although other factors such as poverty and parent education impact ELL achievement, language has the greatest impact, with increasing gaps as language demands increase (Abedi, 2002, 2008b, 2009; Abedi & Gándara, 2006; Abedi et al., 2004). As discussed in Chapter 2, Martiniello (2008) found that measures of lexical density and syntactic complexity could estimate item difficulty for ELLs on the Grade 4 Mathematics MCAS (see also Abedi & Gándara, 2006; Abedi & Lord, 2001), and Abedi and Hejri (2004) found that the ELL achievement gap on the NAEP Grade 8 science assessment widened when items became linguistically complex. Although Huff, Egan, Gaines, and Ferrara (2013) found that reading load was not a consistent predictor of item

difficulty for Grade 4 and Grade 8 mathematics items, the authors believed that more research is needed since their results were inconsistent with an earlier study.<sup>109</sup>

Martiniello (2008) also called for further studies to examine linguistic complexity and differential item functioning for ELLs from all language backgrounds as well as for studies that account for ELL proficiency differences.<sup>110</sup>

This study explored ELL performance on the item attribute of linguistic complexity. Linguistic complexity was operationalized in Phase I. Three linguistic variables—total lexical density (TLD), stem syntactic density (SSD), and reading complexity score (RCS)—were normalized and added to compute a new variable, composite linguistic complexity (CLC). Multiple-choice items were grouped into low CLC, medium CLC, and high CLC based on tertiles, and ELL performance was explored for each of these three groups of linguistic complexity.<sup>111</sup> Table 4.48 summarizes measures of central tendency. The mean percent correct was less than 50% for all three levels of linguistic complexity: low CLC ( $M = 49.59$ ;  $SD = 20.27$ ); medium CLC ( $M = 44.26$ ;  $SD = 20.82$ ); and high CLC ( $M = 44.98$ ;  $SD = 19.97$ ). As item linguistic complexity increased from low CLC to medium CLC, the mean percent correct decreased; however, ELLs performed similarly on items of medium CLC and high CLC with less than a 1% difference in the mean percent correct.

---

<sup>109</sup> The current study used an item's estimated Lexile® measure as the reading load.

<sup>110</sup> Martiniello (2008) acknowledged that a limitation of her study was that it only explored differential item functioning for Spanish-speaking ELLs.

<sup>111</sup> See Appendix I for items classified as low, medium, and high CLC.

Table 4.48  
*Percent Correct by Item Linguistic Complexity*

ELLs (n = 3,315)			
<u>Composite Linguistic Complexity</u>	<u>n</u>	<u>M</u>	<u>SD</u>
Low CLC	13	49.59	20.27
Medium CLC	14	44.26	20.82
High CLC	13	44.98	19.97
	40		

The data suggested that ELLs in general may have found items with medium CLC difficult just as difficult as items of high CLC. As expected, ELL performance was highest on items with low linguistic complexity. The similar performance on items with medium and high linguistic complexity, however, was unexpected. Second language acquisition theories, especially Krashen's comprehensible input hypothesis, and cognitive load theories support decreased ELL performance as linguistic complexity increases. One possible explanation is that 23.1% (n = 3) of the high CLC items were at the foundational cognitive skill level compared to only 7.7% (n = 1) of the medium CLC items.<sup>112</sup> Thus, the lesser cognitive load of the foundational level might have mitigated the increased cognitive load from high CLC so that performance did not decrease between medium and high CLC.

Another possible explanation is that the data support Robinson's cognition hypothesis. The data suggested that higher linguistic complexity items increased noticing and attending to language, which led to increased item comprehension and performance. The finding is also consistent with Leighton and Gokiart's (2005) belief that students can

<sup>112</sup>For both medium CLC and high CLC, there were a similar number of items for both the conceptual and application cognitive skill levels (Table 4.19).

access the knowledge and skills for a correct answer even in the absence of full comprehension of the item. Thus, even though the ELLs are by definition not English proficient, they were able to attend to the higher linguistic demands and perform just as well on high CLC items as they did on medium CLC items. The data, however, were inconsistent and did not suggest a similar increased attending effect when moving from low CLC items to medium CLC items. One possible explanation is what can be thought of as a Goldilocks effect. The ELLs recognized low CLC items as being linguistically simple, and their noticing was not heightened because the language was less of a barrier in demonstration of content performance. When ELLs encountered high CLC items, they recognized that the items were more difficult to comprehend, so they attended more closely to the language of the item. The medium CLC items, however, may have seemed not linguistically simple but also not linguistically complex. The ELLs may have thought the language of medium CLC was “just right” for their English proficiency level, and without the heightened noticing and attending to language, performance was lower than on low CLC items and the same as on high CLC items with increased noticing. This finding should be explored further in future studies.

**English proficiency impact.** As discussed earlier, this study found that English proficiency level contributed 29% of the variance among ELLs on overall Biology MCAS score. The ELL sample was disaggregated by MEPA levels to explore the impact of English proficiency on performance for the three levels of item linguistic complexity. Table 4.49 summarizes measures of central tendency. At the lower end of English proficiency, MEPA Levels 1 and 2 had similar performance for the three levels of item linguistic complexity; mean percent correct ranged from 35% to 39.5%. MEPA Level 3

performed slightly better, with mean percent correct ranging from 40% (medium CLC) to 46.5% (low CLC); the data suggested that medium CLC items were slightly more difficult than high CLC for Level 3 ELLs. MEPA Level 4 had increased performance, with mean percent correct of approximately 50% for medium and high CLC and 56% for low CLC items.<sup>113</sup> Moving from Level 4 to Level 5, the mean percent correct was approximately 10% higher for each level of item linguistic complexity. The data were consistent with the finding that performance increased at successively higher levels of English proficiency.

Table 4.49  
*Linguistic Complexity Percent Correct by English Proficiency Levels*

Linguistic Complexity	English Proficiency				
	Level 1 n = 276 (SD)	Level 2 n = 441 (SD)	Level 3 n = 1,441 (SD)	Level 4 n = 603 (SD)	Level 5 n = 554 (SD)
Low	38.16 (17.96)	39.51 (17.17)	46.50 (17.82)	55.96 (19.31)	64.40 (20.10)
Medium	35.90 (17.35)	35.57 (15.89)	40.04 (18.05)	49.59 (21.47)	60.55 (21.38)
High	34.95 (16.66)	35.44 (16.17)	42.27 (16.75)	49.76 (18.86)	59.43 (19.06)

Within MEPA Level 4 and Level 5, where the majority of ELLs passed the June 2012 Biology MCAS, there was no decrease in performance between medium and high CLC items. As discussed previously, cognitive load theory and Krashen's comprehensible input hypothesis suggest that items with higher linguistic demands would have decreased performance; however, the data did not support this. As discussed

<sup>113</sup> MEPA Level 4 mean percent correct for medium and high linguistic complexity rounded to 50%.

previously, the data suggested that increased noticing and attending may have leveled linguistic complexity challenges between medium CLC items and high CLC items.

One-way ANOVA analyses confirmed that English proficiency had a statistically significant impact on performance on the item attribute of linguistic complexity. The data suggested that the impact of English proficiency was similar at all three level of item linguistic complexity: low CLC,  $F(4, 3310) = 177.60, p < .001$ ; medium CLC,  $F(4, 3310) = 167.96, p < .001$ ; and,  $F(4, 3310) = 170.25, p < .001$ . This was unexpected since English proficiency should have higher impact as linguistic complexity increases (i.e., items with high linguistic complexity require higher levels of English proficiency for access). Scheffé post hoc analyses were conducted, and Table 4.50 summarizes the results. The Scheffé post hoc analyses determined that there was no statistically significant difference in performance for MEPA Levels 1 and 2. This was consistent with the findings that there was indistinguishable performance between these levels on overall MCAS performance, content domain performance, and cognitive skill level performance. The Scheffé post hoc analyses further confirmed that there was a statistically significant increase in mean percent correct between all other adjacent MEPA levels for items at each of the three linguistic complexity levels. The data suggested that items with low linguistic complexity approached accessibility at MEPA Level 3; however, it was at Level 4 that the mean percent correct was greater than 50%. For items with medium and high linguistic complexity, the data suggested that Level 4 was the cusp of accessibility, and these items were generally accessible for Level 5.

Table 4.50

*Summary of Scheffé Post Hoc Analysis on Linguistic Complexity Percent Correct*

	(I) MEPA Level	(J) MEPA Level	Mean Difference (I-J)
Low CLC	Level 1	Level 2	-1.35
	Level 2	Level 3	-6.99***
	Level 3	Level 4	-9.47***
	Level 4	Level 5	-8.44***
Medium CLC	Level 1	Level 2	0.33
	Level 2	Level 3	-4.47**
	Level 3	Level 4	-9.54***
	Level 4	Level 5	-10.96***
High CLC	Level 1	Level 2	-0.49
	Level 2	Level 3	-6.82***
	Level 3	Level 4	-7.50***
	Level 4	Level 5	-9.66***

\*p < .05, \*\*p < .01, \*\*\*p < .001

Simple linear regression analyses were used to test to what extent the independent variable English proficiency (MEPA score) significantly predicted performance on items with: (1) low linguistic complexity, (2) medium linguistic complexity, and (3) high linguistic complexity. The results of the linear regression analyses are summarized in Table 4.51. The models for performance at each of the three levels of linguistic complexity emerged as statistically significant: (1) low CLC,  $F(1, 3313) = 760.60$ ,  $p < .001$ ; (2) medium CLC,  $F(1, 3313) = 583.80$ ,  $p < .001$ ; and (3) high CLC,  $F(1, 3313) = 658.30$ ,  $p < .001$ . The model for low CLC performance had an  $R^2$  value of .187, accounting for approximately 19% of the variation among ELLs. The medium CLC performance model had an  $R^2$  value of .150, accounting for 15% of the variation among ELLs, and the high CLC an  $R^2$  value of .166, accounting for approximately 17% of the variation among ELLs. English proficiency was a strong predictor of performance at all



three levels of linguistic complexity:  $B = .42$  for low CLC;  $B = .39$  for medium CLC; and  $B = .38$  for high CLC. The  $B$  values indicated that for each level of linguistic complexity, an additional 5-point increase in MEPA score would increase the mean percent correct by approximately 2% according to the following models:

$$\% \text{ Correct Low CLC} = .42(\text{MEPA score}) - 153.34$$

$$\% \text{ Correct Medium CLC} = .39(\text{MEPA score}) - 421.38$$

$$\% \text{ Correct High CLC} = .38(\text{MEPA score}) - 135.85$$

These findings were unexpected. The ELLs in the sample were not English proficient, and as discussed in Chapter 2, the construct of academic language has multiple lexical, syntactic, and discourse-level elements. Since second language and cognitive load theories support the prediction that increased linguistic complexity would decrease item accessibility, the similar performance for medium CLC and high CLC items needs further investigation.

Table 4.51

*Summary of Linear Regression between English Proficiency and Item Linguistic Complexity Performance*

	<u>N</u>	<u>Mean</u>	<u>SD</u>	<u><math>\beta</math></u>	<u>t</u>
% Correct Low CLC	3,315	49.59	20.27	.43***	-20.82***
English proficiency	3,315	480.09	20.73		
R <sup>2</sup>	.187				
F	760.60***				
% Correct Low CLC = .42(MEPA score) – 153.34					
% Correct Medium CLC	3,315	44.26	20.82	.39***	-18.41***
English proficiency	3,315	480.09	20.73		
R <sup>2</sup>	.150				
F	583.80***				
% Correct Medium CLC = .39(MEPA score) – 421.38					
% Correct High CLC	3,315	44.98	19.17	.41***	-19.26***
English proficiency	3,315	480.09	20.73		
R <sup>2</sup>	.166				
F	658.30***				
% Correct High CLC = .38(MEPA score) – 135.85					

Note: \*p < .05, \*\*p < .01, \*\*\*p < .001.

To further explore the relationship between English proficiency and performance at the three levels of item linguistic complexity, the simple regression analysis was repeated for the subgroup of ELLs who had a MEPA score of 464 and above (MEPA Levels 3 to 5). The results of the regression analyses are summarized in Table 4.52. The three models emerged as statistically significant: (1) low CLC,  $F(1, 2596) = 518.96$ ,  $p < .001$ ; (2) medium CLC,  $F(1, 2596) = 505.46$ ,  $p < .001$ ; and (3) high CLC,  $F(1, 2596) = 441.25$ ,  $p < .001$ . The model for low CLC performance had an  $R^2$  value of .167, accounting for approximately 17% of the variation among ELLs. The medium CLC performance model had an  $R^2$  value of .163, accounting for approximately 16% of the variation among ELLs, and the high CLC an  $R^2$  value of .145, accounting for

approximately 14% of the variation among ELLs. Compared to the whole sample, English proficiency was a slightly stronger predictor of performance at all three levels of linguistic complexity for ELLs who were at MEPA Levels 3 to 5:  $B = .56$  for low CLC;  $B = .59$  for medium CLC; and  $B = .50$  for high CLC. The  $B$  values indicated that for each level of linguistic complexity, an additional 2-point increase in MEPA score would increase the mean percent correct by approximately 1% according to the following models:

$$\% \text{ Correct Low CLC} = .56(\text{MEPA score}) - 222.10$$

$$\% \text{ Correct Medium CLC} = .59(\text{MEPA score}) - 242.95$$

$$\% \text{ Correct High CLC} = .50(\text{MEPA score}) - 196.09$$

Table 4.52

*Summary of Linear Regression between English Proficiency and Item Linguistic Complexity Performance for MEPA Levels 3 to 5*

	<u>N</u>	<u>Mean</u>	<u>SD</u>	<u><math>\beta</math></u>	<u>t</u>
% Correct Low CLC	2,598	52.51	20.02	.41	-18.42***
English proficiency	2,598	488.20	14.52		
$R^2$	.167				
F	518.96***				
% Correct Low CLC = $.56(\text{MEPA score}) - 222.10$					
% Correct Medium CLC	2,598	46.63	21.27	.40	-18.41***
English proficiency	2,598	488.20	14.52		
$R^2$	.163				
F	505.46***				
% Correct Medium CLC = $.59(\text{MEPA score}) - 242.95$					
% Correct High CLC	2,598	47.67	19.03	.38	-16.89***
English proficiency	2,598	488.20	14.52		
$R^2$	.145				
F	441.25***				
% Correct High CLC = $.50(\text{MEPA score}) - 196.09$					

Note: \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

When the ELL sample was disaggregated into a subgroup of MEPA Levels 3 to 5, the data suggested that the impact of English proficiency decreased slightly between medium CLC items and high CLC items. This decrease, however, was not enough to affect the general model for all three ELP levels of a 2-point increase in MEPA score would increase mean percent correct at each linguistic complexity level by approximately 1%. As discussed previously and in Chapter 5, these findings warrant further investigation in future studies.

**First language family impact.** The sample was disaggregated by first language family to explore the impact of a Latinate or non-Latinate L1 on performance across item linguistic complexity levels. Table 4.53 and Figure 4.19 summarize measures of central tendency. For all three levels of item linguistic complexity, the non-Latinate L1 group had a higher mean percent correct. Performance was highest on items of low linguistic complexity for both subgroups, and within each subgroup, there was similar performance for medium and high CLC items. Latinate L1 ELLs had almost the same performance on medium CLC and high CLC items ( $M = 42.55$ ,  $SD = 19.59$ ; and  $M = 42.93$ ,  $SD = 18.11$ , respectively). Non-Latinate ELLs also had similar performance on medium CLC items ( $M = 48.90$ ,  $SD = 23.21$ ) and high CLC items ( $M = 50.54$ ,  $SD = 20.81$ ).

Table 4.53

*Linguistic Complexity Percent Correct by First Language Family*

Linguistic Complexity	Latinate (n = 2,420)		Non-Latinate (n = 895)		t	Cohen's d	Partial $\eta^2$
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>			
Low	47.81	19.17	54.40	22.31	-7.83***	-.32	.021
Medium	42.55	19.59	48.90	23.21	-7.28***	-.30	.018
High	42.93	18.11	50.54	20.81	-9.67***	-.39	.031

\*p < .05, \*\* p < .01, \*\*\*p < .001

The data suggested that ELLs with a non-Latinate L1 generally performed better than ELLs with a Latinate L1 at all levels of linguistic complexity. It further suggested that within each subgroup, performance was best on low CLC items, but there appeared to be no difference in performance between medium CLC and high CLC items.

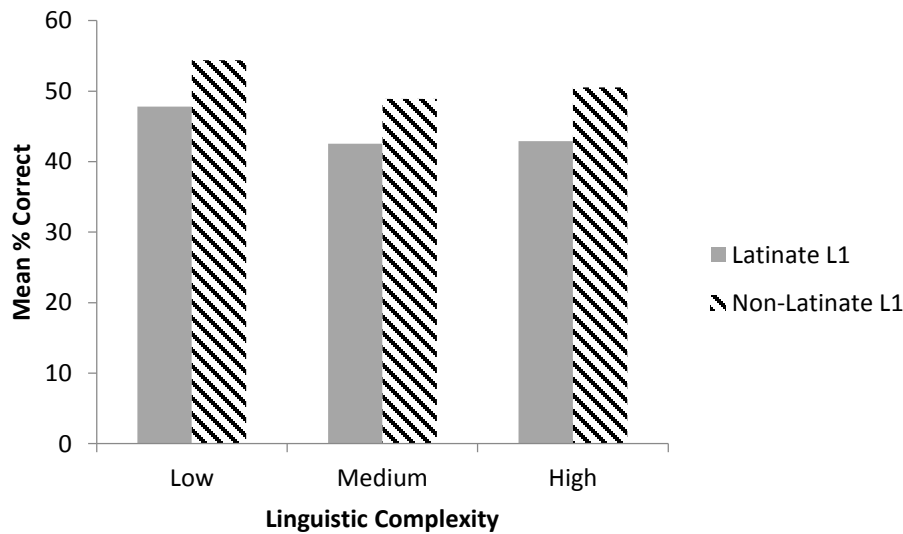


Figure 4.19. Mean Percent Correct for Item Linguistic Complexity by First Language Family. ELLs with a non-Latinate L1 performed better than ELLs with a Latinate L1 at each level of item linguistic complexity.

Independent samples t-tests confirmed that non-Latinate L1 ELLs had statistically significant better performance at each linguistic complexity level. Table 4.53 summarizes these results. Non-Latinate L1 ELLs performed better on items of low linguistic complexity, which had the highest mean percent correct for both groups,  $t(1410) = -7.83$ ,  $p < .001$ . The same was true for medium linguistic complexity items,  $t(1391) = -7.28$ ,  $p < .001$ , and high linguistic complexity items,  $t(1424) = -9.67$ ,  $p < .001$ . Cohen's  $d$  determined that L1 language family had a small effect size, and univariate analyses of variance determined that 2.1% of the variance in performance on low CLC items, 1.8% on medium CLC items, and 3.1% on high CLC items was attributed to having a Latinate or a non-Latinate L1. Although the practical significance was small, the finding that L1 language family contributed the most variance on high CLC item performance (3.1%) and the least on medium CLC item performance (1.8%) was noteworthy. The data suggested that when item linguistic complexity was high, ELLs with a non-Latinate L1 generally performed better. These results appear to support the cognition hypothesis prediction that increased task complexity leads to increased attending and higher L2 output for this subgroup; the increased linguistic complexity may have highlighted the L1 to L2 distance. In terms of the Biology MCAS, the lack of L1-to-English cognates and increased syntactic complexity would have increased cognitive load for non-Latinate L1 ELLs, which may have resulted in more noticing and attending to the item's linguistic features and increased performance on high linguistic complexity items.<sup>114</sup> This might explain the higher performance for non-Latinate L1 ELLs at all levels of linguistic

---

<sup>114</sup> Depending on the non-Latinate L1, increased syntactic complexity could have heightened awareness of the L1-to-L2 syntactic distance.

complexity. It might also explain the finding that within each subgroup, there was no decrease in performance between medium and high CLC items. A similar effect may have happened for Latinate L1 ELLs who did not know the L1-to-English cognates. As discussed in Chapter 5, future research into the Latinate/non-Latinate L1 performance gap should account for L1 literacy, L1 schooling, and L1 content knowledge.

**First language orthography impact.** The sample was disaggregated by first language orthography to explore the impact of an alphabetic or non-alphabetic L1 on performance across item linguistic complexity levels. Table 4.54 and Figure 4.20 summarize measures of central tendency. The performance data were similar to those discussed previously for the Latinate/non-Latinate L1 subgroups. Alphabetic L1 ELLs had a mean percent correct less than 50% for all levels of linguistic complexity. In comparison, ELLs with a non-alphabetic L1 had a mean percent correct higher than 50% for two of the three linguistic complexity levels: low CLC (54.1%) and medium CLC (50.7%). Performance was highest on items of low linguistic complexity for both groups. ELLs with an alphabetic L1 performed nearly the same on items with medium CLC ( $M = 43.39$ ,  $SD = 20.15$ ) and items with high CLC ( $M = 43.74$ ,  $SD = 18.43$ ). ELLs with a non-Latinate L1 also had similar performance on items with medium CLC ( $M = 48.31$ ,  $SD = 23.24$ ) and high CLC ( $M = 50.71$ ,  $SD = 21.38$ ).

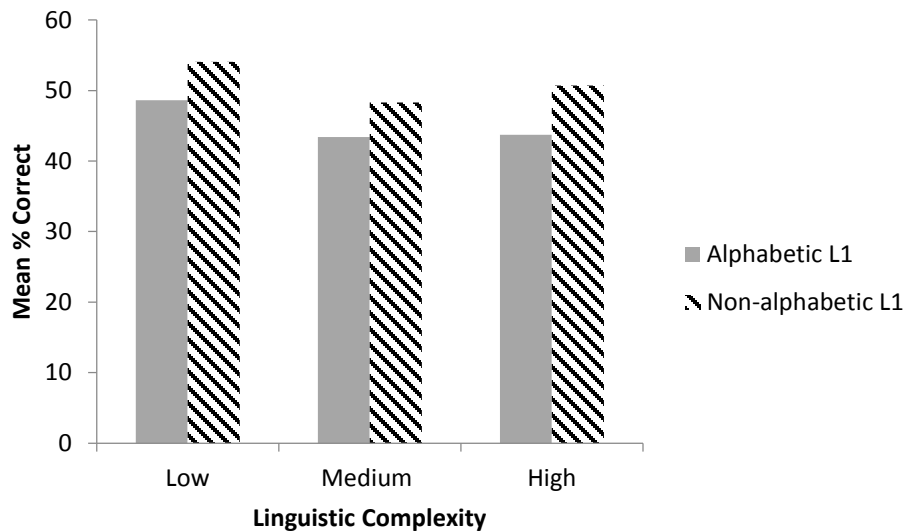
Table 4.54

*Linguistic Complexity Percent Correct by First Language Orthography*

<u>Linguistic Complexity</u>	Alphabetic (n = 2,724)		Non- alphabetic (n = 591)		<u>t</u>	Cohen's <u>d</u>	Partial <u>eta</u> <sup>2</sup>
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>			
Low	48.61	19.61	54.08	22.58	-5.46***	.26	.011
Medium	43.39	20.15	48.31	23.24	-4.77***	.23	.008
High	43.74	18.43	50.71	21.38	-7.36***	.35	.019

\*p < .05, \*\* p < .01, \*\*\*p < .001

The data suggested that both groups generally had higher performance on items with low CLC and that ELLs with a non-alphabetic L1 generally performed better than ELLs with an alphabetic L1 at all levels of linguistic complexity.



*Figure 4.20. Mean Percent Correct for Item Linguistic Complexity by First Language Orthography. ELLs with a non-alphabetic L1 performed better than ELLs with an alphabetic L1 at each level of item linguistic complexity.*



Independent samples t-tests confirmed that there were statistically significant differences in mean percent correct between L1 orthography subgroups. Table 4.54 summarizes the results of these analyses. The results indicated that ELLs with a non-alphabetic L1 performed better at each level of linguistic complexity: low CLC,  $t(794) = -5.46$ ,  $p < .001$ ; medium CLC,  $t(794) = -4.77$ ,  $p < .001$ ; and high CLC,  $t(791) = -7.36$ ,  $p < .001$ . Cohen's  $d$  determined that L1 orthography had a small effect size, and univariate analyses of variance determined that L1 orthography contributed 1.1% of the performance variance on low CLC items, 0.8% on medium CLC items, and 1.9% on high CLC items. The data suggested that L1 orthography impact was negligible for item linguistic complexity performance; however, the pattern was similar to that for L1 language family with approximately the same impact for low and medium CLC items but increased impact for high CLC items. The finding that L1 orthography impact was the greatest when items had high linguistic complexity calls for further study on L1 orthography and cognitive load in the context of standardized assessments.

**Late-entry ELL status impact.** The ELL sample was disaggregated into late-entry and not late-entry ELLs to explore the impact of age of entry on performance across levels of item linguistic complexity. Measures of central tendency are summarized in Table 4.55. Both groups had mean percent correct less than 50% and similar performance, except for late-entry ELLs performing slightly better on items of medium linguistic complexity.

Table 4.55

*Linguistic Complexity Percent Correct by Late-Entry ELL Status*

<u>Linguistic Complexity</u>	Late-entry ELL (n = 2,218)		Not-late-entry ELL (n = 1,097)		<u>t</u>	Cohen's <u>d</u>	Partial <u>eta</u> <sup>2</sup>
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>			
Low	49.67	20.91	49.41	18.92	0.34	.013	
Medium	45.15	21.00	42.47	20.33	-3.54***	.130	.004
High	45.03	19.85	44.88	17.73	0.21	.008	

\*p < .05, \*\* p < .01, \*\*\*p < .001

Independent samples t-tests confirmed that late-entry ELL status only had a statistically significant impact on performance for items with medium CLC. Table 4.55 summarizes the results of these analyses. For items with medium CLC, late-entry ELLs had a higher mean percent correct,  $t(2249) = -3.54$ ,  $p < .001$ ; however, the effect size was negligible (Cohen's  $d = 0.13$ ) with less than 1% of the variance attributed to late-entry ELL status. The data suggested that late-entry ELL status was not a predictive factor in ELL performance on item linguistic complexity. This was an interesting finding because late-entry ELLs should have full L1 acquisition and higher meta-linguistic awareness, which theoretically could be an advantage when encountering complex syntax in the L2 (see Snow & Hoefnagel-Höhle, 1978). As discussed in Chapter 5, future studies should explore L1 literacy and L1 schooling experiences when studying the impact of age of entry on MCAS performance.

**Summary of linguistic complexity performance.** Performance on items at the three levels of linguistic complexity exhibited normal distributions, but there was a slightly exaggerated peak just below the mean for medium linguistic complexity. For the

ELL sample as a whole the mean percent correct for each level of item linguistic complexity (low, medium, and high) was below 50%. In general, ELLs had the best performance on items of low linguistic complexity. Instead of the expected pattern of performance decreasing as linguistic complexity increases, ELLs appeared to have similar performance for both medium and high linguistic complexity items. This finding appeared to support the cognition hypothesis that higher linguistic complexity increases noticing and leads to higher performance; however, more research is needed.

