AN EXPLORATION OF THE ROLE OF ENGLISH LANGUAGE PROFICIENCY IN ACADEMIC ACHIEVEMENT

by

ADAM C. WITHYCOMBE

A DISSERTATION

Presented to the Department of Educational Methodology, Policy, and Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Education

December 2014

DISSERTATION APPROVAL PAGE

Student: Adam C. Withycombe

Title: An Exploration of the Role of English Language Proficiency in Academic Achievement

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Education degree in the Department of Educational Methodology, Policy, and Leadership by:

David Conley	Chairperson
Keith Hollenbeck	Core Member
Aki Kamata	Core Member
Audrey Lucero	Institutional Representative

and

J. Andrew Berglund	Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded December 2014

DISSERTATION ABSTRACT

Adam C. Withycombe

Doctor of Education

Department of Educational Methodology, Policy, and Leadership

December 2014

Title: An Exploration of the Role of English Language Proficiency in Academic Achievement

The purpose of this study was to examine the relationship between English language proficiency scores as measured by the ACCESS for ELLs and achievement and growth scores on the reading subtest of the Measures of Academic Progress (MAP). The sample consisted of 2,006 3rd-5th grade English language learners (ELLs) from a large Midwestern school district. Results confirmed that an increase in English proficiency is associated with higher reading achievement scores. The unique variance explained by each of the domain scores (reading, writing, speaking, and listening) on the ACCESS for ELLs supports the use of a weighted composite score for decision making purposes. When considering within-year MAP growth by differing levels of proficiency, a curvilinear trend emerged. The two lowest proficiency groups demonstrated significantly lower reading growth than the two moderate and two highest proficiency groups. The greatest growth was seen by the two groups in the middle of the proficiency spectrum. Given the increased demands on measuring the achievement and progress of all students, including ELLs, and the use of standardized achievement scores for program and teacher evaluation, the results of this study suggest that a dichotomous classification of ELL/non-ELL might not accurately reflect the variability in growth at various levels of English proficiency.

Implications for interpreting and using scores by ELLs are discussed.

CURRICULUM VITAE

NAME OF AUTHOR: Adam C. Withycombe

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
Willamette University, Salem, Oregon

DEGREES AWARDED:

Doctor of Education, Educational Methodology, Policy, and Leadership, 2014, University of Oregon
Master of Arts in Teaching, 2001, Willamette University
Bachelor of Arts, Spanish and Anthropology, 1998, Willamette University

AREAS OF SPECIAL INTEREST:

Early Learning Assessment in Reading and Math
Assessment Literacy
English Language Learners

PROFESSIONAL EXPERIENCE:

Early Learning Content Specialist, Northwest Evaluation Association, Portland, Oregon, 2012-Present

3rd and 5th Grade Teacher, Knox County Schools, Knoxville