English proficiency had statistically and practically significant impact on performance with respect to item linguistic complexity. Simple linear regression analyses indicated that English proficiency could generally predict performance on items at different levels of linguistic complexity, and it accounted for 15% to 19% of the variance among ELLs. The impact of first language characteristics was small or negligible, with the greatest impact for items of high linguistic complexity. First language family contributed between 1.8% and 3.1% of the variance with non-Latinate L1 ELLs favored. First language orthography impact was less, and it contributed from less than 1% to 1.9% of the variance with non-alphabetic L1 ELLs. Late-entry ELL status appeared to have no statistically significant impact for performance on items of low or high linguistic complexity and a negligible effect size for items of medium linguistic complexity.

### **Summary of Results**

**ELL Biology MCAS performance.** Approximately half (52.8%) of the ELLs passed the June 2012 Biology MCAS; however, only 14.9% scored at the Proficient or higher performance levels. ELLs appeared to do best on ecology and anatomy and physiology items, where they had a mean percent correct of 54.1% and 50.4%,

respectively. The lowest performance was seen on cell biology items (40.4%), followed by genetics items (41.5%). Performance on items at different cognitive skills had the expected highest performance on foundational items (49%), but there was similar performance on conceptual items (46.1%) and application items (45.3%). Performance on items at different levels of linguistic complexity, however, did not follow the expected pattern that performance would decrease as linguistic complexity increased. As expected, ELLs performed best on items with low linguistic complexity (49.6%); however, the data indicated only a minor difference of less than 1% between items with medium linguistic complexity (44.3%) and high linguistic complexity (45%).

**English proficiency impact.** When the ELL sample was disaggregated by English proficiency levels, performance differences emerged. As ELLs became more fluent in English, including academic English, their performance on the June 2012 Biology MCAS increased. ELLs at MEPA Levels 4 and 5 had a passing rate of 79.4%, with 31.7% scoring Proficient or higher. ELLs at MEPA Level 5 had a passing rate of 87.7%, with 43.3% scoring Proficient or higher. English proficiency had the greatest impact on performance, and a positive linear relationship emerged after MEPA Level 3, the level where ELLs begin using academic language. A simple linear regression analysis found that English proficiency accounted for 29% of the variance in performance among the five MEPA Levels. The data also indicated that there were no statistically significant differences in performance between MEPA Level 1 and Level 2.

The same pattern of English proficiency impact was seen on performance on the six content domains. The data indicated that there were no statistically significant differences in performance between MEPA Level 1 and Level 2. Moving from MEPA

Level 2 to Level 3 had an increase in performance for all content domains, and as English proficiency increased through successive levels, performance on all content domains increased. Worth noting were the similarities in performance on the six content domains irrespective of English proficiency (see Figure 4.14). Regardless of English proficiency, performance was highest on ecology items, followed by performance on anatomy and physiology items. The effect sizes for these two domains were negligible, and English proficiency contributed less than 1% of the variance between groups. In contrast, English proficiency had greater impact on the two domains that had the lowest performance, genetics and cell biology. It appeared that for the lower levels of English proficiency (MEPA Levels 1 to 3), genetics items proved the most difficult. For higher levels of English proficiency (MEPA Levels 4 and 5), genetics items still appeared difficult, with second-to-worst performance, but these higher levels of English proficiency had the lowest performance on cell biology items. The effect size for English proficiency approached medium for genetics items and contributed 4.2% of the variance between groups. The effect size for English proficiency was small to medium for cell biology items and contributed 3.4% of the variance between groups. English proficiency had a small effect size on the remaining two domains, biochemistry and evolution. English proficiency contributed 2.1% of variance on biochemistry items and 1.1% of the variance on evolution items. The data suggested that English proficiency had the greatest impact on genetics and cell biology, the two domains that were the most difficult for ELLs.

The impact of English proficiency on performance at the different cognitive skill levels was statistically significant. For each cognitive skill level, performance increased as English proficiency increased; however, post hoc analyses indicated that there were no

statistically significant differences between MEPA Levels 1 and 2. The data suggested that English proficiency had greater impact beyond the foundational level. Further analyses indicated that English proficiency contributed 11.7% of the variance on foundational items, 18.8% of the variance on conceptual items, and 18.1% of the variance on application items. The impact of English proficiency on performance at different levels of item linguistic complexity was also statistically significant. As with its impact on cognitive skill level performance, post hoc analyses indicated that English proficiency had no statistically significant difference between MEPA Levels 1 and 2. Simple linear regression analyses determined that English proficiency contributed 19% of the performance difference among MEPA levels for items with low linguistic complexity, 15% of the performance difference for items with medium linguistic complexity, and 17% for items with high linguistic complexity. Similar English proficiency impact was also seen for the subgroup of ELLs at MEPA Levels 3 to 5, though there was a slightly smaller contribution (16%) for high CLC items. The data suggested that a 2- to 2.5-point increase in MEPA score would increase mean percent correct by approximately 1% for all three levels of item linguistic complexity. English proficiency had both statistical and practical significance for ELL performance on the item attributes of cognitive skill, linguistic complexity, and for two content domains (genetics and cell biology).

**First language family impact.** The disaggregation of the sample by a Latinate or non-Latinate L1 indicated a statistically significant performance gap between these two groups on overall Biology MCAS performance. The data indicated that the L1 language family contributed 4% of the variance with non-Latinate L1 ELLs favored; however, when looking only at ELLs at MEPA Levels 3 to 5, L1 language family contributed 4.9%

of the variance. This performance gap persisted when exploring performance by item attributes. ELLs with a non-Latinate L1 had statistically significant higher performance on each of the six content domains. The practical significance, however, was negligible or small except for cell biology and genetics, where the effect sizes approached medium and L1 language family accounted for 3.4% and 4.2%, respectively, of the variance between the groups. ELLs with a non-Latinate L1 also had statistically significant higher performance on items at each of the three cognitive skill levels. First language family contributed 1.8% to 2.5%, with non-Latinate L1 ELLs favored; however, the practical significance was small. Likewise, ELLs with a non-Latinate L1 had statistically significant higher performance at each of the three levels of item linguistic complexity. The greatest contribution to performance difference was on items with high linguistic complexity (3.1%), followed by low linguistic complexity (2.1%), and then by medium linguistic complexity items (1.8%). The data suggested that for item attributes, a non-Latinate L1 had the greatest practical impact for cell biology items (3.4%), genetics items (4.2%), and items with high linguistic complexity (3.1%).

**First language orthography impact.** The data also indicated a performance gap when the sample was disaggregated by L1 orthography. On overall MCAS performance, the data indicated statistically significant higher performance for ELLs with a non-alphabetic L1 compared to ELLs with an alphabetic L1. The data indicated that L1 orthography contributed 2% of the difference between groups. Thus, practical significance of L1 orthography was less than that of first language family. First language orthography also appeared to have less impact than L1 language family on content domain performance. The data indicated that L1 orthography had a small effect size on

two domains—cell biology and genetics—where it contributed 2.4% and 2.8%, respectively, to the variance between groups. With respect to item cognitive skill, ELLs with a non-alphabetic L1 had statistically significant higher performance on items at all three cognitive skill levels (foundational, conceptual, and application); however, it had little practical impact since it contributed approximately 1% of the variance. Likewise, a non-alphabetic L1 had a minor impact on linguistic complexity performance and contributed 1.9% of the difference on high linguistic complexity items, followed by 1.1% on low linguistic complexity items, and 0.8% on items with medium linguistic complexity. The data suggested that for the item attributes of content domain, cognitive skill level, and linguistic complexity, first language orthography had less impact than first language family.

**Late-entry ELL status impact.** The data suggested that late-entry ELL status did not have an impact on overall Biology MCAS performance. There was either no impact or negligible impact on content domain performance. The data further suggested that late-entry ELL status had no impact on cognitive skill level performance. For performance at the three levels of linguistic complexity, the data suggested no impact or a negligible impact. Late-entry ELL status was a predictor of neither overall Biology MCAS performance nor performance on the item attributes of content domain, cognitive skill, and linguistic complexity.



## CHAPTER 5

### DISCUSSION

This study analyzed the June 2012 Biology MCAS performance for ELLs who had also taken the March 2012 MEPA; 3,315 records met these criteria. It is difficult to match the ELL population of this study to accountability and MCAS reports because of differing reporting criteria. The technical report for the 2012 MCAS exams reported 1,990 ELL test-takers (Appendix B; MA DESE, 2013); however, calculations do not include students who enrolled after October 1 (MA DESE, 2013c, p. 7). Although all ELLs in Grades 10, 11, and 12 who have not passed an STE MCAS must take an exam regardless of time in the United States, scaled scores and achievement levels are not reported if the students are first-year ELLs. In this study, 736 (22.2%) were reported to be in their first year of U.S. schooling.<sup>115</sup> For purposes of MCAS participation, reporting, and accountability, ELLs are considered first-year if they entered after March 1 of the previous school year, and the school site bubbled this designation on the student's test booklet (MA DESE, 2013c).

In addition, statewide 2012 MCAS results were reported for the Class of 2014 and “based on students’ best performance on any STE test taken in grade 9 or grade 10; only

---

<sup>115</sup> See Chapter 3 for how these raw scores were converted into scaled scores to determine performance levels.

students continuously enrolled in Massachusetts public schools from fall of grade 9 through spring of grade 10 are included” (MA DESE, 2012i, p. 34). Thus, ELLs who took the June 2012 Biology MCAS but were not members of the Class of 2014 or who were not in Massachusetts schools for both Grade 9 and Grade 10 were not included in this report of statewide results. More than half of the ELLs (56.8%, n = 1,883) who took the June 2012 Biology MCAS were not in Grade 10, and an additional 302 (9.1%) were Grade 10 ELLs who were in Massachusetts schools for one year.<sup>116</sup> Therefore, this study’s results may differ from those of other MCAS reports, and caution should be used in trying to equate findings that use different sample criteria.

### **Summary of Findings**

This study analyzed secondary ELL performance on the June 2012 Biology MCAS to explore the nature of the achievement gap between ELLs and non-ELLs. The data indicated that reporting performance for the ELL subgroup as a single statistic masked performance gains seen at the higher levels of English proficiency, where the overwhelming majority of ELLs passed. English proficiency at MEPA Level 3 was the cusp, but meaningful participation in the Biology MCAS became evident at MEPA Levels 4 and 5. English proficiency was a strong predictor of ELL Biology MCAS performance, and statistically significant models emerged in which English proficiency contributed 29% of the variance among ELLs at all MEPA levels and 25% of the variance among ELLs at MEPA Levels 3 to 5. The data further indicated that MEPA Levels 1 and 2 had indistinguishable performance. It appeared that although there were

---

<sup>116</sup> There were 679 ELLs in Grades 11 and 12, which would have included ELLs who had failed a previous Biology MCAS and had a retest status. These students were not part of the 2014 cohort reported results.

entry points for all students at the item level, these entry points were insufficient to make the instrument, as a whole, accessible for Level 1 and Level 2 ELLs to demonstrate their content knowledge. Disaggregation by first language characteristics identified a persistent performance gap for secondary ELLs who had a Latinate first language regardless of English proficiency and, to a lesser extent, a gap for ELLs with an alphabetic first language. The disaggregation of late-entry ELLs determined that age of entry was not a predictor of Biology MCAS performance; however, the similar performance of this subgroup to not-late-entry ELLs, who had a mean MEPA scaled score one level higher, suggested one or more mitigating factors.

Analyses of ELL performance on each of the six content domains indicated that ELLs generally found the same content domains easy or difficult as English proficient test-takers; however, differences emerged when results were disaggregated by English proficiency. Consistent with the data on overall MCAS performance, the content domain performance was indistinguishable between MEPA Levels 1 and 2. The higher levels—that is, MEPA Levels 4 and 5—had the same ordinal ranking of domain performance as non-ELLs; MEPA Level 3 performance was in between, with some domain performance rankings the same as the lower levels and some the same as the higher levels of English proficiency. The data further confirmed the construct-irrelevant variance introduced by English proficiency. Within each content domain, the mean percent correct increased as ELLs moved to the next higher level of English proficiency, except for indistinguishable performance between MEPA Levels 1 and 2.

The ELL performance data for item cognitive skill level and item linguistic complexity had limited support for cognitive load theory and Robinson's (2005)

cognition hypothesis predictions. As expected from cognitive load theory, ELLs performed best on items at the lowest cognitive skill level and the lowest level of item linguistic complexity irrespective of English proficiency, first language characteristics, or late-entry ELL status. Performance at intermediate and high levels of cognitive skill and linguistic complexity were similar, which did not support cognitive load theory's expected decrease in performance as cognitive load increases. This data, however, supported the cognition hypothesis, according to which increased task complexity leads to increased performance from increased noticing and attending. This was also consistent with Krashen's monitor hypothesis, though in the context of the current study, reading comprehension, not language output, was monitored. The data further suggested a Goldilocks effect, whereby it was only at the high levels of cognitive skill or linguistic complexity that increased noticing and attending mitigated increased task complexity.

### **Sample Characteristics**

Statewide MCAS results are reported for a particular Grade 10 cohort, the Class of 2014 in the case of the 2012 high school MCAS exams. ELLs who took the June 2012 Biology MCAS differed from the Class of 2014 not only in English proficiency but also in age. As reported in Chapter 4, 38.3% ( $n = 1,273$ ) of the ELLs who took the June 2012 Biology MCAS were the age of 17 years or older—older than the usual age range for students in Grade 9 and Grade 10 (i.e., 14 to 16 years); however, the majority (54.8%,  $n = 697$ ) of these older ELLs were in Grades 9 and 10. The grade-level placement of newly enrolled secondary ELLs is determined by district-level policies, and the data showed that age did not appear to be the determining factor in placing ELLs in a grade.

Descriptive analyses on the number of school years spent by students in Massachusetts

by grade level reported a subgroup of ELLs who had not achieved English language proficiency despite having spent almost all of their school years in the state. Almost one-fifth (19.7%) of the Grade 9 ELLs in this study had been in Massachusetts schools for eight or more years. Long-term ELLs were not a focus of this study, and findings with respect to this ELL subgroup should be interpreted with caution. In my practice, I have taught ELLs who may have started school in the United States but then returned to their home country for a period of months or years. The data used in this study did not reflect whether the years a student spent in Massachusetts schools were continuous, nor did they indicate socioeconomic or other factors that might impact attainment of English proficiency. Any apparent finding with respect to long-term ELLs is incidental and superficial, and no conclusions should be drawn.

**English proficiency.** English proficiency was assessed by the MEPA instrument in March 2012, a few months before the June 2012 Biology MCAS was administered. English proficiency (MEPA Levels 1 to 5) exhibited a normal distribution, with the greatest number (43.5%) at MEPA Level 3, at which a student “communicates using basic English at school, although errors sometimes interfere with communication and understanding” (MA DESE, 2012f, p. 4). The English proficiency of ELLs who took the June 2012 Biology MCAS was generally lower than secondary ELLs statewide. In 2012, 45% of the ELLs in Grades 9 to 12 in Massachusetts were at MEPA Level 4 (21%) or Level 5 (24%) (MA DESE, 2012g). In this sample, 18.2% were at MEPA Level 4 and 16.7% were at MEPA Level 5.

For ELLs in this sample, distribution across the five English proficiency levels was similar for the Latinate L1 and non-Latinate subgroups, except for a slightly higher

percentage of the non-Latinate L1 group at MEPA Level 5 (20.6% compared to 15.3% of the Latinate L1 group). The mean MEPA scaled score between the groups was statistically significant at the  $p < .05$  level; however, since the mean of both groups was at the higher end of the MEPA Level 3 proficiency range, there was no practical significance. Likewise, the distribution across English proficiency levels was similar for the alphabetic L1 and non-alphabetic L1 groups, and there was no statistically significant difference at the  $p < .05$  level for the mean MEPA scaled score.

The late-entry and not-late-entry groups, however, did have statistically significant differences at the  $p < .05$  level in English proficiency. The mean MEPA scaled score for late-entry ELLs (474.68) corresponded to MEPA Level 3, and the mean MEPA scaled score for not-late-entry ELLs (491.02) corresponded to MEPA Level 4. The majority of late-entry ELLs (45.2%) were at MEPA Level 3, and there were similar percentages above and below. In contrast, 55.4% of the not-late-entry ELLs were above MEPA Level 3, and only 4.6% were below. A student moves from MEPA Level 3 to Level 4 when he or she progresses from using *some* to *most* grade-level academic vocabulary (MA DESE, 2012f; emphasis added.). For the ELLs in this sample, the not-late-entry ELLs generally had higher levels of English proficiency. This is not surprising since this group would have been in the United States longer and had more exposure to English (both general and academic), and they likely would have progressed beyond the lowest levels of English proficiency.

**First languages.** The ELLs in this study were linguistically diverse, with over 70 first languages, but the 10 most common first languages accounted for 88.5% ( $n = 2,935$ )

of the participants.<sup>117,118</sup> The majority of the ELLs in the sample (73%, n = 2,420) had a Latinate first language, and Spanish was the first language for 51.2% of the ELL test-takers (n = 1697); in 2012, Spanish-speaking ELLs represented 54% of ELLs statewide (OELA&AA, 2013). Statewide, the 10 most frequent L1s in 2012 did not include Nepali (1.6% in the sample) and French (1.3% in the sample), and Chinese languages represented 3.2% of ELLs (6.1% in the sample).<sup>119</sup> Alphabetic L1s represented 82.2% of the ELLs in the sample (n = 2,724) even though four of the 10 most common first languages (i.e., Arabic, Chinese Languages, Khmer/Khmai, and Nepali) were non-alphabetic. I encountered some difficulty in the classification of orthographies as alphabetic or non-alphabetic. The *Compendium of the World's Languages* (Campbell, 1991a, 1991b) was the primary resource, and it contained information for most of the 70 first languages represented by the test-takers. Some languages, however, were not included in this reference. In these cases, I consulted Internet resources for orthographic samples and made a visual determination as to whether it relied essentially on the Roman alphabet.<sup>120</sup> It should be noted that although this study classified Chinese languages as non-alphabetic, pīnyīn is an official Romanized script (Campbell, 1991a), and as such, ELLs with one of the Chinese languages as a first language likely would have had exposure to the alphabetic principle.<sup>121</sup>

---

<sup>117</sup> Chinese languages were combined into one category (see Table 4.7).

<sup>118</sup> First language classifications are in Appendix F.

<sup>119</sup> Other difference between the ELLs in the sample and statewide ELLs was that Portuguese was the second most common L1 (6.4%) in the state but the fifth most common (4.8%) in the sample. The ninth and tenth most common L1s in the state were the category Other and Russian, which were replaced by Nepali and French in the sample.

<sup>120</sup> Several languages used the Roman alphabet with some additional letters and/or diacritics to represent sounds specific to that language.

<sup>121</sup> See Bialystok, McBride-Chang, and Luk (2005) for the effect of learning to read English for Chinese children who were familiar with pīnyīn compared to those who were not.

**Late-entry ELL status.** The majority of the ELLs (66.9%) were 12-years-old or older when they entered the United States, and for this sample, approximately 10% more non-Latinate L1 ELLs (74.5%) were late-entry compared to Latinate L1 ELLs (64.1%). The highest percentage of first-year ELLs was in Grade 11 (31.5%,  $n = 179$ ). Fewer than half of the ELLs in the sample (41%,  $n = 1,358$ ) entered Massachusetts schools at the secondary level. The percentage of ELLs who entered as high school students was the lowest in Grade 9 (20.6%,  $n = 236$ ), followed by Grade 10 (44.1%,  $n = 657$ ). Over two-thirds of Grade 12 (67.3%,  $n = 74$ ) and Grade 11 (68.8%,  $n = 391$ ) ELLs were in their first through fourth or first through third year, respectively. These students have at most four years to acquire enough English to demonstrate proficiency in English language arts, mathematics, and biology (or another science content area) for high school graduation and for post-secondary studies. Given the research on second language acquisition, especially with respect to academic language, these ELLs face a seemingly insurmountable task.

### **The June 2012 MCAS Instrument**

The June 2012 Biology MCAS followed its technical blueprint for content domains, and multiple-choice items across six domains, three cognitive skill levels, and a range of linguistic complexity ensured that all students had entry points to the assessment.<sup>122, 123</sup> The instrument emphasized three domains slightly more (i.e., ecology, evolution, and genetics), which represented 60% of the multiple-choice items; the

---

<sup>122</sup> Entry points, as used here, are at the item level, and accessibility is a function of sufficient entry points to demonstrate content knowledge. A student could get some items right (entry points), but the overall instrument could be inaccessible if the entry points were few.

<sup>123</sup> Appendix A contains the June 2012 Biology MCAS; Appendix C contains the Massachusetts High School Biology Standards; and Appendix D contains the Cognitive and Quantitative Skills Descriptions for Science and Technology/Engineering MCAS Tests used for the June 2012 Biology MCAS.



remaining 40% comprised anatomy and physiology, biochemistry, and cell biology items.<sup>124</sup> The distribution of multiple-choice items across the six standards allowed test-takers multiple entry points. If a test-taker was particularly weak in one of the six content domains, he or she could still pass the Biology MCAS since no content domain represented more than 20% of the multiple-choice items.

The 40 multiple-choice items represented three cognitive skill levels: foundational (n = 6), conceptual (n = 18), and application (n = 16). The distribution of cognitive skill levels across domains for this instrument did not appear equal. Although only 15% of the instrument's multiple-choice items were foundational, 40% of anatomy and physiology items were foundational, and there were no foundational items for biochemistry or evolution. Three domains (cell biology, ecology, and evolution) contained predominately conceptual items, and biochemistry and genetics included predominantly application items (60% and 62.5%, respectively). A cognitive load perspective suggests that students would perform higher on anatomy and physiology items and lower on biochemistry and genetics items.

The Biology MCAS is written for and normed on English proficient test-takers. The Phase I textual analyses highlighted domains that might have been more challenging for ELLs because of increased cognitive load from linguistic complexity. There was a wide range of values for the lexical, syntactic, and discourse elements in the multiple-choice items. Item stems ranged from nine to 103 words and from one to 15 sentences; the average number of words in a stem's sentences ranged from six to 24. The majority of

---

<sup>124</sup> The five constructed-response items covered the following domains: (1) anatomy and physiology (item 32), (2) cell biology (item 23), ecology (item 45), evolution (item 12), and genetics (item 44).

item stems had 21 to 40 words (52.5%) and one or two sentences in the stem (57.5%). Item answer options ranged widely in average number of words, from one to 22; however, nearly half of the items (47.5%) had an average of one to three words in an answer option. Likewise, the estimated Lexile<sup>®</sup> score for items ranged from 540 to 1,520. Although some items had high linguistic demands, the wide ranges assured entry points for all students.

Martiniello (2008) called for further studies to examine the interaction between different learning strands and linguistic complexity as a source of differential performance for ELLs. For the June 2012 Biology MCAS instrument, the data suggested that the lexical density of the answer options (TALD) differed among the domains. Scheffé post hoc analyses determined that these differences were statistically significant at the  $p < .05$  level between (1) evolution and biochemistry items, and (2) evolution and cell biology items. This suggested that evolution items might be problematic for ELLs, even if they understood the content and the stem, because differentiating between the answer options required parsing more words for comprehension.

For cognitive skill levels, the statistically significant linguistic differences at the  $p < .05$  level were in the item stem (SLD and SS). Scheffé post hoc analyses determined that the number of words in the stem (SLD) had statistically significant differences at the  $p < .05$  level between (1) foundational and application items, and (2) conceptual and application items. Scheffé post hoc analyses also determined that the number of sentences in the stem (SS) had statistically significant differences at the  $p < .05$  level between conceptual and application items. This suggested that the increased cognitive load from linguistic complexity might have made application items more difficult for ELLs.

When the linguistic element differences among domains and cognitive skill levels are taken together, evolution items at the application level could be the most difficult for ELLs since both the stem and answer options contained a statistically significant higher number of words. This would be followed by application items, which had a statistically significant higher number of words and sentences in the stem. Thus, one might expect ELL performance to be lower for evolution items at the application level, as well as for biochemistry and genetics items, where 60% and 62.5%, respectively, were at the application level. These potential areas of difficulty for ELLs represented 11, or 27.5%, of the multiple-choice items.<sup>125</sup>

### **ELL Performance**

This study's findings were consistent with the ELL achievement gap reported in the literature (see Abedi & Dietel, 2004; Cook et al., 2011; Duran, 2008; Xu & Drame, 2008). The 2012 STE MCAS results for ELLs in the 2014 cohort, for all STE exams and all test administrations, reported a 94% passing rate (MA DESE, 2012i). There are no previous studies that examined performance for all ELLs on a single Biology MCAS administration. By using state-level data, this study addressed validity and reliability concerns raised by Abedi et al. (2004) and Solorzano (2008) with respect to the varying ways states define English proficiency and re-designate ELLs as English proficient.

Approximately half of the ELLs (52.8%) passed the June 2012 Biology MCAS, but only 14.9% scored Proficient or higher. In comparison, 69% of the 2014 cohort scored Proficient or higher on the 2012 STE MCAS exams. The achievement gap

---

<sup>125</sup> Three items were evolution items at the application level, three items were biochemistry items at the application level, and five items were genetics items at the application level.

narrowed when ELLs reached higher levels of English proficiency, and the overwhelming majority of Level 4 and Level 5 ELLs (71.8% and 87.7%, respectively) passed. When compared to the 2014 cohort STE passing rate of 94%, Level 5 ELLs almost closed the gap for passing, but with only 43% scoring Proficient or higher, the achievement gap persisted.

**Content domain performance.** Standardized science assessments measure multiple dimensions, which supports the benefit of reporting sub-scores (Leighton et al., 2007). Compared to the statewide mean percent correct for each content domain, ELLs had lower performance, confirming that the ELL achievement gap on overall MCAS performance persisted in each content domain. The mean percent correct gap ranged from 21.7% for cell biology to 32.2% for evolution. Phase I textual analyses revealed that the lexical density of answer options for evolution items was approximately four times greater than for cell biology items, and one-way ANOVA analyses and Scheffé post hoc analyses determined that this difference was statistically significant at the  $p < .05$  level. The greatest gap on evolution and the smallest gap on cell biology are consistent with Abedi and Hejri's (2004) finding that the ELL achievement gap on the Grade 8 Science NAEP widened when item linguistic complexity increased. This also supported Martiniello's (2008) speculation that linguistic complexity was a source of ELL differential functioning among Grade 4 Mathematics MCAS strands. Caution is warranted here because there may have been other factors that impacted performance differences between these two domains. For example, evolution contained more items (37.5%) at the application cognitive skill level than cell biology (16.7%), and there were no evolution items at the foundational cognitive skill level. It is also possible that the gap

was smallest for cell biology because statewide performance was lowest for this domain.<sup>126</sup>

Although the data on evolution and cell biology performance appeared to support item linguistic complexity as a source of ELL differential functioning, the performance data for other content domains were inconsistent. The ranking of ELL content domain performance mirrored the statewide performance across domains, except for a reversal of biochemistry and evolution (see Figure 4.13 and Table 4.33). Both ELLs and all test-takers performed best on ecology and worst on cell biology; the majority of items in both these areas occurred at the conceptual cognitive skill level, but ecology items had higher mean linguistic complexity values than cell biology items (see Table 4.16). If linguistic complexity were the only factor driving content domain performance, ecology performance should have been lower than cell biology. This suggested that differential performance across domains is due to intrinsic content elements. One hypothesis for the high performance on ecology and anatomy and physiology for both ELLs and all test-takers is that these domain concepts build on prior knowledge from elementary and middle school science. Another hypothesis is that these domains are more concrete (e.g., food chains, the water cycle, digestion, etc.) than cell biology or genetics, in which concepts are often abstract (e.g., transcription, enzyme-substrate complex, mutations, etc.).

**Cognitive skill level performance.** The multiple-choice items on the June 2012 Biology MCAS assessed content knowledge across three cognitive skill levels:

---

<sup>126</sup> Except for cell biology (62.3%), the average mean percent for all test-takers was above 70% for each content domain.

foundational, conceptual, and application. Cognitive load theory posits that as tasks become more complex, fewer working memory resources are available for content knowledge and language (see Paas et al., 2010; Shank, 2007; Van Gog et al., 2010). As discussed in Chapter 2, cognitive skill level and item difficulty are not synonymous; however, required skills differ and become more complex. Phase I textual analyses determined that language became more complex across the three cognitive skill levels in terms of number of words (SLD) and sentences (SS) in the stem. This suggested that ELL performance should have been highest on foundational items, then decreased from foundational to conceptual items, and from conceptual to application items.

The data in this study were inconsistent with respect to cognitive load theory. Though ELLs performed highest on foundational items, they performed similarly on conceptual and application items. These findings only partially supported Duran and Moreno's (2004, as cited in Duran, 2008) finding that cognitive load is a factor in ELL performance. Yet, the similar performance on conceptual and application items was consistent with Robinson's (2005) cognition hypothesis prediction that increased task complexity leads to increased accuracy of second language output, and it also supported the previous findings of Michel, Kuiken, and Vedder (2007), and Robinson and Gilabert (2007).<sup>127</sup> Specifically, increased task complexity (i.e., higher cognitive demands of application items) may have led to increased noticing and attending. This may have led to higher application item performance that equaled conceptual item performance; increased noticing and attending may have mitigated increased task complexity.

---

<sup>127</sup> The cognition hypothesis predictions were for oral language output.

Caution, however, should be used in interpreting this study's finding on cognitive skill level performance. Ćurković (2012) found that performance data on the Croatian state-level summative high school math exam did not support the existence of three hierarchal cognitive levels (knowledge, comprehension, and application). It is possible that, like Ćurković's study, this study's cognitive skill levels were not hierarchal. Notwithstanding the question of a hierarchal model, Housen and Kuiken (2009) pointed out that studies have not consistently defined task complexity—another reason to exercise caution in interpreting the current study's findings on cognitive skill level performance. In addition, item cognitive skill levels are not reported publically, so no comparison could be made between ELL and statewide performance. Thus, it is unknown whether the similar performance between conceptual and application skill levels was unique to ELL test-takers.

**Linguistic complexity performance.** Abedi and Lord (2001) called for continued research into the role of language in content assessment. This study analyzed performance for the item attribute of linguistic complexity, which was operationalized in the Phase I textual analyses and classified as low, medium, and high composite linguistic complexity (low CLC, medium CLC, and high CLC, respectively). As discussed in Chapter 2, higher language demands increase cognitive load for ELLs, and performance should decrease in going from low CLC to medium CLC to high CLC. As expected, ELL performance was highest on low CLC items, but there was no decrease in performance between medium CLC and high CLC items. This was similar to Abedi and Hejri's (2004) finding that increased linguistic complexity was inconsistent in widening the ELL achievement gap on the Grade 4 and Grade 8 NAEP, and Schneider et al.'s (2013)

finding that reading load was not a consistent predictor of item difficulty for Grade 4 and Grade 8 mathematics multiple-choice items.<sup>128,129</sup> The current study's findings are also consistent with DeLeeuw and Mayer's (2008) finding that increased item redundancy (similar to this study's lexical density) and sentence complexity were not statistically significant sources of cognitive load for college students.<sup>130</sup>

These findings were not consistent with Martiniello's (2008) finding that linguistic complexity was a source of differential item functioning for ELLs and that measures of lexical and syntactic complexity could estimate item difficulty for ELLs on the 2003 Grade 4 Mathematics MCAS. There are some important differences beyond the content area between Martiniello's study and the current study that may explain the inconsistency. Martiniello only explored linguistic complexity for Spanish-speaking students ( $n = 24$ ) through a think-aloud protocol. The current study explored performance for ELLs with over 70 first languages; Spanish-speaking ELLs represented 51.2% of the sample. The ELLs in the current study were older, with 66.9% of the sample entering the United States at the age of 12 years or later. As discussed in Chapter 2, late-entry ELLs have fully acquired their L1, which is in contrast to Martiniello's Grade 4 ELLs. The higher metalinguistic awareness that comes with full L1 acquisition may have led to the Goldilocks effect discussed in Chapter 4. The ELLs in the current study may have attended more closely to the language of high CLC items, which mitigated the impact of the increased linguistic complexity. This would support the predictions made by the

---

<sup>128</sup> Abedi and Hejri (2004) found that increased linguistic complexity did not widen the ELL achievement gap for Grade 4 science, Grade 8 math, and writing in both Grade 4 and Grade 8, but it did widen the gap for Grade 4 math and Grade 8 science.

<sup>129</sup> The estimated Lexile® score in the current study is a measure of reading load.

<sup>130</sup> Their study was not on ELLs.



cognition hypothesis (Robinson, 2005). Another possible explanation is that because the ELLs in the current study were older, there was positive transference of L1 literacy and academic language skills.

This study found that ELL performance on items of increasing linguistic complexity was both consistent and inconsistent with previous studies, though the findings lent limited, tentative support to the cognition hypothesis. As discussed, content domain performance findings were inconsistent with respect to linguistic complexity. The greatest performance gap between ELLs and all test-takers was on evolution items, which had high answer lexical density. This was inconsistent with the highest ELL performance on ecology items and the lowest ELL performance on cell biology items where the former had higher linguistic complexity. It is possible that these inconsistencies arose from this study's operationalization of linguistic complexity. Although this study captured elements of linguistic complexity, it may not have captured the elements that significantly impacted ELL performance on the Biology MCAS. It appears that this study joins the body of literature with "inconsistency of complexity findings both across and within studies" (Kuiken & Vedder, 2012, p. 277).

### **English Proficiency Impact**

Previous studies have established that English proficiency impacts ELL performance, as does the linguistic complexity of the instrument, resulting in an achievement gap between ELLs and non-ELLs on standardized assessments (Abedi, 2002, 2009; Abedi & Gándara, 2006; Abedi & Hejri, 2004; Abedi et al., 2004; Abedi & Lord, 2001; Martiniello, 2008; Menken, 2008, Chapter 4; Solano-Flores & Li, 2009; Solorzano, 2008). Research identified a need for further study into the role of language in

ELL content assessment (Abedi, 2008b; Abedi et al., 2004; Abedi & Lord, 2001; Martiniello, 2008; Solorzano, 2008) and the extent of impact of English proficiency (Solorzano, 2008) in order to inform the construction of reliable and valid content assessments for ELLs (Abedi, 2008b) and to serve as a source of differential item functioning (Martiniello, 2008).

**When do ELLs have meaningful participation on the Biology MCAS?** Abedi (2008b) argued that ELLs should participate in standardized content assessments only when English proficiency assessments show that language proficiency matches the language of the content assessment. This study disaggregated the ELL sample by English proficiency levels to explore when language proficiency was sufficient for meaningful ELL participation on the Biology MCAS. For Massachusetts ELLs, it was at MEPA Level 3 (at which academic language use begins) that almost half of the ELLs (46.6%) in the sample passed the Biology MCAS. As ELLs became more English proficient, especially in academic English, their performance on the June 2012 Biology MCAS increased. Progressing from Level 3 (some grade-level academic language) to Level 4 (most grade-level academic language) increased the passing rate to nearly three-fourths (71.8%), and it appeared to be the point where approximately one-fifth (21.1%) demonstrated content proficiency. At MEPA Level 5, 87.7% passed and almost one-half (43.3%) scored Proficient or higher. For Level 5 ELLs, the achievement gap almost closed for passing the Biology MCAS, but it still remained for scoring Proficient or higher.

Although MEPA Level 3 was the turning point where ELLs began accessing the Biology MCAS, a passing rate of less than 50% cannot be considered meaningful

participation for the majority. Factors beyond the scope of this study might explain how nearly half of the sample passed with only the beginnings of academic language. Possible factors include L1 content knowledge and literacy skills (including L1 academic language) that helped students access the instrument and participate meaningfully. The data indicated that stakeholders could expect to see the beginnings of meaningful ELL participation at MEPA Level 3, but to expect that all ELLs could access the instrument was premature. The results indicated that at MEPA Levels 4 and 5, the overwhelming majority of ELLs had meaningful participation and could be reasonably expected to pass the MCAS. At MEPA Levels 4 and 5, ELLs could also access most grade-level texts and academic language. Thus, the data supported that the Biology MCAS assessed content knowledge at grade-level discourse.

The data further indicated that stakeholders should not expect passing rates and performance levels equivalent to native-speaker performance. Even though Level 4 and Level 5 ELLs are accessing grade-level academic language, they still do not have native-speaker proficiency.<sup>131</sup> Additionally, some Level 4 and Level 5 ELLs may have educational gaps that no longer affect English proficiency but could affect performance in a content domain or on a specific item. Even without a formal education gap, ELLs at these higher English proficiency levels may have content area deficiencies because they were not fully accessing the curriculum when they were at lower English proficiency levels. For example, some ELLs are in high schools where the Massachusetts biology standards are taught in a two-year course sequence. An ELL who was MEPA Level 3 in

---

<sup>131</sup> Some Level 5 ELLs who are ready for reclassification as English proficient may have acquired native-speaker proficiency, but the ELL designation and proficiency assessment determine whether they can perform ordinary classwork in English, not if they have acquired native-speaker proficiency. It is a subtle distinction, but a distinction nonetheless.

Grade 9 may have accessed less of the first course than he or she did in the second-year course when he or she may have been a Level 4 or 5. Similar reasoning holds for ELLs who experience the Biology standards in a one-year course: As their English proficiency increases throughout the year, they access more and more of the curriculum. These ELLs may be stronger in both content and academic discourse for standards studied in April than those studied in October.

This study showed that a single homogenized ELL statistic masks the gains that are made as ELLs acquire English proficiency. It showed that stakeholders should expect ELLs to demonstrate content knowledge on the Biology MCAS instrument beginning at MEPA Level 3 but only as a turning point. The data indicated that only at the next higher level of English proficiency, MEPA Level 4, did the overwhelming majority of ELLs have meaningful participation and pass, though Proficient and higher performance levels were less than those for English proficient test-takers. The fact that a majority of ELLs passed at the Needs Improvement performance level is not a reason for finger pointing and blaming because the data showed that the percentage of ELLs at the Proficient and higher performance level doubled between Level 4 and Level 5 (21.1% and 43.3%, respectively). In other words, Needs Improvement appeared to result from the confounding of content and language. The data suggested that if the ELLs at the Needs Improvement level retook the MCAS when they gained the next higher English proficiency level—FLEP status, or FLEP exit status—the percentage at Proficient or higher would approach and perhaps equal the performance levels of native speakers. This, of course, is not feasible and only conjecture.

**English proficiency was a strong predictor of Biology MCAS performance.**

Since a test-taker needs to understand the language of the assessment, it was not unexpected that English proficiency impacted Biology MCAS performance. This study went beyond previous studies and quantified the ELL performance variance attributed to differences in English proficiency levels.<sup>132</sup> A strong, positive, linear relationship between English proficiency and Biology MCAS performance emerged around MEPA Level 3, the level where ELLs begin to use academic language.<sup>133</sup> A simple linear regression analysis resulted in a statistically significant model ( $p < .05$ ) in which English proficiency contributed approximately 29% of the variation in Biology MCAS performance among ELLs. English proficiency was a strong predictor of Biology MCAS score, with  $B = .25$ , indicating that for each additional 4-point increase in MEPA score, the Biology MCAS score would increase by one point. To eliminate low English proficiency as a source of validity and reliability issues, the simple regression analysis was repeated for the subgroup of ELLs who had a MEPA score of 464 and above (MEPA Levels 3 to 5). The model emerged as statistically significant ( $p < .05$ ); English proficiency contributed approximately 25% of the variation in Biology MCAS performance among MEPA Level 3 to 5 ELLs. English proficiency remained a strong predictor of Biology MCAS score, with  $B = .33$ , which indicated that for each additional 3-point increase in MEPA score, the Biology MCAS score would increase by one point. Although English proficiency contributed approximately 4% less of the variance among MEPA Levels 3 to 5 than for the entire sample, the B value indicated that each point

---

<sup>132</sup> As discussed in Chapter 4, this was possible because all the ELLs in the sample had been assessed by the same English proficiency instrument (the MEPA) less than three months prior to the June 2012 Biology MCAS.

<sup>133</sup> MEPA Level 3 began at a MEPA scaled score of 464.

increase in MEPA score had a slightly stronger impact for ELLs who were developing academic language proficiency. This suggests that continued language development for ELLs at the higher English proficiency is still critical for closing the achievement gap.

***Impact on content domain performance.*** The same pattern of English proficiency impact was seen on performance within the six content domains, with domain performance increasing as English proficiency increased. One-way ANOVA analyses determined that English proficiency had a statistically significant ( $p < .05$ ) impact on performance across all six content domains, and univariate analyses of variance determined medium or large effect sizes. The greatest impact was on ecology and evolution items, where English proficiency contributed 17% of the variance, followed by genetics, where it contributed 14% of the variance. Scheffé post hoc analyses, however, determined that MEPA Level 1 and Level 2 ELLs had indistinguishable performance for each content domain. This finding validated Abedi and Hejri's (2004) reliability concern at lower English proficiency levels.

Moving from MEPA Level 2 to Level 3 had an increase in performance for all content domains, and as English proficiency increased through successive levels, performance on all content domains increased. The difference in mean percent correct between MEPA Level 3 and Level 5 ranged from 13.8% (cell biology) to 21.58% (genetics), with five of the six domains having mean percent correct differences greater than 17%. This was consistent with this study's finding that MEPA Level 3 was the turning point where ELLs began to access Biology MCAS items.

***Impact on cognitive skill level performance.*** One-way ANOVA analyses determined that English proficiency had a statistically significant impact on performance

at the different cognitive skill levels, and univariate analyses determined a medium effect size at the foundational cognitive skill level, where it contributed 11.7% of the variance. Beyond the foundational level, the impact of English proficiency was large, contributing 18.8% of the variance on conceptual items and 18.1% of the variance on application items. As English proficiency increased, the mean percent correct at each cognitive skill level increased; however, Scheffé post hoc analyses determined that there were no statistically significant performance differences between MEPA Levels 1 and 2, which was consistent with findings for overall performance, content domain performance, and linguistic complexity performance. Scheffé post hoc analyses further determined that English proficiency impact on cognitive skill performance was statistically significant at the  $p < .05$  level once ELLs reached MEPA Level 3. This was consistent with the finding that MEPA Level 3 appeared to be the turning point at which the Biology MCAS became accessible.

As discussed previously, there have been no studies or publically released performance data for the Biology MCAS with respect to the cognitive skill levels of the items. This study had limited support for cognitive load theory. As expected from cognitive load theory, once ELLs reached Level 3 proficiency, each successive level of English proficiency experienced increased performance at a cognitive skill level; the mean percent correct difference was between 9.1% and 10.3%, except when moving from Level 3 to 4 on foundational items, in which case the difference was 7%.

It appeared that as English proficiency increased, the decreased cognitive load from language freed up cognitive resources for content domain skills and knowledge. Performance within an English proficiency level, however, did not support predictions

based on cognitive load theory. At a given English proficiency level, cognitive load theory would predict that performance would decrease from foundational to conceptual items, and from conceptual to application items. Performance decreased slightly when moving from foundational to conceptual items; however, performance was similar for conceptual and application items. This suggested tentative, limited support for the cognition hypothesis; increased noticing and attending resulted in application item performance equaling conceptual item performance.

*Impact on linguistic complexity performance.* Although other factors, such as poverty and parent education, impact ELL achievement, language has the greatest impact, with greater gaps forming as language demands increase (Abedi, 2002, 2008b, 2009; Abedi & Gándara, 2006; Abedi et al., 2004). This study analyzed ELL performance on items at three levels of composite linguistic complexity (i.e., low CLC, medium CLC, and high CLC). One-way ANOVA analyses confirmed that English proficiency had a statistically significant ( $p < .05$ ) impact on performance at each of the three levels of linguistic complexity. Scheffé post hoc analyses determined that there was no statistically significant difference in performance for MEPA Levels 1 and 2, which was consistent with the findings of indistinguishable performance between these levels on overall MCAS performance, content domain performance, and cognitive skill level performance.

Scheffé post hoc analyses confirmed that the difference in mean percent correct was statistically significant at the  $p < .05$  level between all other adjacent MEPA levels. The data suggested that items with low linguistic complexity approached accessibility for ELLs at MEPA Level 3; however, at Level 4 the mean percent correct was greater than 50%. For items with medium and high linguistic complexity, the data suggested that



Level 4 was the cusp of accessibility, and these items were generally accessible at Level 5. These findings were consistent with Abedi and Lord's (2001) finding that linguistic complexity impacted ELL performance on a standardized mathematics assessment (see also Abedi, 2002, 2009; Abedi & Gándara, 2006; Abedi & Hejri, 2004; Abedi et al., 2004; Martiniello, 2008; Menken, 2008, Chapter 4; Solano-Flores & Li, 2009; Solorzano, 2008).

This study went beyond previous studies to determine to what extent English proficiency (MEPA scaled score) significantly predicted performance on items with low, medium, and high linguistic complexity. The models for performance at each level of linguistic complexity emerged as statistically significant at the  $p < .05$  level. English proficiency was a strong predictor of performance at all three levels of linguistic complexity;  $R^2$  values indicated that English proficiency contributed approximately 19% of the variation among ELLs on low CLC items, 15% on medium CLC items, and 17% on high CLC items. The B values indicated that for each level of linguistic complexity, an additional 5-point increase in MEPA score would increase the mean percent correct by approximately 2%. These findings were unexpected. Second language acquisition and cognitive load theories suggest that English proficiency should have higher impact as linguistic complexity increases (i.e., items with high linguistic complexity require higher levels of English proficiency for access).

To eliminate validity and reliability issues at the lower levels of English proficiency, the simple linear regression analyses were repeated for ELLs at MEPA Levels 3 to 5, and the three models emerged as statistically significant at the  $p < .05$  level, and compared to the whole sample, English proficiency was a slightly stronger predictor

for ELL performance. The general model for linguistic complexity performance was that a 2-point increase in MEPA scaled score would increase the mean percent correct at each linguistic complexity level by approximately 1%. This finding is inconsistent with Abedi's (2002) finding that the impact of linguistic complexity increased as the content area's language demands increased. Krahen's input hypothesis and Cummins' four-quadrant model both suggest that increased linguistic complexity becomes more accessible as English proficiency increases. English proficiency should have contributed the most variance for high linguistic complexity performance and the least for items of low linguistic complexity where items should have been accessible to most of the ELLs at MEPA Levels 3 to 5. The absence of a pattern of increasing English proficiency impact as linguistic complexity increases warrants further investigation.

**English proficiency was a construct-irrelevant variance.** Although previous studies have shown that English proficiency impacts ELL achievement on standardized tests, the varying definitions of English proficiency and how ELLs are re-designated as English proficient created concerns for norming, validity, and reliability (Abedi et al., 2004; Solorzano, 2008). The current study addressed these concerns by using state-level data that included a uniform assessment of English language proficiency—that is, the MEPA. This study analyzed the relationship between ELL Biology MCAS performance and English proficiency to inform whether validity and reliability concerns exist at all levels of English proficiency. Scheffé post hoc analyses of ELL performance by English proficiency level consistently showed indistinguishable performance between MEPA Level 1 and Level 2 (the lower end of English proficiency) on overall MCAS score, content domain performance, cognitive skill level performance, and linguistic complexity

performance. This suggested that the Biology MCAS was equally inaccessible to Levels 1 and 2, thereby supporting validity and reliability concerns raised in earlier studies (Abedi, 2002, 2008b; Abedi & Gándara, 2006),

Performance across the six content domains showed a similar pattern for each English proficiency level, and this pattern mirrored the statewide results (see Figures 4.13 and 4.14). Since statewide results and performance for all MEPA levels rose and fell together across the content domains, this suggested that ELLs and non-ELLs had the same domain content challenges (e.g., cell biology was the most difficult domain, and ecology was the easiest domain). This indicated that differential performance among domains was due to intrinsic content demands. Within a domain, however, differential performance was due to English proficiency where each successively higher English proficiency level had higher performance (vertical distance), except for between Level 1 and Level 2. The mean percent correct for a content domain was highest for the statewide reported results, followed by MEPA Level 5, Level 4, Level 3, and indistinguishable lowest performance for MEPA Levels 1 and 2. This supported Abedi's (2002) hypothesis that language factors introduced construct-irrelevant variance.

**ELL Biology MCAS performance supported some aspects of cognitive load theory, the cognition hypothesis, and a Goldilocks effect.** Cognitive load is the amount of mental processing needed for a task and its demands on working memory. Several factors, including the content and the number of items that require attending, affect the cognitive load of a task (Shank, 2007), and "the more difficult one factor is (e.g., the language), the less attention can be dedicated to another (e.g., the content)" (Dickey, 2004, p. 11). The schemas stored in long-term memory determine performance in a given

area or domain, and they reduce demands on working memory because they are processed as one element (Paas & Sweller, 2012; Paas et al., 2010).

***Cognitive load theory.*** Cognitive load theory suggests that ELLs who possess lower levels of English proficiency will experience higher cognitive load because they must attend to and integrate the interacting elements of vocabulary, sentence structure, and syntactic rules for comprehension, which may or may not be successful (C. H. Lee & Kalyuga, 2011). As English proficiency increases, ELLs gain automaticity and can retrieve schemas from long-term memory, which reduces cognitive load (C. H. Lee & Kalyuga, 2011; Sweller & Chandler, 1991). Cognitive load theory maintains that performance increases as English proficiency increases, and as task complexity increases, performance decreases. Accordingly, English proficiency was a strong predictor of overall Biology MCAS performance. The indistinguishable performance difference between MEPA Level 1 and Level 2 ELLs on overall MCAS, content domain, cognitive skill level, and linguistic complexity indicated that the cognitive load from language was overwhelming for these ELLs—results which supported cognitive load theory (see Paas et al., 2010; Shank, 2007).<sup>134</sup> The impact of cognitive load from language was evident in domain performance where ELL and statewide performance rose and fell in similar patterns, but within a domain, performance increased as English proficiency increased. Under cognitive load theory, ELLs with increased language (automaticity and schema) had more working memory resources available for domain knowledge and skills. For

---

<sup>134</sup> As noted previously, an assessment can be overall inaccessible despite a student having an entry point to a few items. If there were no entry points, the score would be zero.

each cognitive skill level and linguistic complexity level, performance also increased with increasing English proficiency, except for between Levels 1 and 2.

Within a given English proficiency level, the cognitive load from language should be similar. Thus, performance should decrease as task complexity increases (in this study, task complexity was measured by item cognitive skill level or linguistic complexity). Decreased performance was seen for all English proficiency levels when moving from the least complex task to the next level of complexity (foundational to conceptual, and low CLC to medium CLC); however, there was no corresponding decrease in performance when moving from intermediate- to high-task complexity. This was consistent with DeLeeuw and Mayer's (2008) finding that there was no statistically significant decrease in performance for items with redundant words (lexical density). It was also consistent with the finding of Schneider, Huff, Egan, Gaines, and Ferrara (2013) that reading load was not a consistent predictor of item difficulty of mathematics multiple-choice items for Grades 4 and 8; in the current study, reading load was measured by the estimated Lexile<sup>®</sup> score.

First language cognates should have lessened reading load and conferred an advantage to ELLs with a Latinate L1 (see Martiniello, 2008); however, this study found that performance was higher for non-Latinate L1 ELLs. DeLeeuw and Mayer (2008) conjectured that prior knowledge might have been a hidden factor in their results, and the same might have been true for the current study. The orthographic distance between a non-alphabetic L1 and alphabetic L2 theoretically should have increased cognitive load; yet, ELLs with a non-alphabetic L1 performed better on the June 2012 Biology MCAS. These findings are not necessarily inconsistent with cognitive load theory. As discussed

previously, MEPA Level 3 was the turning point where almost half of the ELLs passed the Biology MCAS. At this level, alphabetic automaticity may have been sufficient and a non-alphabetic L1 did not increase cognitive load. An alternative explanation is that these finding support Robinson's (2005) cognition hypothesis predictions.

***The cognition hypothesis.*** Based on the cognition hypothesis, Robinson (2005) predicted that increased task complexity leads to increased attending and higher L2 performance. Motivation in the face of increased task complexity leads to the increased noticing and attending. The Biology MCAS is high-stakes since passing it is requisite for a high school diploma. In my experience in three urban high schools, students are highly motivated to pass the Biology MCAS exam. That is not to say that ELLs may not be anxious or even terrified of taking this assessment in English, but they are motivated to do their best; they understand its high-stakes nature. In my experience, the days and weeks leading up to the Grade 10 MCAS exams are analogous to preparing for a big game. Teachers review the content, but just as much, they reassure the students that they are ready and capable. Some schools have motivational assemblies where administrators and teachers cheer on the students the day before the MCAS exams. There is no doubting the importance of these exams. Some schools have programs after school or on weekends to help students further prepare. Schools often have a late start for non-MCAS students with only the MCAS test-takers arriving in the morning. In my practice, my ELL students have been nervous and some have been scared, but they were all motivated—even students who had not expressed high motivation in class. It was not possible to know whether every student, ELL or non-ELL, was motivated; however, it is likely that most were motivated to pass the June 2012 Biology MCAS. With high motivation, the

cognition hypothesis offers a possible explanation for data that was inconsistent with cognitive load theory.

Cognitive load from language should be similar for ELLs with the same level of English proficiency. Within each English proficiency level, performance decreased when proceeding from the least complex task to the next level of cognitive skill complexity (foundational to conceptual); however, there was no corresponding decrease in performance from intermediate (conceptual) to high (application) task complexity. The similar cognitive skill performance on intermediate (conceptual) to high (application) task complexity items supported Robinson's prediction. This pattern was the same for the whole sample and within L1 family and L1 orthographic subgroups. Performance data for item linguistic complexity lent further support. A similar pattern in which medium CLC and high CLC performance was lower than low CLC but similar to each other was seen for the whole sample, within English proficiency subgroups, L1 family subgroups, and L1 orthography subgroups. High CLC items could have highlighted the L1 to L2 distance and increased noticing and attending to language, which may have led to increased item comprehension and performance.

This finding is consistent with Leighton and Gokiert's (2005) belief that students can access the knowledge and skills for a correct answer even in the absence of full comprehension of the item. Thus, even though the ELLs are by definition not English proficient, they were able to attend to the higher linguistic demands and perform just as well on high CLC items as they did on medium CLC items. Higher noticing and attending could also explain the higher performance of non-Latinate L1 ELLs and non-alphabetic L1 ELLs. The lack of L1-to-English cognates and the L1 to L2 orthographic

distance would have increased task complexity, which could have resulted in more noticing and attending to the item's linguistic features and increased performance.

***The Goldilocks effect.*** The Goldilocks effect builds on Robinson's (2005) second prediction that increased task complexity can lead to higher L2 performance. This study's data supported Robinson's cognition hypothesis prediction with respect to similar performance for intermediate and high cognitive skill and linguistic complexity. The data, however, were inconsistent and did not suggest a similar increased attending effect in going from low to medium complex tasks. One possible explanation is what can be thought of as a Goldilocks effect. English language learners recognized low CLC items as linguistically simple, and their noticing was not heightened because the language was less of a barrier in demonstrating content performance. When ELLs encountered high CLC items, they recognized that the items were more difficult to comprehend, so they attended more closely to the language of the item. The medium CLC items, however, may have seemed not linguistically simple but also not linguistically complex. The ELLs may have thought the language of medium CLC was "just right" for their English proficiency level, and without the heightened noticing and attending to language, performance was lower than on low CLC items and the same as on high CLC items with increased noticing. Further study is needed to determine whether this finding was limited to this sample and the June 2012 Biology MCAS.

### **First Language Characteristics Impact**

This study analyzed Biology MCAS performance for ELLs with over 70 first languages. This addressed the limitation of only exploring performance for Spanish-



speaking ELLs in Martiniello's (2008) study of the Grade 4 Mathematics MCAS and Abedi and Hejri's (2004) study of ELL accommodations on the NAEP.

**Lower performance for ELLs with a Latinate L1.** The use of L1-L2 cognates is an L2 reading comprehension strategy, and Martiniello (2008) found that Spanish-speaking ELLs used this strategy on the Grade 4 Mathematics MCAS. The current study explored whether a Latinate L1 impacted performance on the Biology MCAS and found differential performance that favored ELLs with a non-Latinate L1. On the June 2012 Biology MCAS, ELLs with a Latinate L1 had statistically significant ( $p < .05$ ) lower performance than non-Latinate L1 ELLs, not only on overall performance but also on the multiple-choice item attributes of content domain, cognitive skill level, and linguistic complexity. The mean MEPA scaled score for both groups was at the higher end of MEPA Level 3, and statistical analysis determined that any effect size favoring non-Latinate L1 ELLs was minor (Cohen's  $d = .19$ ). Therefore, lower performance by ELLs with Latinate L1 was not attributable to differences in English proficiency.

For overall MCAS score, the performance gap between Latinate L1 and non-Latinate L1 ELLs widened at the Proficient and Advanced MCAS performance levels. On overall MCAS score, L1 family contributed 4% of the variance, with non-Latinate L1 ELLs favored. When the subgroups were disaggregated into MEPA Levels 1 to 2 and MEPA Levels 3 to 5, differences emerged. First language family impact was not statistically significant at the  $p < .05$  level for ELLs at low levels of English proficiency (MEPA Levels 1 to 2), which was consistent with this study's finding of indistinguishable performance for these lower levels. In contrast, L1 family impact was statistically significant at the  $p < .05$  level for ELLs at MEPA Levels 3 and above, and

the variance attributed to the L1 family increased to 4.9% with non-Latinate L1 ELLs favored.

Lower Latinate L1 performance persisted on multiple-choice item attributes. Latinate L1 ELLs had statistically significant ( $p < .05$ ) lower performance for each content domain; however, impacts ranged from negligible to contributing 4.2% of the variance. There was negligible impact on the two domains where all ELLs performed highest (ecology and anatomy and physiology). First language family contributed the most variance on the two most challenging domains for ELLs and all test-takers: 3.4% of the variance on cell biology and 4.2% on genetics. This suggested that as content became more challenging, there was an increasing positive impact of a non-Latinate L1.

The Latinate L1 gap was also evident on item linguistic complexity performance, where 1.8% to 3.1% of the variance was attributable to a Latinate or non-Latinate L1, with the greatest impact on items of high linguistic complexity. First language family had the least impact on cognitive skill performance, where it contributed approximately 2% of the variance between groups. Although individual performance variances may seem small, the cumulative performance variances over content domains, cognitive skills, and linguistic complexity translated into lower performance for Latinate L1 ELLs.

The findings in this study cannot be equated with the Hispanic achievement gap, which is well-established in the literature (Murphy, 2010).<sup>135</sup> This study did not examine performance by race, and the Latinate L1 subgroup in this study included both Hispanic and non-Hispanic ELLs. Although the majority of the current study's Latinate L1

---

<sup>135</sup> Although ELLs would be included within the Hispanic subgroup, the Hispanic subgroup is much larger and contains Hispanic students who have always been English proficient. English proficiency is not equivalent to academic language proficiency, and it is the lack of the latter that may contribute to the Hispanic achievement gap.

subgroup (70.1%,  $n = 1697$ ) were Spanish speakers, this subgroup also included ELLs with other Latinate L1s. Thus, caution should be used when interpreting findings in relation to previous studies on standardized test performance for Spanish-speaking ELLs.

In their study of Spanish-speaking ELLs, Solano-Flores and Li (2009) found inconsistent performance across items and the language of the items (i.e., standard Spanish, Spanish dialect, or standard English); they also found that score dependability varied across and within the languages. Since the majority of the Latinate L1 ELLs in this study were Spanish speakers, this study tentatively suggested corroboration of Solano-Flore and Li's findings. Their finding that standard Spanish produced more dependable scores than the Spanish dialect suggests that prior schooling may have had hidden impact (i.e., prior schooling would have resulted in more exposure to content in standard Spanish), and this may have also been a hidden factor in this study. Now that differential performance based on first language family has been identified within ELLs, more study of the Latinate L1 subgroup is needed to explore sources of differential performance.

In the present study, Latinate-English cognates did not appear to confer an advantage. This was inconsistent with Martiniello's (2008) finding that Spanish-speaking ELLs used knowledge of Spanish-English cognates for unknown vocabulary, a strategy she believed was generalizable to ELLs with other Latinate L1s.<sup>136</sup> Martiniello used a think-aloud protocol with Grade 4 ELLs ( $n = 24$ ) who had two to four years of schooling in Massachusetts, and two of the six school sites offered dual immersion (Spanish-English) programs; she did not include ELLs who could not read or

---

<sup>136</sup> Martiniello (2008) also believed that her finding on cognates was also generalizable to Germanic languages.

communicate in English. Almost one-quarter of the Latinate L1 ELLs in the current study (23.2%,  $n = 562$ ) were at MEPA Levels 1 and 2 (levels prior to the use of academic language), and 22.2% ( $n = 736$ ) were in their first year in Massachusetts schools. Although this study's Latinate L1 performance gap appeared inconsistent with Martiniello's finding, future studies that account for language proficiency, L1 literacy, and prior schooling may inform the nature of Latinate L1 lower performance and support Martiniello's finding for ELLs at higher levels of English proficiency and with L1 literacy.

A Latinate L1 can confer an L1-English cognate advantage, and higher Latinate L1 performance was expected though not seen. A first language does not confer greater or lesser science aptitude. Thus, the consistently lower performance for ELLs with a Latinate L1 within the same English proficiency level must be attributed to a factor beyond the scope of this study and therefore is only speculative. In my urban school practice, almost all of my ELLs have low socioeconomic status and many experience week-to-week financial insecurity.<sup>137</sup> It is well-established in the literature that low income or poverty impacts academic performance (Murphy, 2010), but I have often found that a student's socioeconomic status in his or her home country can impact academic achievement as well. Students who had educational opportunities in their home countries, especially academic L1, can often demonstrate content knowledge, though they are not English proficient, irrespective of first language.

Two of my former students, both Latinate L1 ELLs, illustrate that home country socioeconomic opportunities have impact. One had a good education in her country, but

---

<sup>137</sup> In 2011, 79% of ELLs in Massachusetts were low-income (MA DESE, 2012j).

her family struggled financially to meet basic needs here in the United States. Her father had difficulty finding full-time employment, and her mother earned what she could by cleaning homes part-time, though she had been a nurse in her country. The second student attended a private school in his home country. In the United States, his family lived on his mother's wages from a fast-food restaurant. Both of these students were low-income here, but they had higher socioeconomic status in their home countries. Both were also late-entry ELLs who scored Proficient on the Biology MCAS while in self-contained SEI content classes and were not English proficient; both had a Latinate L1.

Prior educational experiences vary greatly, and home country socioeconomic status often translates into L1 schooling opportunities. Differences in educational opportunities exist in many ELL home countries, not just those with a Latinate language. Some ELLs have spent years in refugee camps with little access to schooling, been denied schooling because of gender, or were unable to pay for schooling. I have found that among students from the same home country, students from rural schools can be below grade-level compared to their urban or private-school peers because of inequities in that country's educational system. More investigation is needed into which Latinate L1 ELLs are at increased academic risk.

**Weak impact of L1 orthography.** This study also found a performance gap when the sample was disaggregated by L1 orthography, but the impact was less than that of L1 family. On overall MCAS performance, ELLs with a non-alphabetic L1 had a statistically significant ( $p < .05$ ) higher performance compared to ELLs with an alphabetic L1. First language orthography contributed 2% of the difference between groups, which was half of the variance attributable to L1 family on overall MCAS

performance. As with L1 family, further disaggregation into subgroups based on English proficiency showed that the impact of L1 orthography was (1) not statistically significant for ELLs at MEPA Levels 1 to 2, and (2) increased to 3.4% for ELLs at MEPA Levels 3 to 5, though this was still less than the 4.9% variance attributed to L1 family for this subgroup.

First language orthography also impacted performance for the item attributes of content domain, cognitive skill level, and linguistic complexity, but the impact also was less than that of L1 family. First language orthography had no impact on two domains (ecology and anatomy and physiology), negligible or minor impact on two domains (evolution and biochemistry), and a small effect size on two domains (cell biology and genetics), where it contributed 2.4% and 2.8%, respectively, to the variance between groups. Likewise, a non-alphabetic L1 had a minor impact on linguistic complexity performance and contributed 1.9% of the difference on high linguistic complexity items, followed by 1.1% on low linguistic complexity items, and 0.8% on items with medium linguistic complexity. With respect to item cognitive skill, ELLs with a non-alphabetic L1 had statistically significant higher performance on items at all three cognitive skill levels (foundational, conceptual, and application); however, this had little practical impact since it contributed approximately 1% of the variance. These results suggested that a Latinate L1 and an alphabetic orthography had confounding effects.

The review of the literature did not find studies that specifically explored the impact of L1 orthography on ELL performance on wide-scale science assessments, so a comparison to previous research findings was not possible. Earlier studies have, however, explored L1 orthographic impact on L2 reading and found positive L1-L2 transfer of

phonological skills when both languages were alphabetic; phonological skills are requisite for English literacy and important when encountering unknown or low-frequency words (Wang, Koda, & Perfetti, 2003). This study found that ELLs with a non-alphabetic L1 performed better, especially at MEPA Level 3 and above, which suggested that at MEPA Level 3, there was no additional cognitive load from L1-L2 orthographic distance and that resources were available for content and attention to other item attributes. This finding supported Wang, Koda, and Perfetti's (2003) postulation that as English proficiency increases for ELLs with a non-alphabetic L1, phonological processing becomes dominant in their approach to reading English.

With respect to cognitive load theory, however, this study's findings are inconclusive. Cognitive load theory suggests that the cognitive load for ELLs with a non-alphabetic L1 might lead to decreased performance; however, the opposite was observed in this study: Non-alphabetic L1 ELLs had statistically significant higher performance for overall MCAS performance, some content domain performance, cognitive skill performance, and high linguistic complexity performance. Cognitive load from a non-alphabetic L1 may not have been an issue in the context of an untimed assessment, during which test-takers can write notes (such as translations into the L1) in the test booklet. Another possible explanation is that after a certain level of English proficiency, automaticity with the alphabetic principle makes moot any cognitive load from L1-L2 orthographic distance. An alternative interpretation is that the findings support the cognition hypothesis: The L1-L2 orthographic distance led to higher noticing and attending and, in turn, to higher performance. The possible support for the cognition

hypothesis, however, is weakened by the inconsistent impact of L1 orthography across the content domains.

### **Late-Entry ELL Impact**

This study found that late-entry ELL status was neither a predictor of overall Biology MCAS performance nor a predictor for performance on the item attributes of content domain, cognitive skill, and linguistic complexity. As discussed in Chapter 4, the not-late-entry ELLs had a mean English proficiency one level higher than late-entry ELLs. Since this study found that English proficiency was a strong predictor of Biology MCAS performance, it was expected that late-entry ELLs would have lower performance. The data, however, indicated no statistically significant difference at the  $p < .05$  level in performance between the two groups. This suggested a mitigating factor for late-entry ELLs. Though more substantive study is needed regarding this unexpected finding, one interpretation supports the common underlying proficiency model (Cummins, 2000). Biology content knowledge in the first language may have mitigated the lower English proficiency for late-entry ELLs, and thus, their performance was similar to that of not-late-entry ELLs with a higher mean English proficiency level.

As discussed in Chapter 2, Hakuta, Butler, and Witt (2000) found that it took four to seven years to achieve academic language proficiency in English for students who began in kindergarten, but the literature demonstrates little consensus about older second language learners (see Long, 2007, Chapter 3).<sup>138</sup> Snow and Hoefnagel-Höhle (1978)

---

<sup>138</sup> Hakuta, Butler, and Witt (2000) used existing data from two Canadian ELL studies and new data from two U.S. schools to explore English proficiency as a function of length of exposure to English. The Canadian datasets did not address the issue of socioeconomic status (SES). The U.S. datasets, however, did show that SES was an important variable in predicting the rate of becoming proficient in English. The current study did not have SES data.



found that the 12- to 15-year-old group exhibited the most rapid acquisition of the language skills tested compared to younger groups. However, Collier (1987a) found that the 12- to 15-year-old group exhibited the most difficulty and took the longest—six to eight years—to achieve parity with native speakers. These studies explored the rate of language acquisition. Although this study explored performance on a content assessment, the data indicated support for Snow and Hoefnagel-Höhle's (1978) finding when the sample was disaggregated into two subgroups based on English proficiency: (1) MEPA Levels 1 to 2, and (2) MEPA Levels 3 to 5. For MEPA Levels 1 to 2, the passing rate of late-entry ELLs (22.7%) was twice that of not-late-entry ELLs (10%). When late-entry ELLs possessed sufficient English proficiency to access the assessment (MEPA Levels 3 to 5), nearly two-thirds (64.8%) passed, with 21.3% scoring Proficient or higher. In contrast, 56% of the not-late-entry ELLs at MEPA Levels 3 to 5 passed, with 13.4% scoring Proficient or higher. Although performance was statistically significant at the  $p < .05$  level, partial eta-squared for both subgroups determined that late-entry ELL status had a minimal impact on performance, contributing 1% of the variance at MEPA Levels 1 and 2, and 1.4% of the variance at MEPA Levels 3 to 5. This study was limited to age of entry at 12 years or later. Future studies that account for age of entry and years of English instruction might produce results that lend more support to Snow and Hoefnagel-Höhle's findings.

### **Implications of the Study**

Though this study analyzed a highly specific subset of data, the findings have broad implications for policy, practice, and future research. The data informed analyses about when ELLs experienced meaningful participation on the instrument, pointed to

practice areas that might increase ELL academic success, and suggested areas that warrant future study.

**Policy implications.** The Biology MCAS is a competency exam, serving as one of several high school graduation requirements in Massachusetts. Yet, there is a dialectical tension between the Commonwealth's interest in ensuring a meaningful education to prepare high school students for post-secondary career and college opportunities and what the research shows as feasible for students who are not proficient in English. On the one hand, the STE MCAS graduation requirement ensures that Massachusetts high school graduates, including ELLs, possess grade-level knowledge and skills in science and technology. On the other hand, the data showed that the majority of ELLs pass the Biology MCAS at MEPA Levels 4 and 5, and that even though they have passed, there still exists a performance gap. Some have argued that content exams should be linguistically simplified for ELLs (see Abedi & Hejri, 2004). Linguistic simplification might be a more accurate assessment of content knowledge; however, without the appropriate level of academic discourse, a linguistically simplified content exam would still not measure grade-level competency. The underlying question, then, is "What does a Massachusetts high school diploma mean in terms of knowledge and skills?" Does it imply the acquisition of subject matter knowledge or, rather, subject matter knowledge in English at a level requisite to be an informed citizen and to access post-secondary educational and workplace opportunities? Though this question goes beyond the scope of this study, it is a crucial one for the various stakeholders in Massachusetts public schools to address.

I do not suggest exempting ELLs from STE competency exams. That is a slippery slope, and if ELLs were exempt from the STE MCAS high school graduation requirement, this group might get left behind. ELLs must have access to grade-level content knowledge. In the past, I have seen ELL biology classes taught by an ESL teacher who had not studied biology since his or her own high school biology class. The standards-based STE MCAS exams have focused teachers on ensuring that all students in their classes, including ELLs and former ELLs, have access to the same grade-level curricula. MCAS accountability for ELLs has made school administrators recognize the importance of dually certified (subject matter and ESL) teachers for ELLs in content classes. Project SAEL (Successful Advancement of English Learners in Gateway Cities in Massachusetts; <http://www.projectsael.org>) is one example of this growing systemic awareness that ELLs need access to teachers who are trained in both ESL and science, mathematics, technology, and engineering. Project SAEL is a federally funded program at Salem State University that prepares undergraduate STEM and education majors, as well as in-service teachers, to teach ELLs either through the SEI Endorsement or a Graduate Licensure Only ESL program.<sup>139</sup> The graduation requirement of passing an ELA, math, and STE MCAS has resulted in resources, especially teacher professional development and the hiring of highly qualified dual-certified teachers, to ensure that ELLs not only access the curricula but also graduate with the requisite skills to pursue post-secondary educational and employment opportunities.

---

<sup>139</sup> Massachusetts requires all core academic teachers who do not have an ESL/ELL license and have one or more ELLs in their classes to earn the Sheltered English Instruction endorsement to their teaching license by July 1, 2016 (<http://www.doe.mass.edu/retell>).

Although I believe that all ELLs should take the MCAS exams, I also believe that accountability measures need to be reviewed both at the federal and state levels. Schools must provide ELLs with a meaningful education—that is, schools need to provide instruction not only in English language development (ELD) but also in grade-level content. Title III is a federal grant program reauthorized under NCLB that provides funds to districts to support programs for ELLs (Boyle, Taylor, Hurlburt, & Soga, 2010), and it requires state ELL accountability measures known as Annual Measurable Achievement Objectives (AMAOs). There are three AMAO goals; the first two focus on ELD, and the third focuses on MCAS performance (for Massachusetts).<sup>140</sup> In the 2007-2008 school year, 95% of ELLs in Massachusetts attended schools in districts that received Title III funding (Boyle et al., 2010). School systems may meet their first and second AMAO goals with respect to English proficiency but still struggle to meet the third goal for ELL MCAS performance.<sup>141</sup>

The data from this study indicated that English proficiency is a strong predictor of Biology MCAS performance, accounting for 29% of the variance among ELLs. The data further indicated that only after attaining MEPA Level 4 English proficiency do the majority of ELLs pass the Biology MCAS. Content assessment accountability measures, such as the third AMAO goal, should take into account that ELLs at the equivalent of MEPA Levels 1 through 3 may not have sufficient English proficiency to demonstrate

---

<sup>140</sup>The first AMAO is the annual increase in the number or percentage of students making progress in learning English, and the second is annual increase in the number or percentage of students attaining English proficiency. The third AMAO is making AYP for ELLs (Tanenbaum et al., 2012).

<sup>141</sup> It should be noted that the ELL subgroup for accountability differs from other accountability subgroups in that the composition is constantly changing. Higher performing members are taken out when they reach English proficiency and replaced by ELLs who are entering with minimal or lower English proficiency. This is unlike most subgroups (e.g., racial) where members do not cycle out because they no longer meet the definition.

content knowledge, and therefore schools should not be penalized. All ELLs should have the opportunity to take the science MCAS exams to meet this graduation requirement, especially since approximately 19% of the Level 1 ELLs, 24% of the Level 2 ELLs, and 47% of the Level 3 ELLs pass the Biology MCAS. The policy implication is how the Biology MCAS results for ELLs should be used to calculate Annual Yearly Progress (AYP) and the third AMAO 3 because the data suggested that English proficiency introduced sufficient construct-irrelevant variance to question the validity of the assessment for Level 1 to 2 ELLs.

Validity concerns for testing ELLs on standardized content exams before they attain English proficiency are not limited to the literature. In a study of the implementation of Title III across states, seven of 12 case-study districts voiced similar validity concerns in their testing of ELLs on content exams (Tanenbaum et al., 2012). This study demonstrated that English proficiency was a source of construct-irrelevant variance on the item attributes of content domain, cognitive skill level, and linguistic complexity. It also showed that the majority of ELLs passed the Biology MCAS once they attained MEPA Level 4 English proficiency. It is unreasonable to expect the majority of ELLs who lack academic language to pass the Biology MCAS. Likewise, it is also unreasonable to hold a school or district accountable for test results where the instrument was not valid. As discussed previously, Massachusetts requires all ELLs to take the Biology MCAS; however, there is no scaled score or performance level for ELLs who are in their first year in the United States, and these test results are not included in AYP. This first-year exemption is based on time, not when the instrument is valid. In the sample, approximately 43% of ELLs in their first year were at Levels 1 and 2. The first-

year exemption from AYP, however, is not sufficient. For ELLs who were not in their first year, 15.6% were at Levels 1 and 2, and an additional 45.6% were at Level 3; their MCAS scores could be counted for AYP.

I believe that with respect to the STE MCAS exams, accountability should be for Level 4 and Level 5 ELLs.<sup>142</sup> The data suggested that ELLs at these English proficiency levels had meaningful participation on the Biology MCAS, and the majority passed. In addition, the accountability measures should note that even at these higher English proficiency levels, ELLs are not English proficient, and thus, achieving Proficient or Advanced performance levels is still a challenge. Such a policy shift would not pose a slippery slope resulting in the educational needs of ELLs being ignored because schools would still be held accountable for ELL content assessment at the higher end of English proficiency, and the first two AMAO goals ensure accountability for ELLs reaching these higher English proficiency levels.

The question arises about alternative competency determinations for Level 1 and 2 ELLs. As previously stated, I believe that all ELLs should take the Biology MCAS. For the Level 1 and Level 2 ELLs who fail, the data suggested that it is just a matter of time until their academic language proficiency is sufficient, and then they will pass. I have heard colleagues debate whether Level 1 or Level 2 ELLs should take a science class because at those English proficiency level, they probably will not pass a science MCAS. Some of these ELLs undoubtedly are required to take the STE MCAS despite having no access to the curriculum. This inequitable access to grade-level curricula is just one

---

<sup>142</sup> In January 2013, Massachusetts began using the ACCESS instrument to assess English proficiency. The corresponding ACCESS levels should be the point where ELL MCAS Biology accountability should begin.

example of how the slippery slope can begin if some ELLs are not required to take the MCAS.

Some have suggested a portfolio or other type of assessment. As stated in Chapter 1, alternative and performance science assessments were beyond the scope of this study. Massachusetts, however, has two alternative routes to a competency determination: (1) MCAS Performance Appeal and (2) MCAS Portfolio Appeal (MA DESE, 2014). Both alternative routes require that a student take the MCAS exam.<sup>143</sup> The Performance Appeal uses a student's course grades in relation to a cohort of students who took the same course(s) and scored between 220 and 228 (Needs Improvement) on the same exam. "In a small number of cases, it may not be possible to use a student's course grades for the purpose of filing an MCAS cohort appeal ... [and an] MCAS Portfolio Appeal consist[s] of the student's current or cumulative work in the content area of the appeal" (MA DESE, 2014). I have observed colleagues prepare ELA Portfolio Appeals, and it required many hours for a single appeal. If policy were to change and allow Portfolio Appeals in lieu of taking the MCAS exam for Level 1 and 2 ELLs, this would place a burden on ELL teachers, especially in urban districts where resources are already limited.

It must be remembered that the science MCAS is only one of three MCAS graduation requirements. MEPA Performance-Level Descriptors (Appendix B) describe a Level 1 ELL as a student who "cannot yet communicate in English, and errors almost always interfere with communication" and a Level 2 ELL as a student who "communicates using simple written and spoken English at school, with errors that often interfere with communication and understanding" (MA DESE, 2012f, p. 5). It is unlikely

---

<sup>143</sup> There are other requirements such as passing the subject matter course and attendance.

that ELLs at Levels 1 and 2 would pass the ELA MCAS, especially since academic language begins with Level 3. In my practice, many of my ELL students passed the science before passing ELA MCAS. Level 1 and Level 2 ELLs may not pass the additional ELA and mathematics ELA MCAS exams until the following year or two later when their English proficiency, including academic language is sufficient, and at which point they would also likely pass the science MCAS. Thus, developing an alternative assessment for Level 1 and 2 ELLs would require resources at the state, district, school, and classroom level and would not guarantee graduation by itself. This also goes to the issue raised at the beginning of this section: Do stakeholders want ELLs graduating from high school when they cannot communicate in English? How is this giving them equitable opportunities?

**Practice implications.** This study's goal was to analyze ELL Biology MCAS performance to inform the nature of the achievement gap reported as a single statistic. The data indicated that most ELLs experienced meaningful participation at MEPA Level 4 or higher. By definition, at this English proficiency level, ELLs have acquired sufficient academic language to access content in grade-level texts. This reinforces the importance of providing ELLs with focused, explicit instruction to accelerate academic language development. The data also indicated that the lack of academic language at Levels 1 and 2 was accompanied by only approximately 20% passing. The greatest increase in percent passing was seen when progressing from Level 2 to Level 3 (23.8% to 46.8%, an increase of 96.6%). Clearly, getting ELLs to Level 3 proficiency is critical, but it is just as critical that they attain Level 4 proficiency so that most of them can pass this high-stakes assessment; in moving from Level 3 to Level 4, the passing rate increased from 46.8% to



71.8%, an increase of 53.4%. Equally critical for secondary ELLs is to incorporate academic language into classroom discourse from the beginning. For example, Level 1 and 2 ELLs should learn *autotroph* as well as the simpler word *producer*. When talking about enzymes, teachers should use *optimal* as well as *best*. If academic language is developed alongside content in the SEI classroom, it may be disjointed at first, but as English development progresses, the students at the lower end of English proficiency will have a more robust vocabulary, though probably robust only in the context of the particular content area until they reach Level 3 and beyond.

The MA DESE guidelines decrease English language development (ELD) instructional minutes from 2.5 hours a day for Level 1 and 2 ELLs to one to two hours a day for Level 3 ELLs (MA DESE, 2012j).<sup>144</sup> In my experience, some school systems integrate Level 3 ELLs in general education with push-in language support. This may be because ELLs at Level 3 are beginning to access grade-level content and can participate in classwork with native-speaking peers, or perhaps the decrease in ELD instructional minutes is misconstrued as ELD becoming less important than content. I have observed that many push-in teachers shelter the grade-level content, but they do not focus on ELD as a separate curriculum. To be fair, it is hard to be a push-in teacher sitting with a group of ELLs and delivering instruction on a curriculum different from what is going on in the rest of the room. Too often, ELL push-in teachers take on the role of a para-professional, not an instructor. Push-in models can work if there is adequate common planning time

---

<sup>144</sup> The instructional minutes recommended by MA DESE for MEPA Levels 1-5 were the same as the current ACCESS Levels 1-5.

and synergy between the content and ELL teachers, but the focus is naturally on content in a content classroom.

The data in this study suggested that continued, focused ELD instruction to move ELLs from Level 3 to Level 4 will ultimately result in a majority of them passing the Biology MCAS, and continued ELD focus will increase the percentage of ELLs who reach the Proficient and higher performance levels. Massachusetts schools, however, face a challenge with respect to an ELD curriculum. In 2003, Massachusetts published its *English Language Proficiency Benchmarks and Outcomes for English Language Learners* (ELPBO) (MA DESE, 2003), yet these standards have not been in use since Massachusetts adopted the Common Core. Nothing has replaced the ELPBO as a curriculum framework for ELD; the WIDA standards are not standards in the traditional sense but rather indicators of what students are able to do at different proficiency levels. Since Massachusetts has such well-defined frameworks for content areas, this transition period while MA DESE develops Common Core-aligned ELD standards is challenging to school districts, schools, and individual teachers who are faced with developing their own ELD curriculum. In speaking with ELL colleagues in several districts, I noted among them a “wait and see” attitude in light of the time and resources needed for a district to develop curricula. Given that the data indicated that English proficiency had strong impact on Biology MCAS performance and that reaching Level 5 essentially closed the passing gap and narrowed the performance level gap, there is a significant need for a well-defined ELD framework to replace the ELPBO.

The data indicated that ELLs with a Latinate L1 had consistently lower performance when compared to non-Latinate L1 ELLs, and this subgroup appeared at

risk for academic failure beyond their ELL status and English proficiency level. Although this finding warrants more research, the implications are that schools and teachers should not assume that Latinate-English cognates confer an advantage and that vocabulary development is not as critical as it is for non-Latinate L1 ELLs. The data showed that irrespective of a cognate advantage (which was beyond the scope of this study), Latinate L1 ELLs performed at consistently lower levels. Prior educational experiences and L1 literacy could be a hidden factor, and schools need to identify students that may have gaps in their education. Schools must also look at other factors (such as attendance and poverty) that may impact academic achievement for this subgroup. The data showed that even if targeted policies and programs are in place, they may not be enough. Future study for this subgroup may inform practice so that all ELLs are supported for academic success.

The data indicated two at risk groups: (1) Level 1 and Level 2 ELLs, and (2) ELLs with a Latinate L1. The data showed that approximately one-fifth of the Level 1 ELLs and one-fourth of the Level 2 ELLs passed the Biology MCAS. So even at these lower levels of English proficiency, ELLs can pass the MCAS. Content teachers should keep their expectations high and help Level 1 and Level 2 ELLs by having a dual purpose to their lessons, whereby they teach not only content but also scaffold and promote academic language in everyday classroom interactions. The data further suggested that teachers should not be discouraged if a Level 1 or Level 2 ELL does not pass the Biology MCAS. As language proficiency increases, they will more than likely pass; classroom expectations should not be lowered. For these ELLs, it is a matter of English proficiency, which is a function of time. The second at-risk group, ELLs with a Latinate L1, need

further investigation. The data suggested that teachers cannot assume L1-L2 cognates are known. Perhaps teaching cognate strategies and making explicit reference to the L1 cognates might be a place to start.

**Future research implications.** Analysis of ELL performance elucidated areas for further investigation. A Hispanic or Latino achievement gap on standardized assessments is well-established in the literature (Murphy, 2010). This study showed lower performance for ELLs with a Latinate first language (i.e., Spanish, Haitian Creole, Portuguese, French, and Italian) when compared to ELLs with a non-Latinate first language, and the data indicated that lower performance persisted within English proficiency levels. This study only explored performance differences between these groups for two English proficiency subgroups: (1) MEPA Levels 1 and 2, and (2) MEPA Levels 3 to 5. Future research should investigate the performance of the Latinate L1 group by first language and within each English proficiency level. Additionally, studies that account for language proficiency, L1 literacy, prior schooling, and years of English instruction may inform the nature of the Latinate L1 gap and support Martiniello's (2008) finding for cognates as a reading comprehension strategy for Latinate L1 ELLs on MCAS items. Future studies might also explore whether some non-Latinate languages adopt or adapt the English version of scientific terms and whether this extends a Tier III cognate advantage to some non-Latinate L1 ELLs.

This study found that first language family (Latinate/non-Latinate) and first language orthography (alphabetic/non-alphabetic) impacts were generally similar but diminished for orthography, often to a point of negligible impact. The data also suggested that L1 orthography impact was the greatest when items had high linguistic complexity;

therefore, additional study is needed on L1 orthography and cognitive load in the context of standardized assessments. The diversity of ELLs in this sample resulted in languages that were linguistically distant but which used the Roman alphabet. Orthography's lesser impact than first language family (Linate/non-Linate) argues for combining the two first language characteristics into three groups for future research: (1) Linate L1; (2) non-Linate, alphabetic L1; and (3) non-Linate, non-alphabetic L1.

That there was no statistically significant difference in performance between late-entry ELLs and not-late-entry ELLs, despite the latter having higher levels of English proficiency, is a finding that warrants more study. The data suggested that a fully acquired L1 and higher meta-linguistic awareness did not confer an advantage for items with a high linguistic complexity (see Snow & Hoefnagel-Höhle, 1978). Future studies should explore L1 literacy and L1 schooling experiences, as well as years of English instruction, when examining the impact of age of entry on MCAS performance. Biology content knowledge in the first language may have mitigated the lower English proficiency for late-entry ELLs, and thus their performance was similar to that of not-late-entry ELLs with a higher mean English proficiency level.

The primary finding of this study was the extent of English proficiency impact, not only on overall Biology MCAS performance but also on content domains, cognitive skill level, and linguistic complexity. More research is needed to determine if these findings can be replicated with other Biology MCAS test administrations or with ELL performance on the Physics MCAS—the STE MCAS with the next highest percentage of test-takers. This study explored the individual impact of learner (i.e., English proficiency and late-entry ELL status) and first language (i.e., language family and orthography)

characteristics. The next step is to study interactions between these factors, including the effect late-entry ELL status might have on and within these interactions.

Task complexity data for cognitive skill level and item linguistic complexity was inconclusive. The data were inconsistent with respect to cognitive load theory. The cognition hypothesis offered a possible explanation for the high task complexity data, though any support for the cognition hypothesis was tentative and limited. Motivation plays a role in increased attending and noticing with higher complexity tasks. A future qualitative study could explore the motivation of ELLs with respect to the Biology MCAS and at what point frustration could negate motivational impact. Such a study could inform the inconsistencies of this study's findings.

This study explored the nature of the ELL achievement gap within a sample of ELLs on one administration of a standardized assessment. Further investigation is needed into the transiency or persistence of the gap. Future studies should also compare the performance of ELLs, former ELLs, and never ELLs. Former ELLs would include not only those students who are designated as FLEP (former limited English proficient) but also those former ELLs who are beyond the two-year monitoring window; the latter are included with non-ELLs for reporting purposes. A longitudinal study of ELL performance over the Grade 5, Grade 8, and high school science MCAS exams would show ELL performance on standardized science assessments over time as English proficiency increased. Although the starting cohort would consist only of ELLs, some ELLs would be English proficient by the Grade 8 or the high school science exam (Grade 9 or Grade 10). Such studies could inform whether ELLs who entered in the lower grades and were reclassified as English proficient achieved academic parity with native speakers

in high school. They could also inform whether some former ELL subgroups continued to remain at risk despite attaining English proficiency.

### **Validity, Reliability, and Limitations**

As a secondary data analysis, this study inherited the validity and reliability of the underlying data. As discussed in Chapter 3, the Biology MCAS was a valid and reliable instrument that assessed the Massachusetts Biology standards. Likewise, the MEPA was a valid and reliable instrument that assessed English proficiency level. The ELL data in this study are based on English proficiency level designations, including re-designation as English proficient, used uniformly throughout Massachusetts. By using state-level data, this study addressed the reliability concerns raised by Abedi et al. (2004) and Solorzano (2008) with respect to the varying ways that states define English proficiency and re-designate ELLs as English proficient.

**Threats to construct validity.** There were validity threats in this study arising from the constructs of late-entry ELLs and linguistic complexity. The study disaggregated the late-entry ELL subgroup by assuming that the number of years enrolled in Massachusetts public schools was a proxy for years in the United States (which was used to calculate age of arrival). Transiency in the ELL population may have affected accurate disaggregation of late-entry ELLs. This study acknowledges this threat to its construct of late-entry ELLs.

Linguistic complexity was a proxy for the construct of academic language, which comprises lexical, syntactic, and discourse elements. At the lexical level, this study defined lexical density as the number of words. All words, however, are not equal; some (e.g., nominalizations) are more informative than others (e.g., articles). Some words

occur in the everyday register (e.g., *both*), while others occur in the academic register (e.g., *illustrations*), and some are content-specific (e.g., *prokaryotic*). At the syntactic level, this study used the mean number of words per sentence as a proxy for syntactic complexity. Although the number of words per sentence is an element of syntactic complexity, the construct represents more than just how many words are in a sentence (Kucer, 2005, Chapter 6). This study used the estimated Lexile<sup>®</sup> reading complexity scores as a proxy for academic discourse. The Lexile Analyzer<sup>®</sup> uses a proprietary algorithm based on native speaker reading levels; however, reading level and discourse are not synonymous. Words, including content-specific vocabulary, that are below grade-level for a native speaker may be unknown to ELLs, and thus the reading complexity may be higher for ELLs. In addition, the Lexile Analyzer<sup>®</sup> can only analyze text, and Biology MCAS items included diagrams, illustrations, labels, or presented information in a tabular format. This study minimized these construct-validity threats to by normalizing values and treating the calculated composite linguistic complexity as a continuum that was categorized as low, medium, and high. This study further minimized validity threats by using the entire sample of ELLs in Massachusetts who had June 2012 MCAS and spring 2012 MEPA scores.

**Limitations.** Construct confounding, which Shadish, Cook, and Campbell (2002) define as “failure to describe all the constructs ... result[ing] in incomplete construct inferences” (p. 73), was a limitation to the current study. The effects of poverty may confound ELL performance. Background knowledge has “overriding influence on the reading process” (Kucer, 2005, p. 120) and affects comprehension, and prior academic experience, such as limited or interrupted schooling, may also confound ELL



performance (see Abedi & Lord, 2001). This study did not explore the impact of factors—for example, poverty, L1 biology knowledge, and L1 academic language—that may confound ELL Biology MCAS performance. Because this study used secondary data, other limitations included the inability to determine whether the ELLs in the dataset had uniform access to (1) the biology curriculum, (2) teachers trained to teach ELLs and biology, and (3) the permitted accommodations on the Biology MCAS. As a secondary data analysis, this study could not account for confounding factors that may affect generalizability to all second language learners.

### **Summary**

This intent of this study was to describe and analyze the phenomenon of secondary ELL Biology MCAS performance beyond a single homogenized statistic. English proficiency appeared to be the linguistic factor with the most impact on Biology MCAS score: As English proficiency increased, Biology MCAS scores increased, as did performance within content domains, cognitive skills levels, and item linguistic complexity levels. As ELLs approached reclassification as English proficient, the gap for passing the Biology MCAS essentially closed, but the gap for attaining content proficiency remained, though it had narrowed. The indistinguishable performance for ELLs at the lower end of English proficiency (i.e., MEPA Levels 1 and 2) indicated that they did not understand the language of the test to have meaningful participation; this raises validity issues for using the Biology MCAS with these ELLs. This study also identified consistently lower performance for ELLs with a Latinate L1, and hinted at a mitigating factor for late-entry ELLs. This study demonstrated that ELL performance is

more complex than a single reported statistic and identified several areas that warrant further study.

APPENDIX A

JUNE 2012 BIOLOGY MCAS

---

XVIII. Biology, High School

## *High School Biology Test*

The spring 2012 high school Biology test was based on learning standards in the Biology content strand of the Massachusetts *Science and Technology/Engineering Curriculum Framework* (2006). These learning standards appear on pages 54–58 of the *Framework*.

The *Science and Technology/Engineering Curriculum Framework* is available on the Department website at [www.doe.mass.edu/frameworks/current.html](http://www.doe.mass.edu/frameworks/current.html).

Biology test results are reported under the following five MCAS reporting categories:

- Biochemistry and Cell Biology
- Genetics
- Anatomy and Physiology
- Ecology
- Evolution and Biodiversity

### **Test Sessions**

The high school Biology test included two separate test sessions, which were administered on consecutive days. Each session included multiple-choice and open-response questions.

### **Reference Materials and Tools**

The high school Biology test was designed to be taken without the aid of a calculator. Students were allowed to have calculators with them during testing, but calculators were not needed to answer questions.

The use of bilingual word-to-word dictionaries was allowed for current and former English language learner students only, during both Biology test sessions. No other reference tools or materials were allowed.

### **Cross-Reference Information**

The table at the conclusion of this chapter indicates each item's reporting category and the framework learning standard it assesses. The correct answers for multiple-choice questions are also displayed in the table.

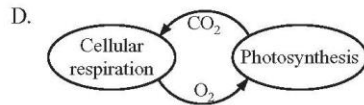
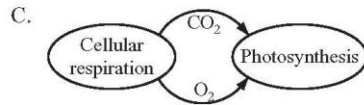
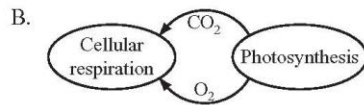
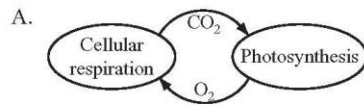
# Biology

## SESSION 1

### DIRECTIONS

This session contains twenty-one multiple-choice questions and two open-response questions. Mark your answers to these questions in the spaces provided in your Student Answer Booklet. You may work out solutions to multiple-choice questions in the test booklet.

- 1 Which of the following diagrams accurately represents the use of gases in both cellular respiration and photosynthesis?



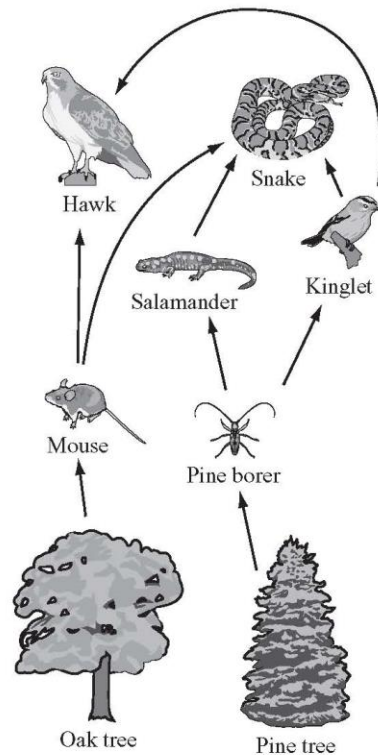
- 2 Hepatitis is a disease of the liver. Which of the following happens as a result of decreased liver function?

- A. Carbon dioxide builds up in the liver.
- B. Toxic compounds build up in the blood.
- C. The kidneys take over the functions of the liver.
- D. The stomach produces the enzymes needed for digestion.

- 3 A species of newt produces a toxin that can kill predators. Scientists have observed that some garter snakes can feed on the newts because they have a natural resistance to the toxin.
- In areas where populations of newts and garter snakes interact, which of the following predictions is **best** supported by evolutionary theory?
- A. The garter snakes with resistance to the toxin will successfully reproduce and pass the trait on to their offspring.
  - B. The garter snakes without resistance to the toxin will acquire resistance by increasing the rate at which they feed on the newts.
  - C. The newts that produce low levels of toxin will also develop camouflage adaptations that allow them to hide from the garter snakes.
  - D. The newts will stop making the toxin rather than continue to use energy to make a toxin that is ineffective against the garter snakes.

- 4 At one time, all the continents on Earth were joined in a supercontinent called Pangaea. Over time Pangaea split into separate continents.
- Which of the following statements describes a result of this split?
- A. All fossil evidence of species from Pangaea was lost.
  - B. Organisms on the separated continents no longer migrated for breeding.
  - C. Ancestral organisms evolved into different species on the separated continents.
  - D. Evolution in species proceeded more slowly on the separate continents than it had on Pangaea.

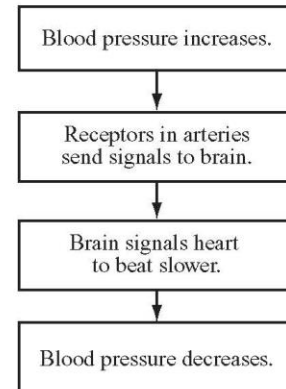
- 5 A food web is shown below.



Which of the following organisms compete for the mouse as a food source?

- A. hawk and snake
- B. snake and kinglet
- C. oak tree and pine tree
- D. pine borer and salamander

- 6 The diagram below shows one response pathway the human body uses to control blood pressure.



Which body systems work together in this response pathway to control blood pressure?

- A. digestive and nervous
- B. nervous and circulatory
- C. respiratory and digestive
- D. circulatory and excretory

- 7 The table below presents a variety of mRNA three-base sequences (codons) and the amino acids for which these sequences code.

First Base of mRNA	Second Base of mRNA	Third Base of mRNA	Amino Acid
G	A	A	glutamic acid
		C	aspartic acid
		G	glutamic acid
		U	aspartic acid
	G	A	glycine
		C	glycine
		G	glycine
		U	glycine
	U	A	valine
		C	valine
		G	valine
		U	valine

Based on the information in the table, which of the following changes is **least likely** to produce a phenotypic change in an organism?

- A. GAU to GGU
- B. GAU to GUU
- C. GAU to GAA
- D. GAU to GAC



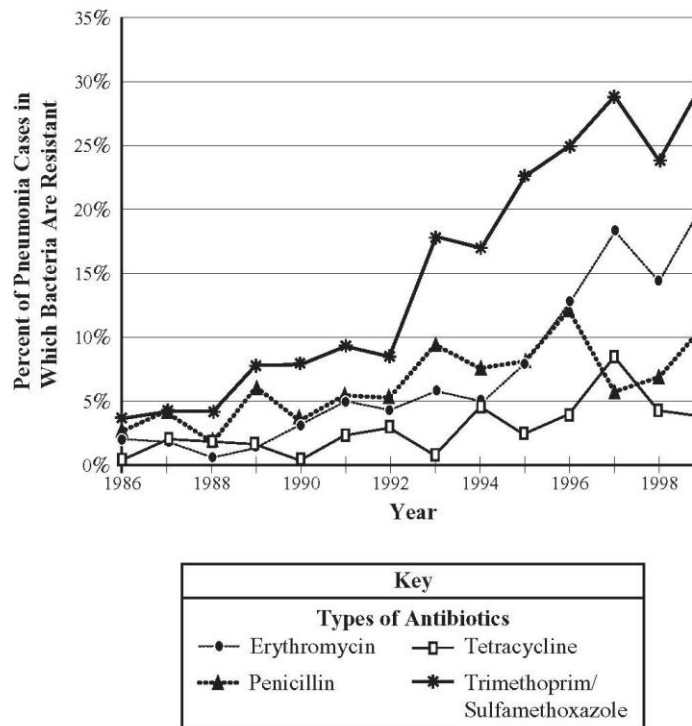
The following section focuses on bacterial resistance to several antibiotics.

Read the information below and use it to answer the four multiple-choice questions and one open-response question that follow.

One of the most important developments in modern medicine was the discovery of antibiotics. Antibiotics are used to treat infections caused by bacteria. However, strains of bacteria that are resistant to antibiotics are emerging. The rate of increase in infections caused by these antibiotic-resistant strains of bacteria is a concern for human health.

The bacterium *Streptococcus pneumoniae* is a major cause of the respiratory disease pneumonia. The graph below shows trends in bacterial resistance to different antibiotics in pneumonia cases from 1986 to 1999.

Trends in Bacterial Resistance



Mark your answers to multiple-choice questions 8 through 11 in the spaces provided in your Student Answer Booklet. Do not write your answers in this test booklet, but you may work out solutions to multiple-choice questions in the test booklet.

- 8 Antibiotics are helpful in treating an infection when the number of bacteria becomes too large for the body's immune system to fight on its own. What process enables the bacteria to multiply inside the body?
- A. binary fission
  - B. fertilization
  - C. meiosis
  - D. nitrogen fixation
- 9 Some antibiotics work by disrupting ATP production in bacteria. Which of the following will the bacteria lack when ATP production is disrupted?
- A. genetic material for reproduction
  - B. energy to perform life processes
  - C. nucleic acids to make proteins
  - D. cytoplasm to diffuse oxygen
- 10 When *Streptococcus pneumoniae* are exposed to an antibiotic, the bacteria try to pump the antibiotic out of their cells. Which of the following mechanisms is **most likely** used by the *Streptococcus pneumoniae* to pump the antibiotic out of their cells?
- A. active transport
  - B. diffusion
  - C. facilitated diffusion
  - D. osmosis
- 11 Resistance to antibiotics results from variations in the genetic code of *Streptococcus pneumoniae*. Which type of molecule encodes genetic information in *Streptococcus pneumoniae*?
- A. carbohydrate
  - B. fatty acid
  - C. nucleic acid
  - D. protein

Question 12 is an open-response question.

- BE SURE TO ANSWER AND LABEL ALL PARTS OF THE QUESTION.
- Show all your work (diagrams, tables, or computations) in your Student Answer Booklet.
- If you do the work in your head, explain in writing how you did the work.

Write your answer to question 12 in the space provided in your Student Answer Booklet.

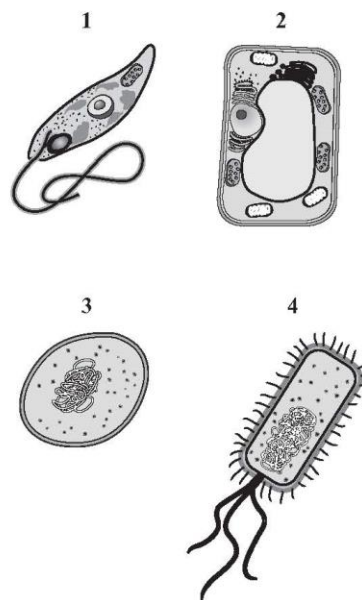
- 12 The graph shows the changes in antibiotic resistance of *Streptococcus pneumoniae* over time.
- a. Based on the graph, which antibiotic had *Streptococcus pneumoniae* become **most** resistant to by 1999?
  - b. Describe what usually happens to a population of *Streptococcus pneumoniae* immediately after it is exposed to a **new** antibiotic.
  - c. Explain, in detail, how antibiotic-resistant populations of *Streptococcus pneumoniae* develop over time as a result of the process of natural selection.

Mark your answers to multiple-choice questions 13 through 22 in the spaces provided in your Student Answer Booklet. Do not write your answers in this test booklet, but you may work out solutions to multiple-choice questions in the test booklet.

- 13 Long-tailed weasels and European otters are both classified into the family Mustelidae. Which of the following types of information was **most likely** used to classify these animals into the same family?
- A. food source
  - B. skeletal structure
  - C. location of habitat
  - D. method of movement

- 14 Which of the following is currently a primary cause of species decline worldwide?
- A. habitat destruction
  - B. intraspecific competition
  - C. random mating
  - D. viral outbreaks

- 15 Each of the illustrations below shows either a prokaryotic cell or a eukaryotic cell. Each cell is numbered.



(Not to scale)

Which two cells should be classified as prokaryotic cells?

- A. 1 and 2
- B. 1 and 3
- C. 2 and 4
- D. 3 and 4

- 16 Which of the following statements describes a DNA molecule?
- A. It contains the base uracil.
  - B. It has a double helix shape.
  - C. It contains five phosphate groups per nucleotide.
  - D. It has a backbone of twenty different nucleotides.
- 17 Height is a polygenic trait in humans. Which of the following statements **best** explains the genetics of this trait?
- A. Height is controlled by more than one gene.
  - B. Height is controlled by a single dominant gene.
  - C. The gene for height is located on the X chromosome.
  - D. The gene for height is located on the Y chromosome.

- 18 Spruce budworms are a type of moth. For every 100 budworm eggs, only about 1% reach adulthood. The table below shows the average number of budworms that survive and the main cause of death at each life cycle stage prior to the adult stage.

Stage in Life Cycle	Average Number Alive at Start of Stage	Main Cause of Death during Stage
egg	100	parasite
early larva	85	dispersal to unsuitable habitat
late larva	17	parasite, disease
pupa	2	parasite
adult	1	

Based on the data in the table, which of the following changes would **most** improve the percentage of budworms surviving to adulthood?

- A. a thinner cocoon wall in the pupal stage
- B. a slower rate of development in the late larval stage
- C. a decrease in exposure to disease in the pupal stage
- D. an increase in resistance to parasites during the egg stage

- 19 In mussels, the allele for brown coloring (**B**) is dominant, and the allele for blue coloring (**b**) is recessive. For which parental genotypes are 100% of the offspring expected to be blue?

A. **Bb** × **Bb**  
B. **BB** × **bb**  
C. **bb** × **bb**  
D. **BB** × **BB**

- 20 Specific DNA sequences called “promoters” provide binding sites for the enzyme that synthesizes RNA. Promoters are directly involved in which cellular process?

A. active transport  
B. crossing over  
C. replication  
D. transcription

- 21 Students digging near their school unearthed four objects. One of the objects was part of the exoskeleton of an insect. The table below shows the results of a chemical analysis of the objects.

Object	Chemical Composition
W	chlorine, sodium
X	oxygen, silicon
Y	carbon, hydrogen, nitrogen, oxygen
Z	aluminum, silicon, oxygen, hydrogen

Based on the chemical analysis, which object is most likely from the exoskeleton?

A. object W  
B. object X  
C. object Y  
D. object Z

- 22 In the human respiratory system, the contraction and relaxation of a muscle called the diaphragm helps move air through which of the following structures?

A. artery, capillary, and vein  
B. larynx, pharynx, and trachea  
C. atrium, trachea, and ventricle  
D. esophagus, kidney, and pharynx

Question 23 is an open-response question.

- **BE SURE TO ANSWER AND LABEL ALL PARTS OF THE QUESTION.**
- **Show all your work (diagrams, tables, or computations) in your Student Answer Booklet.**
- **If you do the work in your head, explain in writing how you did the work.**

Write your answer to question 23 in the space provided in your Student Answer Booklet.

- 23** Various types of evidence can be used to distinguish organisms in different kingdoms.
- a. Describe **two** ways to distinguish bacteria from protists, using cell structures or means of obtaining nourishment.
  - b. Describe **two** ways to distinguish fungi from plants, using cell structures or means of obtaining nourishment.



# Biology

## SESSION 2

### DIRECTIONS

This session contains nineteen multiple-choice questions and three open-response questions. Mark your answers to these questions in the spaces provided in your Student Answer Booklet. You may work out solutions to multiple-choice questions in the test booklet.

- 24 Scientific evidence shows that modern dogs, wolves, and foxes all have a common ancestor. Further evidence shows that dogs are more closely related to wolves than to foxes.

Which of the following observations provides the **best** evidence that dogs are more closely related to wolves than to foxes?

- A. The diets of dogs and wolves are more similar than the diets of dogs and foxes.
- B. The lifespans of dogs and wolves are more similar than the lifespans of dogs and foxes.
- C. The genetic sequences of dogs and wolves are more similar than the genetic sequences of dogs and foxes.
- D. The body sizes of dogs and wolves are more similar than the body sizes of dogs and foxes.

- 25 In which part of the human digestive system do both physical breakdown and chemical breakdown of food first begin?

- A. esophagus
- B. mouth
- C. large intestine
- D. small intestine

- 26 The human body regularly sheds and replaces its skin cells. Which of the following processes is directly responsible for replacing these cells?

- A. meiosis
- B. mitosis
- C. osmosis
- D. transcription

- 27 The diagram below shows a pair of DNA nucleotides. The nitrogenous base guanine (G) is labeled.

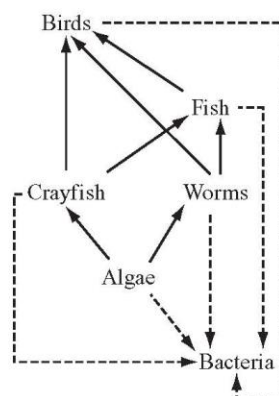


Which nitrogenous base pairs with guanine?

- A. adenine (A)
- B. cytosine (C)
- C. thymine (T)
- D. uracil (U)

- 28 An animal population decreases from 800 individuals to 600 individuals. Which of the following could explain this change in population size?
- A. The population size of the animal's predator increased.
  - B. The emigration rate of the animals from the population decreased.
  - C. The number of breeding pairs in the animal's population increased.
  - D. The number of species competing with the animal for food decreased.

- 29 Part of a marsh food web is shown below.



Which of the following statements correctly describes organisms in this food web?

- A. The birds are producers.
- B. The algae are consumers.
- C. The worms are carnivores.
- D. The bacteria are decomposers.

- 30 A student is preparing to run in a school track competition. For the quickest source of energy, the student should eat a food that contains a high percentage of

- A. carbohydrates.
- B. fat.
- C. proteins.
- D. sodium.

- 31 Which of the following is the **best** example of natural selection?

- A. The lifespan of a chimpanzee is extended to 60 years in captivity.
- B. The population size of giraffes changes over time as a result of immigration.
- C. The bone density of a human increases significantly as a result of participation in sports.
- D. The average toxin level in a poisonous frog population increases over time in response to high predation.

Question 32 is an open-response question.

- BE SURE TO ANSWER AND LABEL ALL PARTS OF THE QUESTION.
- Show all your work (diagrams, tables, or computations) in your Student Answer Booklet.
- If you do the work in your head, explain in writing how you did the work.

Write your answer to question 32 in the space provided in your Student Answer Booklet.

- 32 The illustrations below show a smooth muscle cell and a skeletal muscle cell.



Smooth muscle cell



Skeletal muscle cell

- Identify one location where smooth muscle is found in the human body **and** whether smooth muscle is under voluntary or involuntary control.
- Identify one location where skeletal muscle is found in the human body **and** whether skeletal muscle is under voluntary or involuntary control.

The third type of muscle in the human body is cardiac muscle.

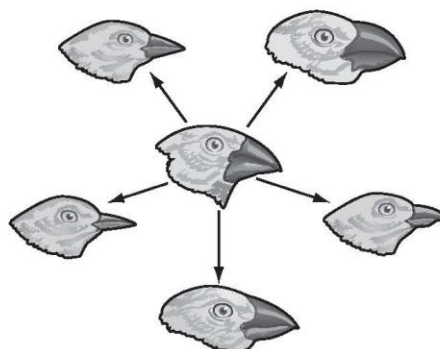
- Identify whether cardiac muscle is more similar to smooth muscle or skeletal muscle. Provide **two** reasons to support your answer.

Mark your answers to multiple-choice questions 33 through 43 in the spaces provided in your Student Answer Booklet. Do not write your answers in this test booklet, but you may work out solutions to multiple-choice questions in the test booklet.

- 33 The number of monarch butterflies counted in one location in the western United States dropped from 354,300 to 50,853 over a 10-year period. Which of the following statements **best** explains the drop in the number of monarch butterflies counted?

- A. The death rate was greater than the birth rate.
- B. The emigration rate was greater than the death rate.
- C. The birth rate was greater than the immigration rate.
- D. The immigration rate was greater than the emigration rate.

- 34 The diagram below shows many finch species that originated from a single ancestral finch species in the Galápagos Islands.



Which of the following statements best explains why many different finch species originated from the single ancestral species?

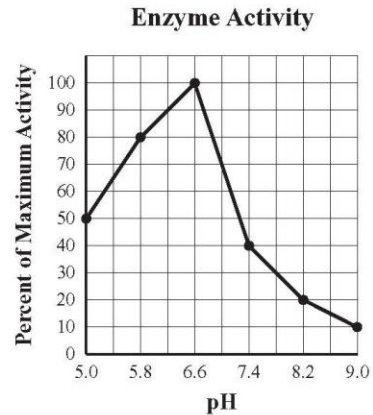
- A. Populations adapted to environmental pressures.
- B. Recessive traits in populations were eliminated over time.
- C. Individuals acquired unique characteristics during their lifetimes.
- D. Random mutation caused some individuals to have harmful traits.

- 35 Many lichens are composed of fungi and algae. The fungi get sugars from the algae, and the algae get water, minerals, and proteins from the fungi.

Which of the following terms **best** describes the relationship between the organisms in the lichens?

- A. commensalism
- B. competition
- C. mutualism
- D. parasitism

- 36 The graph below shows how the activity of an enzyme changes over a range of pH values.



Which of the following conclusions is supported by the data?

- A. The optimum pH of the enzyme is 6.6.
- B. The optimum pH of the enzyme is 5.8.
- C. The enzyme's activity is greater around pH 8.0 than around pH 5.0.
- D. The enzyme's activity continually increases as pH increases from 5.0 to 9.0.

- 37 In guinea pigs, the allele for black hair (**B**) is dominant to the allele for brown hair (**b**). Two black-haired guinea pigs are crossed. One of the guinea pigs is homozygous for black hair and one is heterozygous.

What percentage of the offspring are expected to have black hair?

- A. 25%
- B. 50%
- C. 75%
- D. 100%

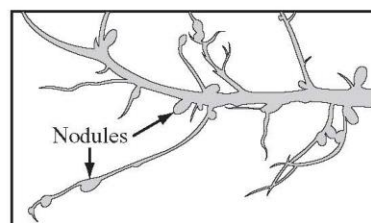
- 38 Acetylcholine is a neurotransmitter in the human body. As a neurotransmitter, acetylcholine is directly responsible for which of the following?

- A. speeding up the rate of biochemical reactions in cells
- B. assisting in the transport of nutrients in the bloodstream
- C. carrying the signal for a nerve impulse from one neuron to the next
- D. facilitating diffusion of amino acids across the plasma membrane of cells

- 39 Which of the following statements is correct about the hierarchy of the taxonomic system currently used to classify organisms?

- A. All organisms of a given order belong to the same species.
- B. Many different classes of organisms belong to the same order.
- C. All organisms of a given phylum belong to the same kingdom.
- D. Many different families of organisms belong to the same genus.

- 40 The illustration below shows part of a clover root system. Two root nodules are labeled.



The nodules contain which of the following to fix nitrogen for the plant?

- A. bacteria
- B. gases
- C. hormones
- D. worms

- 41 The table below shows the genotypes that result in four different blood types in humans.

Genotype	Blood Type
$I^A I^A$ , $I^A i$	A
$I^B I^B$ , $I^B i$	B
$I^A I^B$	AB
ii	O

Based on the information in the table, which of the following describes alleles  $I^A$  and  $I^B$ ?

- A. The  $I^A$  and  $I^B$  alleles show sex linkage.
- B. The  $I^A$  allele is recessive to the  $I^B$  allele.
- C. The  $I^A$  allele is dominant to the  $I^B$  allele.
- D. The  $I^A$  and  $I^B$  alleles show codominance.

- 42 Whale fins and bat wings are anatomically similar. Which of the following does this suggest about the animals?

- A. Whales and bats move in the same way.
- B. Whales and bats have a common ancestry.
- C. Whales and bats have existed for the same amount of time.
- D. Whales and bats were once adapted to the same environment.

- 43 An amoeba in a pond engulfs and consumes a paramecium. The amoeba uses which of the following to quickly break down the organic molecules in the paramecium?

- A. enzymes
- B. glucose
- C. polysaccharides
- D. water



Questions 44 and 45 are open-response questions.

- **BE SURE TO ANSWER AND LABEL ALL PARTS OF EACH QUESTION.**
- **Show all your work (diagrams, tables, or computations) in your Student Answer Booklet.**
- **If you do the work in your head, explain in writing how you did the work.**

Write your answer to question 44 in the space provided in your Student Answer Booklet.

- 44** In tomato plants, the allele for red fruit color (**R**) is dominant to the allele for yellow fruit color (**r**). The allele for round-shaped fruit (**F**) is dominant to the allele for pear-shaped fruit (**f**).

Two tomato plants, heterozygous for fruit color and fruit shape, are crossed. The Punnett square for this dihybrid cross is shown below.

	<b>RF</b>	<b>Rf</b>	<b>rF</b>	<b>rf</b>
<b>RF</b>	RRFF	RRFf	RrFF	RrFf
<b>Rf</b>	RRFf	RRff	RrFf	Rrff
<b>rF</b>	RrFF	RrFf	rrFF	rrFf
<b>rf</b>	RrFf	Rrff	rrFf	rrff

- For this cross, identify all the possible phenotypes of the offspring.
- Considering only fruit color, determine the ratio of offspring with red fruit to offspring with yellow fruit predicted by the Punnett square.
- Considering only fruit shape, determine the ratio of offspring with round-shaped fruit to offspring with pear-shaped fruit predicted by the Punnett square.
- Explain what is meant by independent assortment **and** describe one way in which your answers to parts (a), (b), and (c) support the conclusion that the genes for fruit color and fruit shape sort independently.



**Write your answer to question 45 in the space provided in your Student Answer Booklet.**

**45** The trees in tropical rain forests are important to nutrient cycling in the biosphere.

- a. Describe one role of the trees in the carbon cycle.
- b. Describe one role of the trees in the oxygen cycle.
- c. Describe one role of the trees in the water cycle.

Some rain forest trees are destroyed by burning, while some others are cut down and left on the forest floor.

- d. Describe one way that burning rain forest trees affects nutrient cycling differently than cutting them down and leaving them on the forest floor.

**High School Biology**  
**Spring 2012 Released Items:**  
**Reporting Categories, Standards, and Correct Answers\***

Item No.	Page No.	Reporting Category	Standard	Correct Answer (MC)*
1	306	<i>Biochemistry and Cell Biology</i>	2.4	A
2	306	<i>Anatomy and Physiology</i>	4.2	B
3	307	<i>Evolution and Biodiversity</i>	5.3	A
4	307	<i>Evolution and Biodiversity</i>	5.2	C
5	308	<i>Ecology</i>	6.3	A
6	308	<i>Anatomy and Physiology</i>	4.8	B
7	309	<i>Genetics</i>	3.3	D
8	311	<i>Biochemistry and Cell Biology</i>	2.3	A
9	311	<i>Biochemistry and Cell Biology</i>	2.5	B
10	311	<i>Biochemistry and Cell Biology</i>	2.1	A
11	311	<i>Biochemistry and Cell Biology</i>	1.2	C
12	312	<i>Evolution and Biodiversity</i>	5.3	
13	313	<i>Evolution and Biodiversity</i>	5.2	B
14	313	<i>Ecology</i>	6.2	A
15	313	<i>Biochemistry and Cell Biology</i>	2.2	D
16	314	<i>Genetics</i>	3.1	B
17	314	<i>Genetics</i>	3.4	A
18	315	<i>Ecology</i>	6.1	D
19	316	<i>Genetics</i>	3.6	C
20	316	<i>Genetics</i>	3.2	D
21	316	<i>Biochemistry and Cell Biology</i>	1.1	C
22	316	<i>Anatomy and Physiology</i>	4.3	B
23	317	<i>Biochemistry and Cell Biology</i>	2.3	
24	318	<i>Evolution and Biodiversity</i>	5.1	C
25	318	<i>Anatomy and Physiology</i>	4.1	B
26	318	<i>Biochemistry and Cell Biology</i>	2.6	B
27	318	<i>Genetics</i>	3.1	B
28	319	<i>Ecology</i>	6.2	A
29	319	<i>Ecology</i>	6.3	D
30	319	<i>Biochemistry and Cell Biology</i>	1.2	A
31	319	<i>Evolution and Biodiversity</i>	5.3	D
32	320	<i>Anatomy and Physiology</i>	4.5	
33	321	<i>Ecology</i>	6.1	A
34	321	<i>Evolution and Biodiversity</i>	5.1	A
35	322	<i>Ecology</i>	6.3	C
36	322	<i>Biochemistry and Cell Biology</i>	1.3	A
37	323	<i>Genetics</i>	3.6	D
38	323	<i>Anatomy and Physiology</i>	4.7	C
39	323	<i>Evolution and Biodiversity</i>	5.2	C

Item No.	Page No.	Reporting Category	Standard	Correct Answer (MC)*
40	323	<i>Ecology</i>	6.4	A
41	324	<i>Genetics</i>	3.4	D
42	324	<i>Evolution and Biodiversity</i>	5.1	B
43	324	<i>Biochemistry and Cell Biology</i>	1.3	A
44	325	<i>Genetics</i>	3.5	
45	326	<i>Ecology</i>	6.4	

\* Answers are provided here for multiple-choice items and short-answer items only. Sample responses and scoring guidelines for open-response items, which are indicated by shaded cells, will be posted to the Department's website later this year.

## APPENDIX B

### MEPA PERFORMANCE-LEVEL DESCRIPTORS

Source: Guide to interpreting the spring 2012 MEPA Reports for schools and districts.

#### Grade Spans 3–4, 5–6, 7–8, and 9–12

A student at **Level 1** cannot yet communicate in English, and errors almost always interfere with communication. Comprehension is demonstrated either without words, through a few basic words, or in a language other than English. A student performing at this level typically

- reads only a few simple written words or phrases, with help;
- writes only a few simple words and a few short sentences with errors;
- speaks using only a few English words with common errors, and is not easily understood;
- understands only a little spoken English.

A student at **Level 2** communicates using simple written and spoken English at school, with errors that often interfere with communication and understanding. A student performing at this level typically

- reads and understands simple words, phrases, and a few simple sentences with help, but shows little awareness of features of written English;
- writes one or more simple sentences with some understanding of purpose and audience, but shows little control of grade-level standard English writing conventions;
- speaks using basic English words and phrases, and is generally difficult to understand;
- understands some basic spoken vocabulary, phrases, and simple questions, with frequent repetition and explanation.

A student at **Level 3** communicates using basic English at school, although errors sometimes interfere with communication and understanding. A student performing at this level typically

- reads and understands many common words and some grade-level academic vocabulary; can understand the main idea of some grade-level texts, and understands some grade-level features of written English;
- writes and edits simple sentences and paragraphs to suit an audience, and uses basic grade-level vocabulary; shows some control of grade-level standard English writing conventions;
- speaks using many basic English words and some grade-level academic

vocabulary, creating original sentences, with some errors and pauses in conversation;

- understands most spoken English sentences and questions, some basic grade-level academic vocabulary, and grade-level texts read aloud, with some repetition and explanation.

A student at **Level 4** is generally fluent in English at school, and oral and written communication is mostly correct and usually understandable, with few or minor errors. A student performing at this level typically

- reads and understands most grade-level texts, including academic vocabulary and most grade-level features of written English;
- writes and edits short texts with few errors using basic grade-level academic vocabulary, and shows basic control of grade-level standard English writing conventions;
- speaks English with basic fluency, using grade-level words and sentences, with occasional errors;
- understands most spoken English during classroom discussions, with only occasional repetition and explanation.

A student at **Level 5** communicates effectively in English across all academic subjects, with few errors. The student shows control of standard English. Oral and written communication is correct and understandable. A student performing at this level typically

- reads and understands most grade-level texts, including a range of academic vocabulary;
- writes and edits texts of different lengths, giving details and descriptions to suit the purpose and audience, and shows a general control of standard grade-level English writing conventions;
- speaks English with grade-level fluency, using academic language and descriptive vocabulary in conversations and classroom discussions;
- understands spoken English during nearly all conversations and classroom discussions.

## APPENDIX C

### MASSACHUSETTS HIGH SCHOOL BIOLOGY STANDARDS

## **Biology, High School**

### Learning Standards for a Full First-Year Course

#### **I. CONTENT STANDARDS**

##### **1. The Chemistry of Life**

*Central Concept:* Chemical elements form organic molecules that interact to perform the basic functions of life.

- 1.1 Recognize that biological organisms are composed primarily of very few elements. The six most common are C, H, N, O, P, and S.
- 1.2 Describe the basic molecular structures and primary functions of the four major categories of organic molecules (carbohydrates, lipids, proteins, nucleic acids).
- 1.3 Explain the role of enzymes as catalysts that lower the activation energy of biochemical reactions. Identify factors, such as pH and temperature, that have an effect on enzymes.

##### **2. Cell Biology**

*Central Concepts:* Cells have specific structures and functions that make them distinctive. Processes in a cell can be classified broadly as growth, maintenance, and reproduction.

- 2.1 Relate cell parts/organelles (plasma membrane, nuclear envelope, nucleus, nucleolus, cytoplasm, mitochondrion, endoplasmic reticulum, Golgi apparatus, lysosome, ribosome, vacuole, cell wall, chloroplast, cytoskeleton, centriole, cilium, flagellum, pseudopod) to their functions. Explain the role of cell membranes as a highly selective barrier (diffusion, osmosis, facilitated diffusion, active transport).
- 2.2 Compare and contrast, at the cellular level, the general structures and degrees of complexity of prokaryotes and eukaryotes.
- 2.3 Use cellular evidence (e.g., cell structure, cell number, cell reproduction) and modes of nutrition to describe the six kingdoms (Archaea, Bacteria, Eubacteria, Protista, Fungi, Plantae, Animalia).
- 2.4 Identify the reactants, products, and basic purposes of photosynthesis and cellular respiration. Explain the interrelated nature of photosynthesis and cellular respiration in the cells of photosynthetic organisms.
- 2.5 Explain the important role that ATP serves in metabolism.
- 2.6 Describe the cell cycle and the process of mitosis. Explain the role of mitosis in the formation of new cells, and its importance in maintaining chromosome number during asexual reproduction.
- 2.7 Describe how the process of meiosis results in the formation of haploid

---

## Biology, High School

### Learning Standards for a Full First-Year Course

---

cells. Explain the importance of this process in sexual reproduction, and how gametes form diploid zygotes in the process of fertilization.

- 2.8 Compare and contrast a virus and a cell in terms of genetic material and reproduction.

### 3. Genetics

*Central Concepts:* Genes allow for the storage and transmission of genetic information. They are a set of instructions encoded in the nucleotide sequence of each organism. Genes code for the specific sequences of amino acids that comprise the proteins characteristic to that organism.

- 3.1 Describe the basic structure (double helix, sugar/phosphate backbone, linked by complementary nucleotide pairs) of DNA, and describe its function in genetic inheritance.
- 3.2 Describe the basic process of DNA replication and how it relates to the transmission and conservation of the genetic code. Explain the basic processes of transcription and translation, and how they result in the expression of genes. Distinguish among the end products of replication, transcription, and translation.
- 3.3 Explain how mutations in the DNA sequence of a gene may or may not result in phenotypic change in an organism. Explain how mutations in gametes may result in phenotypic changes in offspring.
- 3.4 Distinguish among observed inheritance patterns caused by several types of genetic traits (dominant, recessive, codominant, sex-linked, polygenic, incomplete dominance, multiple alleles).
- 3.5 Describe how Mendel's laws of segregation and independent assortment can be observed through patterns of inheritance (e.g., dihybrid crosses).
- 3.6 Use a Punnett Square to determine the probabilities for genotype and phenotype combinations in monohybrid crosses.

### 4. Anatomy and Physiology

*Central Concepts:* There is a relationship between the organization of cells into tissues and the organization of tissues into organs. The structures and functions of organs determine their relationships within body systems of an organism. Homeostasis allows the body to perform its normal functions.

- 4.1 Explain generally how the digestive system (mouth, pharynx, esophagus, stomach, small and large intestines, rectum) converts macromolecules from food into smaller molecules that can be used by cells for energy and for repair and growth.
- 4.2 Explain how the circulatory system (heart, arteries, veins, capillaries, red blood cells) transports nutrients and oxygen to cells and removes cell

---

## Biology, High School

### Learning Standards for a Full First-Year Course

---

wastes. Describe how the kidneys and the liver are closely associated with the circulatory system as they perform the excretory function of removing waste from the blood. Recognize that kidneys remove nitrogenous wastes, and the liver removes many toxic compounds from blood.

- 4.3 Explain how the respiratory system (nose, pharynx, larynx, trachea, lungs, alveoli) provides exchange of oxygen and carbon dioxide.
- 4.4 Explain how the nervous system (brain, spinal cord, sensory neurons, motor neurons) mediates communication among different parts of the body and mediates the body's interactions with the environment. Identify the basic unit of the nervous system, the neuron, and explain generally how it works.
- 4.5 Explain how the muscular/skeletal system (skeletal, smooth and cardiac muscles, bones, cartilage, ligaments, tendons) works with other systems to support the body and allow for movement. Recognize that bones produce blood cells.
- 4.6 Recognize that the sexual reproductive system allows organisms to produce offspring that receive half of their genetic information from their mother and half from their father, and that sexually produced offspring resemble, but are not identical to, either of their parents.
- 4.7 Recognize that communication among cells is required for coordination of body functions. The nerves communicate with electrochemical signals, hormones circulate through the blood, and some cells produce signals to communicate only with nearby cells.
- 4.8 Recognize that the body's systems interact to maintain homeostasis. Describe the basic function of a physiological feedback loop.

#### **5. Evolution and Biodiversity**

*Central Concepts:* Evolution is the result of genetic changes that occur in constantly changing environments. Over many generations, changes in the genetic make-up of populations may affect biodiversity through speciation and extinction.

- 5.1 Explain how evolution is demonstrated by evidence from the fossil record, comparative anatomy, genetics, molecular biology, and examples of natural selection.
- 5.2 Describe species as reproductively distinct groups of organisms. Recognize that species are further classified into a hierarchical taxonomic system (kingdom, phylum, class, order, family, genus, species) based on morphological, behavioral, and molecular similarities. Describe the role that geographic isolation can play in speciation.
- 5.3 Explain how evolution through natural selection can result in changes in biodiversity through the increase or decrease of genetic diversity within a population.



---

## Biology, High School

### Learning Standards for a Full First-Year Course

---

#### **6. Ecology**

*Central Concept:* Ecology is the interaction among organisms and between organisms and their environment.

- 6.1 Explain how birth, death, immigration, and emigration influence population size.
- 6.2 Analyze changes in population size and biodiversity (speciation and extinction) that result from the following: natural causes, changes in climate, human activity, and the introduction of invasive, non-native species.
- 6.3 Use a food web to identify and distinguish producers, consumers, and decomposers, and explain the transfer of energy through trophic levels. Describe how relationships among organisms (predation, parasitism, competition, commensalism, mutualism) add to the complexity of biological communities.
- 6.4 Explain how water, carbon, and nitrogen cycle between abiotic resources and organic matter in an ecosystem, and how oxygen cycles through photosynthesis and respiration.

## APPENDIX D

### COGNITIVE AND QUANTITATIVE SKILLS DESCRIPTIONS FOR SCIENCE AND TECHNOLOGY/ENGINEERING MCAS TESTS

Only one cognitive skill will be designated for a common item, although several different cognitive skills may apply to a single item. In addition to the identified cognitive skill, an item may also be identified as having a quantitative component.

<div style="display: flex; flex-direction: column; align-items: center;"> <div>Basic Skills</div> <div style="margin: 20px 0;">↓</div> <div>More Demanding Skills</div> <div>-----</div> <div>Other Skills</div> </div>	Cognitive Skill	Description
	Foundational	<ul style="list-style-type: none"> <li>-Declarative knowledge</li> <li>-Recall of facts</li> <li>-Definition/vocabulary</li> </ul>
	Conceptual	<ul style="list-style-type: none"> <li>-Recognition of a concept</li> <li>-Description of a principle</li> <li>-Description of a process</li> </ul>
	Application	<ul style="list-style-type: none"> <li>-Procedural knowledge</li> <li>-Application of conceptual knowledge to a novel situation</li> <li>-Use predetermined models to devise a solution</li> <li>-Classification diverse objects into unifying groups</li> </ul> <p>*Note: This cognitive level does not automatically include all practical contexts for a concept; the application/situation for the concept must be a new, different example for the concept, not the example used in most textbooks.</p>
	Constructive/ Synthetic	<ul style="list-style-type: none"> <li>-Synthesis of a novel response (by pulling several different pieces of knowledge together)</li> <li>-Application of multi-step problem solving</li> <li>-Application of experimental design and critique</li> <li>-Formulation of a hypothesis</li> <li>-Application of predictive reasoning</li> <li>-Interpretation of experimental data analysis</li> <li>-Application of scientific inquiry or engineering design process</li> </ul>
	Quantitative	<ul style="list-style-type: none"> <li>-Analysis of data</li> <li>-Computation of numerical solution</li> <li>-Graphical interpretation and interpretation of data in tables</li> <li>-Predictive calculations</li> </ul>

Source: Massachusetts Department of Elementary and Secondary Education. (n.d.). *Cognitive and quantitative skills descriptions for science and technology/engineering MCAS tests*. Malden, MA.

## APPENDIX E

### LEXILE ANALYZER® RESULTS

June 2012 Biology MCAS, question 18 item stem.

Lexile Analyzer: Results - Windows Internet Explorer

https://www.lexile.com/analyzer/results/1653368/

Lexile Analyzer: Results

File Edit View Favorites Tools Help

Google Search Share Translate More Sign In

ETS Researcher Maria Marti... Language, Power, and Ped... Belmont\_MA Message Scien... ALLIES Resources

Page Safety Tools

**LEXILE**® The Lexile® Framework for Reading

Quick Book Search:  
Title, Author, or ISBN [Advanced Search](#)  
Put an exact title or author in quotes (ex: "new moon")

About Lexile Measures Using Lexile Measures Common Core Lexile Tools Lexile Training

Lexile® Measure  
**1060L**

Mean Sentence Length  
**16.50**

Mean Log Word Frequency  
**3.48**

Word Count  
**66**

Lexile Analyzer

**Lexile Analyzer: Results**

These results are not saved in any retrievable way. You should print this screen and note your filename or the title of your sample text. If you do not print or record the results, you will have to re-analyze your sample text to know its Lexile measure.

Submit another file

Usage history  
Review your [usage of the Lexile Analyzer®](#)

10:09 AM 2/11/2013

## APPENDIX F

### FIRST LANGUAGE CHARACTERISTICS: LANGUAGE FAMILY AND ORTHOGRAPHY

Language	ELLs (n)	Latinate Yes/No	Alphabetic Yes/No
Afridaans	1	No	Yes
Albanian	18	No	Yes
Amharic	12	No	No
Arabic	107	No	No
Armenian	3	No	No
Bantu	1	No	No
Basque	1	No	Yes
Bengali	7	No	No
Bulgarian	2	No	Yes
Burmese	24	No	No
Canton	39	No	No
Cape Verdean	227	Yes	Yes
Chichewa	1	No	Yes
Chinese, Not Mandarin or Cantonese	103	No	No
Crioulo	18	No	Yes
Danish	2	No	Yes
Farsi	3	No	No
French	44	Yes	Yes
French Patois	5	Yes	Yes
German	1	No	Yes
Greek	3	No	Yes
Guarani	14	No	Yes
Haitian Creole	285	Yes	Yes
Hakka Dialect	1	No	No
Hebrew	2	No	No
Hindi	4	No	No
Hmong	1	No	Yes
Ibo	1	No	Yes
Italian	1	Yes	Yes
Jamaican Creole	6	No	Yes
Japanese	5	No	No

Language	ELLs (n)	Latinate Yes/No	Alphabetic Yes/No
Khaikha Mongolian	1	No	Yes
Khmer/Khmai	78	No	No
Kinyarwandu	3	No	Yes
<i>Kirundi</i>	7	<i>No</i>	<i>Yes</i>
Korean	9	No	No
Kpelle	2	No	Yes
Krio	8	No	Yes
Kurdish	1	No	No
Lao	4	No	No
Luganda	3	No	Yes
Mandarin	60	No	No
Mende	1	No	Yes
Nepali	52	No	No
Niger-Congo	6	No	Yes
Patois	18	No	Yes
Pilipino	5	No	Yes
Polish	3	No	Yes
Portuguese	160	Yes	Yes
Punjabi/Panjabi	3	No	No
Pushtu/Pashtu	7	No	No
Quechua	3	No	Yes
Romanian	1	Yes	Yes
Russian	10	No	Yes
Serbo-Croatian	2	No	Yes
Somali	29	No	Yes
Spanish	1697	Yes	Yes
Swahili	17	No	Yes
Tagalog	5	No	Yes
Telegu	1	No	No
Thai	6	No	No
Tibetan	6	No	No
Tigre	1	No	No
Tigrinya	2	No	No
Turkish	5	No	Yes
Twi	20	No	Yes
Ukranian	3	No	Yes
Urdu	15	No	No

Language	ELLs (n)	Latinate Yes/No	Alphabetic Yes/No
Vietnamese	82	No	Yes
Yoruba	2	No	Yes
Other	35	No	No
	3315		

#### Latinate and Non-Latinate Languages

Latinate Languages	ELLs (n)
Cape Verdean	227
French	44
French Patois	5
Haitian Creole	285
Italian	1
Portuguese	160
Romanian	1
Spanish	1697
	2420

Non-Latinate Languages	ELLs (n)
Afrikaans	1
Albanian	18
Amharic	12
Arabic	107
Armenian	3
Bantu	1
Basque	1
Bengali	7
Bulgarian	2
Burmese	24
Canton	39
Chichewa	1
Chinese, Not Mandarin or Cantonese	103
Crioulo	18
Danish	2
Farsi	3
	344

Non-Latinate Languages	ELLs (n)
German	1
Greek	3
Guarani	14
Hakka Dialect	1
Hebrew	2
Hindi	4
Hmong	1
Ibo	1
Jamaican Creole	6
Japanese	5
Khaikha Mongolian	1
Khmer/Khmai	78
Kinyarwandu	3
Kirundi	7
Korean	9
Kpelle	2
Krio	8
Kurdish	1
Lao	4
Luganda	3
Mandarin	60
Mende	1
Nepali	52
Niger-Congo	6
Patois	18
Pilipino	5
Polish	3
Punjabi/Panjabi	3
Pushtu/Pashtu	7
Quechua	3
Russian	10
Serbo-Croatian	2
Somali	29
Swahili	17
Tagalog	5
Telegu	1
Thai	6
Tibetan	6

Non-Latinate Languages	ELLs (n)
Tigre	1
Tigrinya	2
Turkish	5
Twi	20
Ukranian	3
Urdu	15
Vietnamese	82
Yoruba	2
Other	35
	895

#### Alphabetic and Non-Alphabetic Languages

Alphabetic Languages	ELLs (n)
Afridaans	1
Albanian	18
Basque	1
Bulgarian	2
Cape Verdean	227
Chichewa	1
Crioulo	18
Danish	2
French	44
French Patois	5
German	1
Greek	3
Guarani	14
Haitian Creole	285
Hmong	1
Ibo	1
Italian	1
Jamaican Creole	6
Khaikha Mongolian	1
Kinyarwandu	3
Kirundi	7
Kpelle	2
Krio	8
	346



Alphabetic Languages	ELLs (n)
Luganda	3
Mende	1
Niger-Congo	6
Patois	18
Pilipino	5
Polish	3
Portuguese	160
Quechua	3
Romanian	1
Russian	10
Serbo-Croatian	2
Somali	29
Spanish	1697
Swahili	17
Tagalog	5
Turkish	5
Twi	20
Ukranian	3
Vietnamese	82
Yoruba	2
	2724

Non-Alphabetic Languages	ELLs (n)
Amharic	12
Arabic	107
Armenian	3
Bantu	1
Bengali	7
Burmese	24
Canton	39
Chinese, Not Mandarin or Cantonese	103
Farsi	3
Hakka Dialect	1
Hebrew	2
Hindi	4
Japanese	5
Khmer/Khmai	78

Non-Alphabetic Languages	ELLs (n)
Korean	9
Kurdish	1
Lao	4
Mandarin	60
Nepali	52
Punjabi/Panjabi	3
Pushtu/Pashtu	7
Telegu	1
Thai	6
Tibetan	6
Tigre	1
Tigrinya	2
Urdu	15
Other	35
	591

#### Non- Latinate and Alphabetic Languages

Alphabetic Languages	ELLs (n)
Afridaans	1
Albanian	18
Basque	1
Bulgarian	2
Chichewa	1
Crioulo	18
Danish	2
German	1
Greek	3
Guarani	14
Hmong	1
Ibo	1
Jamaican Creole	6
Khaikha Mongolian	1
Kinyarwandu	3
Kirundi	7
Kpelle	2
Krio	8

Alphabetic Languages	ELLs (n)
Luganda	3
Mende	1
Niger-Congo	6
Patois	18
Pilipino	5
Polish	3
Quechua	3
Russian	10
Serbo-Croatian	2
Somali	29
Swahili	17
Tagalog	5
Turkish	5
Twi	20
Ukranian	3
Vietnamese	82
Yoruba	2
	<hr/> 304 <hr/>

## APPENDIX G

### TEXTUAL ANALYSES OF JUNE 2012 BIOLOGY MCAS MULTIPLE-CHOICE

#### ITEMS

Item	Domain	Cognitive Skill Level	Linguistic Complexity Analyses						
			Stem Lexical Density (SLD)	Total Answer Lexical Density (TALD)	Answer Lexical Density (ALD)	Total Lexical Density (TLD)	Stem Syntax (SS)	Stem Syn-tactic Density (SSD)	Reading Complexity Score (RCS)
1	CB	Conceptual	17	16	4.0	33	1	17.0	1220
2	AP	Application	19	31	7.8	50	2	9.5	670
3	EV	Application	53	87	21.8	140	3	17.7	1240
4	EV	Application	33	44	11.0	77	3	11.0	730
5	EC	Application	38	15	3.8	53	3	12.7	620
6	AP	Application	47	12	3.0	59	6	7.8	730
7	GE	Application	85	12	3.0	97	15	5.7	1260
8	CB	Foundational	35	6	1.5	41	2	17.5	1150
9	CB	Conceptual	22	18	4.5	40	2	11.0	890
10	CB	Application	40	6	1.5	46	2	20.0	1190
11	BC	Conceptual	23	6	1.5	29	2	11.5	960
13	EV	Conceptual	31	10	2.5	41	2	15.5	1020
14	EC	Foundational	13	8	2.0	21	1	13.0	960
15	CB	Conceptual	35	12	3.0	47	4	8.8	700
16	GE	Foundational	9	26	6.5	35	1	9.0	690
17	GE	Conceptual	19	36	9.0	55	2	9.5	710
18	EC	Application	103	38	9.5	141	10	10.3	1520
19	GE	Application	32	12	3.0	44	2	16.0	1140
20	GE	Application	22	6	1.5	28	2	11.0	710
21	BC	Application	65	8	2.0	73	9	7.2	1180
22	AP	Foundational	24	16	4.0	40	1	24.0	1350

Item	Domain	Cognitive Skill Level	Linguistic Complexity Analyses						
			Stem Lexical Density (SLD)	Total Answer Lexical Density (TALD)	Answer Lexical Density (ALD)	Total Lexical Density (TLD)	Stem Syntax (SS)	Stem Syn-tactic Density (SSD)	Reading Complexity Score (RCS)
24	EV	Conceptual	48	68	17.0	116	3	16.0	1110
25	AP	Foundational	19	6	1.5	25	1	19.0	1220
26	CB	Conceptual	22	4	1.0	26	2	11.0	910
27	GE	Foundational	23	4	1.0	27	4	5.8	550
28	EC	Conceptual	21	39	9.8	60	2	10.5	790
29	EC	Conceptual	27	16	4.0	43	3	9.0	810
30	BC	Application	29	4	1.0	33	2	14.5	910
31	EV	Application	11	60	15.0	71	1	11.0	620
33	EC	Conceptual	39	36	9.0	75	2	19.5	1180
34	EV	Conceptual	38	29	7.3	67	3	12.7	1310
35	EC	Conceptual	41	4	1.0	45	3	13.7	970
36	BC	Application	35	40	10.0	75	3	11.7	910
37	GE	Application	49	4	1.0	53	4	12.3	850
38	AP	Conceptual	20	42	10.5	62	2	10.0	540
39	EV	Conceptual	19	42	10.5	61	1	19.0	1200
40	EC	Conceptual	30	4	1.0	34	4	7.5	750
41	GE	Application	44	33	8.3	77	7	6.3	1040
42	EV	Conceptual	18	36	9.0	54	2	9.0	630
43	BC	Conceptual	27	4	1.0	31	2	13.5	900

# APPENDIX H

## ITEM COMPOSITE LINGUISTIC COMPLEXITY

Item	TLD <sub>N</sub>	SSD <sub>N</sub>	RCS <sub>N</sub>	CLC Value	CLC Category
1	.10	.62	.69	1.41	High CLC
2	.24	.21	.13	.58	Low CLC
3	.99	.65	.71	2.36	High CLC
4	.47	.29	.19	.95	Medium CLC
5	.27	.38	.08	.73	Low CLC
6	.32	.12	.19	.63	Low CLC
7	.64	.00	.73	1.38	High CLC
8	.17	.64	.62	1.43	High CLC
9	.16	.29	.36	.80	Medium CLC
10	.21	.78	.66	1.65	High CLC
11	.07	.32	.43	.81	Medium CLC
13	.17	.53	.49	1.19	Medium CLC
14	.00	.40	.43	.83	Medium CLC
15	.22	.17	.16	.55	Low CLC
16	.12	.18	.15	.45	Low CLC
17	.28	.21	.17	.66	Low CLC
18	1.00	.25	1.00	2.25	High CLC
19	.19	.56	.61	1.37	High CLC
20	.06	.29	.17	.52	Low CLC
21	.43	.08	.65	1.17	Medium CLC
22	.16	1.00	.83	1.98	High CLC
25	.03	.73	.69	1.45	High CLC
26	.04	.29	.38	.71	Low CLC
27	.05	.00	.01	.06	Low CLC
28	.33	.26	.26	.84	Medium CLC
29	.18	.18	.28	.64	Low CLC
30	.10	.48	.38	.96	Medium CLC
31	.39	.29	.08	.76	Medium CLC
33	.45	.75	.65	1.86	High CLC
34	.43	.47	.79	1.68	High CLC
35	.20	.43	.44	1.07	Medium CLC

Item	TLD <sub>N</sub>	SSD <sub>N</sub>	RCS <sub>N</sub>	CLC Value	CLC Category
36	.44	.31	.38	1.13	Medium CLC
37	.27	.36	.32	.94	Medium CLC
38	.34	.23	.00	.58	Low CLC
39	.33	.73	.67	1.73	High CLC
40	.11	.10	.21	.42	Low CLC
41	.47	.03	.51	1.01	Medium CLC
42	.28	.18	.09	.55	Low CLC
43	.08	.43	.37	.88	Medium CLC

## APPENDIX I

### MCAS RAW-TO-SCALED SCORE CONVERSION FOR JUNE 2012 BIOLOGY

#### MCAS

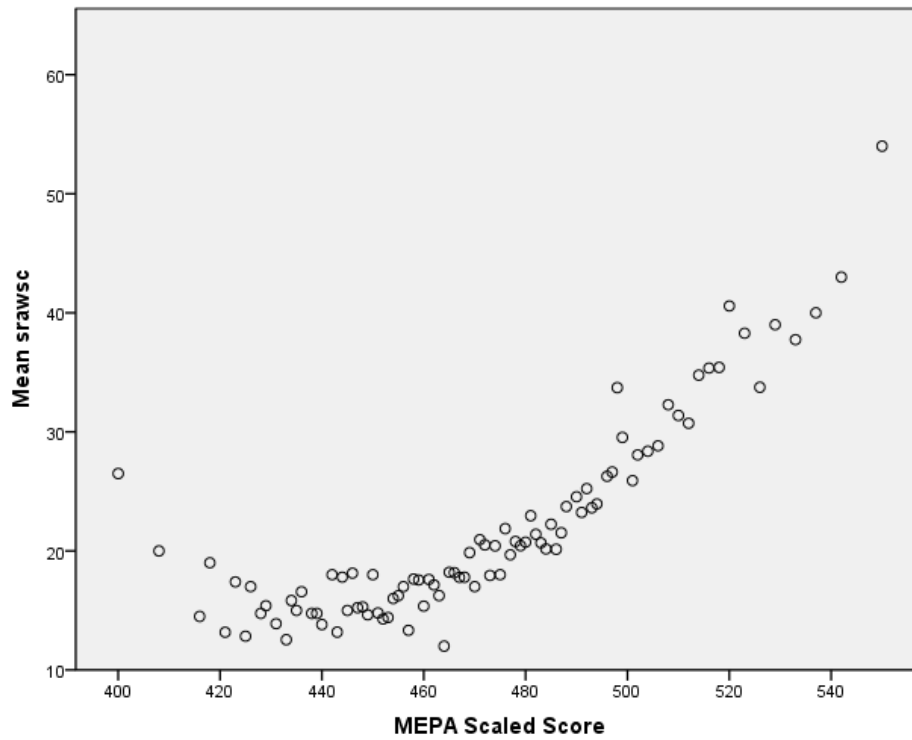
Raw	Scaled	Raw	Scaled	Raw	Scaled	Raw	Scaled
0	200	20	220	33	240	48	260
1	200	21	220	34	242	49	262
2	202	22	220	35	242	50	262
3	202	23	222	36	244	51	264
4	204	24	224	37	244	52	266
5	204	25	226	38	246	53	268
6	206	26	228	39	248	54	268
7	206	27	230	40	248	55	270
8	208	28	232	41	250	56	272
9	208	29	234	42	252	57	276
10	210	30	234	43	252	58	278
11	210	31	236	44	254	59	280
12	212	32	238	45	256	60	280
13	214			46	258		
14	214			47	258		
15	216						
16	216						
17	218						
18	218						
19	218						

Source: *Spring 2012 MCAS raw-to-scaled score conversion: Science and technology/engineering* (MA DESE, 2012h).



## APPENDIX J

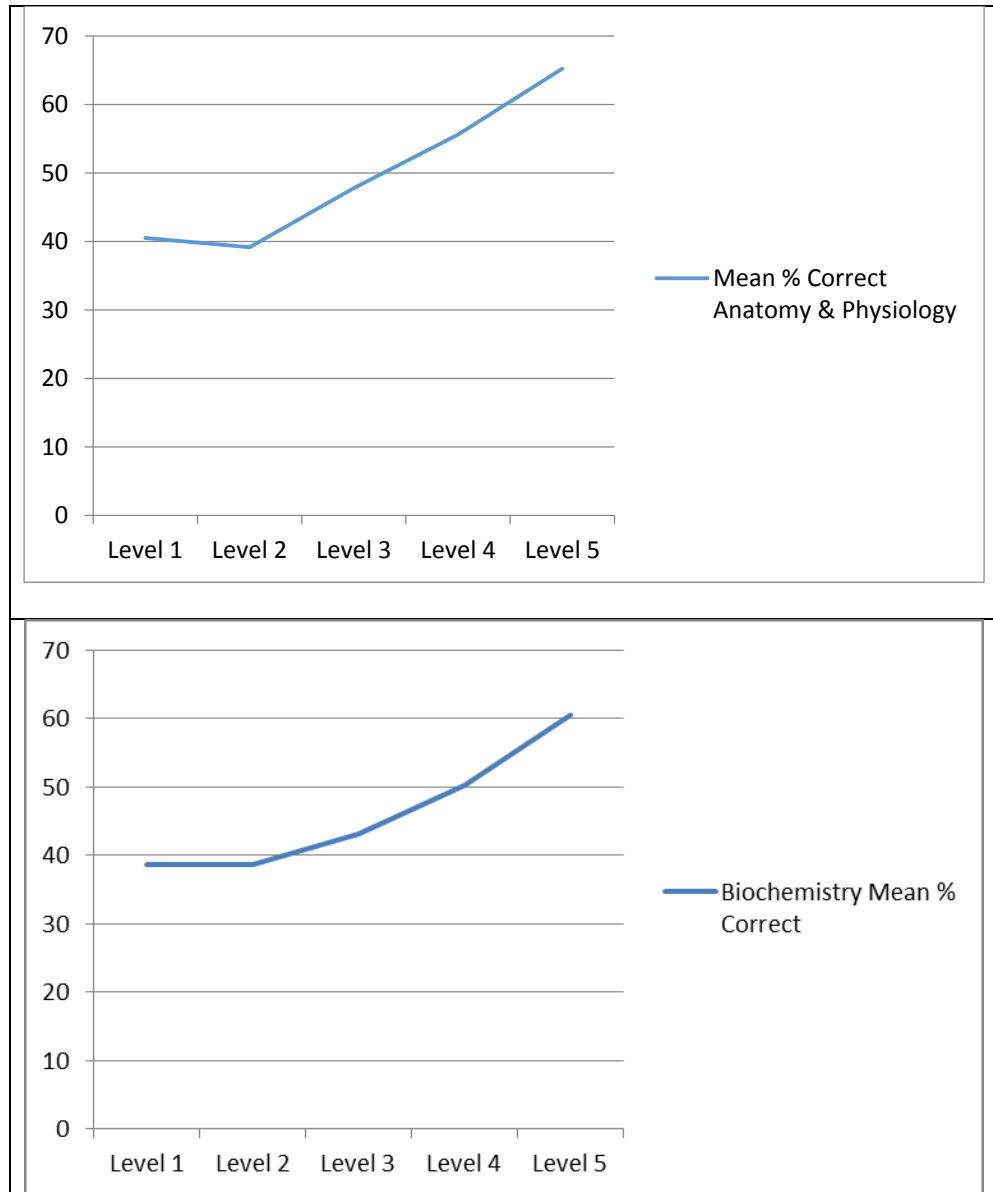
### MCAS PERFORMANCE BY MEPA SCORE

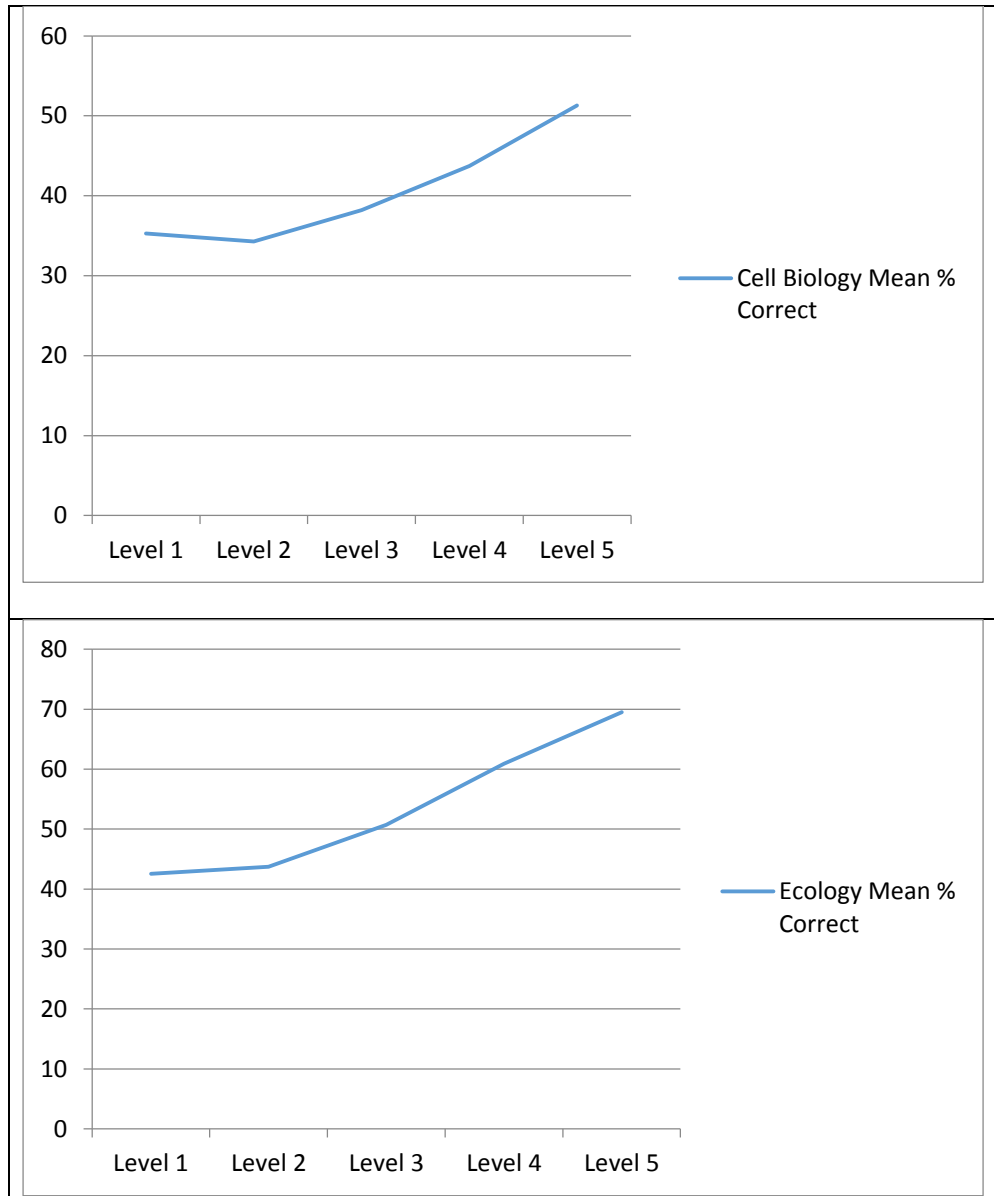


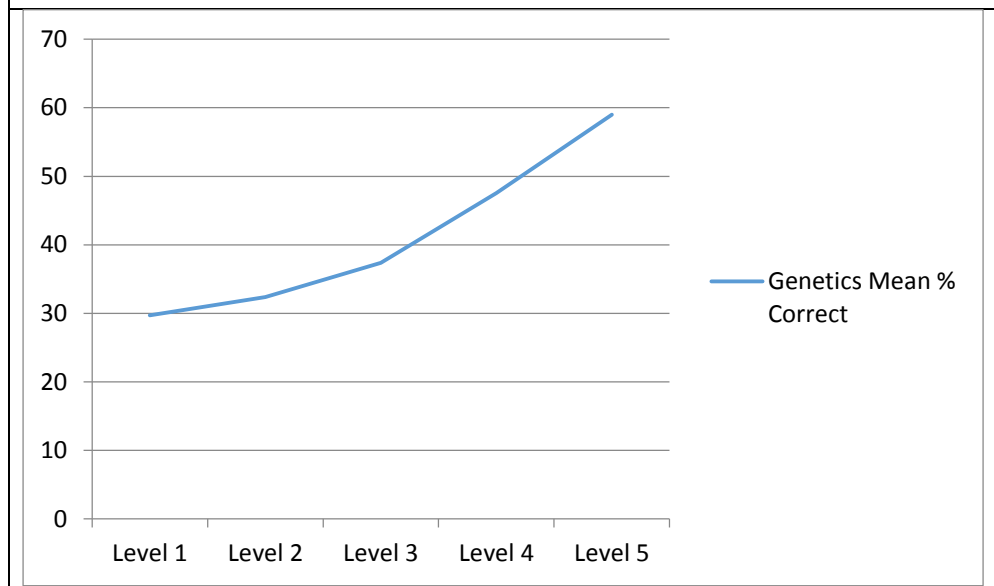
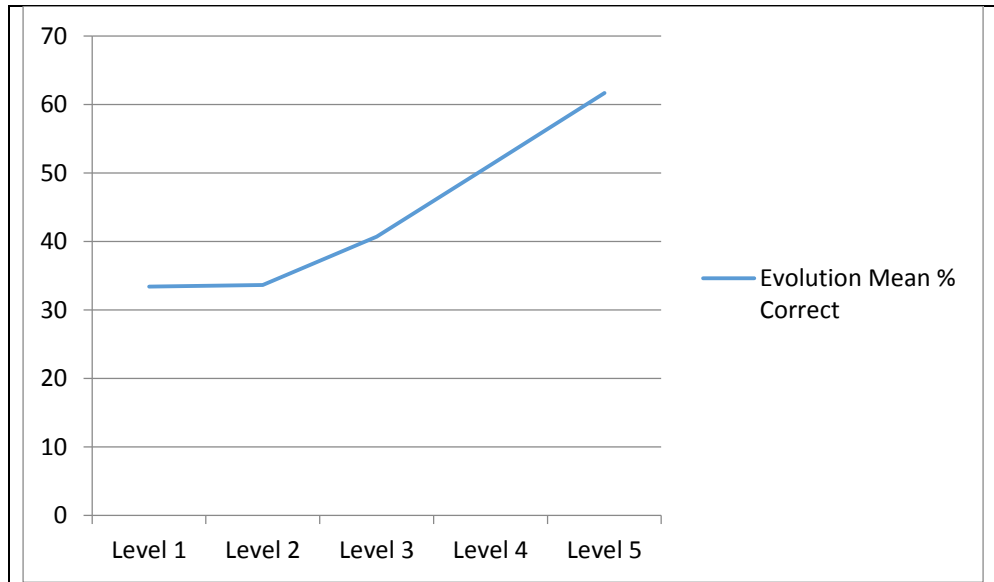
Mean raw Biology MCAS score against MEPA scaled score (English proficiency).  
Linearity emerged around MEPA Level 3, which began at a MEPA scaled score of 464.

## APPENDIX K

### CONTENT DOMAIN PERFORMANCE BY ENGLISH PROFICIENCY







## REFERENCE LIST

- Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment*, 8(3), 231-257.
- Abedi, J. (2008a). Classification system for English language learners: Issues and recommendations. *Educational Measurement: Issues & Practice*, 27(3), 17-31. doi:10.1111/j.1745-3992.2008.00125.x
- Abedi, J. (2008b). Measuring students' level of English proficiency: Educational significance and assessment requirements. *Educational Assessment*, 13(2/3), 193-214. doi: 10.1080/10627190802394404
- Abedi, J. (2009). Computer testing as a form of accommodation for English language learners. *Educational Assessment*, 14(3/4), 195-211. doi:10.1080/10627190903448851
- Abedi, J., & Dietel, R. (2004). Challenges in the No Child Left Behind Act for English language learners. Retrieved from [http://www.cse.ucla.edu/products/policy/cresst\\_policy7.pdf](http://www.cse.ucla.edu/products/policy/cresst_policy7.pdf)
- Abedi, J., & Gándara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues & Practice*, 25(4), 36-46. doi:10.1111/j.1745-3992.2006.00077.x
- Abedi, J., & Hejri, F. (2004). Accommodations for students with limited English proficiency in the National Assessment of Educational Progress. *Applied Measurement in Education*, 17(4), 371-392.
- Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, 74(1), 1-28.
- Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, 14(3), 219-234.
- Academic language. (2011) Retrieved October 14, 2011, from <http://www.academiclanguage.wceruw.org>
- August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English language learners. *Learning Disabilities Research & Practice*, 20(1), 50-57. doi: 10.1111/j.1540-5826.2005.00120.x

- Aukerman, M. (2007). A culpable CALP: Rethinking the conversational/academic language proficiency distinction in early literacy instruction. *Reading Teacher*, 60(7), 626-635. doi: 10.1598/rt.60.7.3
- Baik, C., & Greig, J. (2009). Improving the academic outcomes of undergraduate ESL students: The case for discipline-based academic skills programs. *Higher Education Research & Development*, 28(4), 401-416. doi:10.1080/07294360903067005
- Bailey, A. L., & Huang, B. H. (2011). Do current English language development/proficiency standards reflect the English needed for success in school? *Language Testing*, 28(3), 343-365.
- Bialystok, E., McBride-Chang, C., & Luk, G. (2005). Bilingualism, language proficiency, and learning to read in two writing systems. *Journal of Educational Psychology*, 97(4), 580-590.
- Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, 9, 2-20. doi:10.1016/j.jeap.2010.01.001
- Birch, B. M. (2002). *English L2 reading*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Blume, H. (2012, October 22). Fact check: Massachusetts student achievement is first in nation. *The Los Angeles Times*. Retrieved from <http://articles.latimes.com/2012/oct/22/news/la-pn-fact-check-debate-massachusetts-20121022>
- Boyle, A., Taylor, J., Hurlburt, S., & Soga, K. (2010). *Title III accountability: Behind the numbers*. Washington, DC: U.S. Department of Education. Retrieved from <http://www2.ed.gov/rschstat/eval/title-iii/behind-numbers.pdf>
- Boyson, B. A., & Short, D. J. (2012). Helping newcomer students succeed in secondary schools and beyond. Washington, DC: Center for Applied Linguistics.
- Brown, H. D. (2000). *Principles of language teaching* (4th ed.). White Plains, NY: Addison Wesley Longman, Inc.
- Bunch, G. C. (2006). "Academic English" in the 7th grade: Broadening the lens, expanding access. *Journal of English for Academic Purposes*, 5(4), 284-301. doi:10.1016/j.jeap.2006.08.007
- Campbell, G. L. (1991a). *Compendium of the world's languages. Abaza to Lusatian* (Vol. I). London & New York: Routledge.

- Campbell, G. L. (1991b). *Compendium of the world's languages. Maasai to Zuni* (Vol. II). London & New York: Routledge.
- Carlo, M. S., August, D., McLaughlin, B., Snow, C., Dressler, C., Lippman, D., . . . White, C. E. (2008). Closing the gap: Addressing the vocabulary needs of English-language learners in bilingual and mainstream classrooms. *Journal of Education*, 189(1/2), 57-76.
- Chitiri, H.-F., & Willows, D. M. (1994). Word recognition in two languages and orthographies: English and Greek. *Memory & Cognition*, 22(3), 313-325.
- Collier, V. (1987a). Age and rate of acquisition of second language for academic purposes. *TESOL Quarterly*, 21, 617-641.
- Collier, V. (1987b). *The effect of age on acquisition of a second language for school*. Wheaton, MD: National Clearinghouse for Bilingual Education.
- Conrad, S. M. (1996). Investigating academic texts with corpus-based techniques: An example from biology. *Linguistics and Education*, 8(3), 299-326.
- Cook, H. G., Boals, T., & Lundberg, T. (2011). Academic achievement for English learners: What can we reasonably expect? *Phi Delta Kappan*, 93(3), 66-69.
- Cosentino de Cohen, C., Deterding, N., & Clewell, B. C. (2005). Who's left behind? Immigrant children in high and low LEP schools. Washington DC: Urban Institute.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34, 213-238.
- Cummins, J. (2000). *Language, power, and pedagogy: Bilingual children in the crossfire*. Clevedon, England: Multilingual Matters, Ltd.
- Cummins, J. (2003). BICS and CALP: Origins and rationale for the distinction. In C. Bratt Paulston & G. R. Tucker (Eds.), *Sociolinguistics the essential readings* (pp. 322-328). Malden, MA: Blackwell Publishing.
- Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. Retrieved October 6, 2011, from [http://www.wisd.us/campus/whs/social\\_studies/edd/Fall09/8344/Articles/CumminsBICSCALPSpringer2007.pdf](http://www.wisd.us/campus/whs/social_studies/edd/Fall09/8344/Articles/CumminsBICSCALPSpringer2007.pdf)
- Ćurković, N. (2012). Using of structural equation modeling techniques in cognitive levels validation. *Interdisciplinary Description of Complex Systems*, 10(3), 270-283. doi:10.7906/indecs.10.3.5

- Dalton, T. A. (2011). Comparison of two approaches to improving cognitive academic language proficiency for school-aged English language learners: Two-group, pretest/posttest design. *All Graduate Plan B and other Reports* (Paper 41). Retrieved from <http://digitalcommons.usu.edu/gradreports/41>
- DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology, 100*(1), 223-234.
- Deng, N., Sukin, T., & Hambleton, R. K. (2009). *Judging the content and statistical equivalence of MCAS operational and linking items*. MCAS Validity Report 20. Amherst, MA: University of Massachusetts Amherst, Center for Educational Assessment.
- Dickey, R. J. (2004). Content (adj) or content (n) with your English classes. *Education International, 1*(3), 10-15.
- Dodonova, Y. A., & Dodonov, Y. S. (2013). Faster on easy items, more accurate on difficult ones: Cognitive ability and performance on a task of varying difficulty. *Intelligence, 41*(1), 1-10. doi:10.1016/j.intell.2012.10.003
- Duran, R. P. (2008). Assessing English-language learners' achievement. *Review of Research in Education, 32*(1), 292-327.
- Echevarria, J., Vogt, M., & Short, D. J. (2013). *Making content comprehensible for English learners* (4th ed.). Boston, MA: Pearson.
- Education Reform Act, Massachusetts General Laws c.71 (1993).
- Ellis, R. (2003). *Second language acquisition*. Oxford, England: Oxford University Press.
- English Language Education in Public Schools, Massachusetts General Laws c. 71A § 2 (2002).
- Finegan, E. (2004). *Language: Its structure and use* (4th ed.). Boston, MA: Wadsworth Thomson.
- Francis, D. J., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical guidelines for the education of English language learners: Research-based recommendations for instruction and academic interventions*. Portsmouth, NH: RMC Research Corporation, Center on Instruction.
- Freeman, D., & Freeman, Y. (1988). Sheltered English instruction. Washington, DC: Eric Clearinghouse on Languages and Linguistics. Retrieved <http://www.eric.ed.gov/PDFS/ED301070.pdf>



- Gee, J. (1999). *An introduction to discourse analysis theory and method*. London, England: Routledge.
- Gierl, M. J., Alves, C., & Majeau, R. T. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment. *International Journal of Testing*, 10(4), 318-341. doi:10.1080/15305058.2010.509554
- Hakuta, K., Butler, Y. G., & Witt, D. (2000). How long does it take English learners to attain proficiency? Santa Barbara: University of California Linguistic Minority Research Institute.
- Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2), 101-123.
- Hambleton, R., Zhao, Y., Smith, Z., Lam, W., & Deng, N. (2008). *Psychometric analyses of the 2006 MCAS high school science and technology/engineering tests. MCAS Validity Report 17*. Amherst, MA: University of Massachusetts Amherst, Center for Educational Assessment.
- Hersi, A. A. (2012). Transnational immigration and education: A case study of an Ethiopian immigrant high school student. *Creative Education*, 3(1), 149-154.
- Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461-473. doi:10.1093/applin/amp048
- Hymes, D. (2003). The interaction of language and social life. In C. Bratt Paulston & G. R. Tucker (Eds.), *Sociolinguistics: The essential readings*. Malden, MA: Blackwell Publishing.
- Isabelli, C. (n.d.). Chapter 2. Retrieved September 29, 2012, from <http://www.unr.edu/cla/fll/people/facultyPages/isabelli/FLL703/Chapter%202.pdf>
- Jirka, S. J., & Hambleton, R. K. (2005). Cognitive complexity levels for the MCAS assessments. *MCAS Validity Report 10*. Amherst, MA: University of Massachusetts Amherst, Center for Educational Assessment.
- Johnstone, B. (2002). *Discourse analysis*. Malden, MA: Blackwell Publishing.
- Kim, J. Y. (2008). *The effects of English language learning (ELL) factors on the White-Asian academic achievement gap: An analysis of National Assessment of Educational Progress (NAEP) in math and reading* (Master's thesis). The State University of New York at Buffalo, Buffalo, NY.

- Kong, S., & Hoare, P. (2011). Cognitive content engagement in content-based language teaching. *Language Teaching Research*, 15(3), 307-324.
- Kucer, S. B. (2005). *Dimensions of literacy: A conceptual base for teaching reading and writing in school settings* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Kuiken, F., & Vedder, I. (2012). Complexity and second language learning (CoSeLL). *International Journal of Applied Linguistics*, 22(2), 277-279. doi:10.1111/j.1473-4192.2012.00320.x
- Larsen-Freeman, D. (2000). *Techniques and principles in language teaching* (2nd ed.). New York, NY: Oxford University Press.
- Lawrence, J. F., White, C., & Snow, C. E. (2010). The words students need. *Educational Leadership*, 68(2), 23-26.
- Lee, C. H., & Kalyuga, S. (2011). Effectiveness of different pinyin presentation formats in learning Chinese characters: A cognitive load perspective. *Language Learning*, 61(4), 1099-1118. doi:10.1111/j.1467-9922.2011.00666.x
- Lee, N. (2011, April). *District understandings of academic language in 16 states*. Paper presented at the AERA Conference, New Orleans.
- Leighton, J. P., & Gokiert, R. J. (2005, April). *The cognitive effects of test item features: Informing item generation by identifying construct irrelevant variances*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.
- Leighton, J. P., Gokiert, R. J., & Cui, Y. (2007). Using exploratory and confirmatory methods to identify the cognitive dimensions in a large-scale science assessment. *International Journal of Testing*, 7(2), 141-189. doi:10.1080/15305050701193520
- Lightbown, P. M., & Spada, N. (2004). *How languages are learned* (2nd ed.). Oxford, England: Oxford University Press.
- Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes*, 31(1), 25-35. doi:10.1016/j.esp.2011.07.002
- Long, M. H. (2007). *Problems in SLA*. New York, NY: Lawrence Erlbaum Associates.
- Madyarov, I. (2009). Widening access to education: A case for bilingual distance curriculum. *International Journal of Instructional Technology & Distance Learning*, 6(3). Retrieved from [http://www.itdl.org/journal/mar\\_09/article02.htm](http://www.itdl.org/journal/mar_09/article02.htm)

- Marinova-Todd, S. H., Marshall, D. B., & Snow, C. E. (2000). Three misconceptions about age and L2 learning. *TESOL Quarterly*, 34(1), 9-31.
- Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review*, 78(2), 333-429.
- Massachusetts Department of Elementary and Secondary Education. (2003). *English language proficiency benchmarks and outcomes for English language learners*. Retrieved from <http://www.doe.mass.edu/frameworks/benchmark.pdf>
- Massachusetts Department of Elementary and Secondary Education. (2006). *Massachusetts science and technology/engineering curriculum framework*. Malden, MA: Author.
- Massachusetts Department of Elementary and Secondary Education. (2010). Education Board adopts Common Core standards to keep Massachusetts students national leaders in education. Retrieved from <http://www.doe.mass.edu/news/news.aspx?id=5634>
- Massachusetts Department of Elementary and Secondary Education. (2011a). *2010-11 Selected populations report (district)*. Retrieved from [http://profiles.doe.mass.edu/state\\_report/selectedpopulations.aspx](http://profiles.doe.mass.edu/state_report/selectedpopulations.aspx)
- Massachusetts Department of Elementary and Secondary Education. (2011b). *2011 MCAS and MCAS-alt technical report*. Malden, MA: Author.
- Massachusetts Department of Elementary and Secondary Education. (2011c). *High school graduation requirements, scholarships, and academic support opportunities*. Retrieved from <http://www.doe.mass.edu/mcas/graduation.html>
- Massachusetts Department of Elementary and Secondary Education. (2011d). *Massachusetts Comprehensive Assessment System overview*. Retrieved from <http://www.doe.mass.edu/mcas/overview.html>
- Massachusetts Department of Elementary and Secondary Education. (2011e). *Massachusetts English proficiency assessment (MEPA) statewide results: Spring 2011*. Retrieved from <http://www.doe.mass.edu/mcas/mepa/2011/results/11state.pdf>
- Massachusetts Department of Elementary and Secondary Education. (2012a). *2011-12 Selected populations report (district)*. Retrieved from [http://profiles.doe.mass.edu/state\\_report/selectedpopulations.aspx](http://profiles.doe.mass.edu/state_report/selectedpopulations.aspx)

- Massachusetts Department of Elementary and Secondary Education. (2012b). 2012-2013 MCAS/Access for ELLs testing schedule and administration deadlines. Retrieved from <http://www.doe.mass.edu/mcas/1213schedule.pdf>
- Massachusetts Department of Elementary and Secondary Education. (2012c). 2012 Item by item results for grade HS biology. Retrieved from <http://profiles.doe.mass.edu/mcas/mcasitems2.aspx?grade=HS&subjectcode=BIO&linkid=21&orgcode=00350000&fycode=2012&orgtypecode=5&>
- Massachusetts Department of Elementary and Secondary Education. (2012d). Accountability reports. Retrieved from <http://www.doe.mass.edu/apa/accountability/default.html>
- Massachusetts Department of Elementary and Secondary Education. (2012e). Bilingual dictionaries and glossaries authorized for use by English language learners on MCAS tests. Retrieved from <http://www.doe.mass.edu/mcas/testadmin/lep-bilingual-dictionary.pdf>
- Massachusetts Department of Elementary and Secondary Education. (2012f). *Guide to interpreting the spring 2012 MEPA reports for schools and districts*. Retrieved from [http://www.doe.mass.edu/mcas/mepa/2012/interp\\_guide.pdf](http://www.doe.mass.edu/mcas/mepa/2012/interp_guide.pdf)
- Massachusetts Department of Elementary and Secondary Education. (2012g). *Massachusetts English proficiency assessment (MEPA) statewide results: Spring 2012*. Retrieved from <http://www.doe.mass.edu/mcas/mepa/2012/results/12state.pdf>
- Massachusetts Department of Elementary and Secondary Education. (2012h). *Spring 2012 MCAS raw-to-scaled score conversion: Science and technology/engineering*. Retrieved from <http://www.doe.mass.edu/mcas/2012/results/spring-conversion.pdf>
- Massachusetts Department of Elementary and Secondary Education. (2012i). *Spring 2012 MCAS tests: Summary of state results*. Retrieved from <http://www.doe.mass.edu/mcas/2012/results/summary.pdf>
- Massachusetts Department of Elementary and Secondary Education. (2012j). *Transitioning English language learners in Massachusetts: An exploratory data review*. Retrieved from [http://www.nciea.org/publications/Transitioning%20ELL\\_CD12.pdf](http://www.nciea.org/publications/Transitioning%20ELL_CD12.pdf)
- Massachusetts Department of Elementary and Secondary Education. (2012k). *SIMS data handbook version 3.1*. Retrieved from <http://www.doe.mass.edu/infoservices/data/sims/DataHandbook.pdf>

- Massachusetts Department of Elementary and Secondary Education. (2013a). *2012 MCAS and MCAS-alt technical report*. Retrieved from [http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2012\\_Tech/2011-12%20MCAS%20Tech%20Rep.pdf](http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2012_Tech/2011-12%20MCAS%20Tech%20Rep.pdf)
- Massachusetts Department of Elementary and Secondary Education. (2013b). *Appendix G -- Differential item functioning results Massachusetts English Proficiency Assessment (MEPA) 2011–12 technical report*. Retrieved from <http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2012/Appendix%20G%20-%20DIF%20Results.pdf>
- Massachusetts Department of Elementary and Secondary Education. (2013c). *Appendix R: Analyses and reporting decision rules – MCAS*. Retrieved from [http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2012\\_Tech/Appendix%20R%20-%20Analyses%20and%20Reporting%20Decision%20Rules%20-%20MCAS.pdf](http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2012_Tech/Appendix%20R%20-%20Analyses%20and%20Reporting%20Decision%20Rules%20-%20MCAS.pdf)
- Massachusetts Department of Elementary and Secondary Education. (2013d). *Massachusetts English Proficiency Assessment (MEPA) 2011–12 technical report*. Retrieved from <http://www.mcasservicecenter.com/documents/MA/Technical%20Report/2012/MEPA%202011-12%20Technical%20Report.pdf>
- Massachusetts Department of Elementary and Secondary Education. (2013e). MCAS results. Retrieved from <http://www.doe.mass.edu/mcas/results.html>
- Massachusetts Department of Elementary and Secondary Education. (2014). MCAS performance appeals. Retrieved from <http://www.doe.mass.edu/mcasappeals/filing/portfolio/>
- McMillan, J. H. (2012). *Educational research fundamentals for the consumer* (6th ed.). Boston, MA: Pearson.
- Menken, K. (2008). *English learners left behind: Standardized testing as language policy*. Clevedon, England: Multilingual Matters, Ltd.
- Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *IRAL: International Review of Applied Linguistics in Language Teaching*, 45(3), 241-259.  
doi:10.1515/iral.2007.011
- Murphy, J. (2010). *The educators handbook for understanding and closing achievement gaps*. Thousand Oaks, CA: Corwin.

- Nevárez-La Torre, A. A. (2012). Transiency in urban schools: Challenges and opportunities in educating ELLs with a migrant background. *Education & Urban Society*, 44(1), 3-34. doi:10.1177/0013124510380911
- Norris, S. P., & Phillips, L. M. (2009). Scientific literacy. In D. Olson & N. Torrance (Eds.), *The Cambridge handbook of literacy* (pp. 271-285). New York, NY: Cambridge University Press.
- Office of English Language Acquisition & Academic Achievement. (2012, May). *The state of the state: A report on English language learners in Massachusetts*. Paper presented at the MATSOL 2012 Conference, Framingham, MA.
- Office of English Language Acquisition & Academic Achievement. (2013, May). *The state of the state: A report on English language learners in Massachusetts*. Paper presented at the MATSOL 2013 Conference, Framingham, MA.
- Paas, F., & Sweller, J. (2012). An evolutionary upgrade of cognitive load theory: Using the human motor system and collaboration to support the learning of complex cognitive tasks. *Educational Psychology Review*, 24(1), 27-45. doi:10.1007/s10648-011-9179-2
- Paas, F., Van Gog, T., & Sweller, J. (2010). Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational Psychology Review*, 22(2), 115-121. doi:10.1007/s10648-010-9133-8
- Powers and Duties of the Department of Elementary and Secondary Education, Massachusetts General Laws c. 69 § 1 (1993).
- Ripley, A. (2010). Your child left behind. *The Atlantic*. Retrieved from <http://www.theatlantic.com/magazine/archive/2010/12/your-child-left-behind/308310/>
- Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *IRAL: International Review of Applied Linguistics in Language Teaching*, 43(1), 1-32.
- Robinson, P., & Gilabert, R. (2007). Task complexity, the Cognition Hypothesis and second language learning and performance. *IRAL: International Review of Applied Linguistics in Language Teaching*, 45(3), 161-176. doi:10.1515/iral.2007.007
- Rossell, C. (2005). Teaching English through English. *Educational Leadership*, 62(4), 32-36.

- Schaap, P. (2011). The differential item functioning and structural equivalence of a nonverbal cognitive ability test for five language groups. *SAJIP: South African Journal of Industrial Psychology*, 37(1), 137-152. doi:10.4102/sajip.v37i1.881
- Schlepppegrell, M. J. (2001). Linguistic features of the language of schooling. *Linguistics and Education*, 12(4), 431-459. doi:10.1016/s0898-5898(01)00073-0
- Schlepppegrell, M. J. (2004). *The language of schooling*. New York, NY: Routledge.
- Schneider, M. C., Huff, K. L., Egan, K. L., Gaines, M. L., & Ferrara, S. (2013). Relationships among item cognitive complexity, contextual demands, and item difficulty: Implications for achievement-level descriptors. *Educational Assessment*, 18(2), 99-121. doi:10.1080/10627197.2013.789296
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth.
- Shank, P. (2007). (Not) making it hard(er) to learn, Part 5. *Online Classroom*, 4-5.
- Short, D. J. (1991). *How to integrate language and content instruction* (2nd ed.). Washington, DC: Center for Applied Linguistics.
- Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4), 487-512.
- Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49(1), 93-120. doi:10.1111/1467-9922.00071
- Snow, C. E. (2010). Academic language and the challenge of reading for learning about science. *Science*, 328(5977), 450-452.
- Snow, C. E., & Hoefnagel-Höhle, M. (1978). The critical period for language acquisition: Evidence from second language learning. *Child Development*, 49(4), 1114-1128. doi:10.1111/1467-8624.ep10713138
- Snow, C. E., & Uccelli, P. (2009). The challenge of academic language. In D. Olson & N. Torrance (Eds.), *The Cambridge handbook of literacy* (pp. 112-133). New York, NY: Cambridge University Press.
- Solano-Flores, G., & Li, M. (2009). Language variation and score variation in the testing of English language learners, native Spanish speakers. *Educational Assessment*, 14, 180-194. doi: 10.1080/10627190903422880



- Solorzano, R. E. (2008). High stakes testing: Issues, implications, and remedies for English language learners. *Review of Educational Research*, 78(2), 260-329. doi:10.3102/0034654308317845
- Standards in your state. (2012). Retrieved from <http://www.corestandards.org/in-the-states>
- Stapp, Y. (2003). Facilitating the aquisition of science concepts in L2. *TEFL Web Journal*, 2(1), 31-48. Retrieved from [http://www.teflweb-j.org/v2n1/science\\_concepts.pdf](http://www.teflweb-j.org/v2n1/science_concepts.pdf)
- Stuftt, D. L., & Brogadir, R. (2011). Urban principals' facilitation of English language learning in public schools. *Education & Urban Society*, 43(5), 560-575. doi:10.1177/0013124510380720
- Sweller, J., & Chandler, P. (1991). Evidence for cognitive load theory. *Cognition & Instruction*, 8(4), 351.
- Tan, M. (2011). Mathematics and science teachers' beliefs and practices regarding the teaching of language in content learning. *Language Teaching Research*, 15(3), 325-342.
- Tanenbaum, C., Boyle, A., Soga, K., Floch, K. C. L., Golden, L., Petroccia, M., . . . O'Day, J. (2012). *National evaluation of Title III implementation — report on state and local implementation*. Washington, DC: U.S. Department of Education. Retrieved from <http://www2.ed.gov/rschstat/eval/title-iii/state-local-implementation-report.pdf>
- Turk, C., & Kirkman, J. (1989). *Effective writing: Improving scientific, technical and business communication* (2nd ed.). New York, NY: E & FN Spon.
- Van Gog, T., Paas, F., & Sweller, J. (2010). Cognitive load theory: Advances in research on worked examples, animations, and cognitive load measurement. *Educational Psychology Review*, 22(4), 375-378. doi:10.1007/s10648-010-9145-4
- van Goor, R., & Heyting, F. (2008). Negotiating the world: Some philosophical considerations on dealing with differential academic language proficiency in schools. *Educational Philosophy & Theory*, 40(5), 652-665. doi:10.1111/j.1469-5812.2008.00452.x
- Van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, 17(2), 147-177. doi:10.1007/s10648-005-3951-0



- Wang, M., Koda, K., & Perfetti, C. A. (2003). Alphabetic and nonalphabetic L1 effects in English word identification: A comparison of Korean and Chinese English L2 learners. *Cognition*, 87(2), 129-149.
- Wardhaugh, R. (2006). *An introduction to sociolinguistics* (5th ed.). Malden, MA: Blackwell Publishing.
- WIDA. (n.d.). Consortium members. Retrieved December 22, 2012, from <http://www.wida.us/membership/states/>
- Xu, Y., & Drame, E. (2008). Culturally appropriate context: Unlocking the potential of response to intervention for English language learners. *Early Childhood Education Journal*, 35(4), 305-311. doi:10.1007/s10643-007-0213-4
- Zafar, M. (2009). Monitoring the “monitor”: A critique of Krashen's five hypotheses. *The Dhaka University Journal of Linguistics*, 2(4), 139-146.
- Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment*, 9(1/2), 61-78.
- Zubrzycki, J. (2011, September 28). Feds prompt Massachusetts to review ELL programs.. *Education Week*, 31(5), 16.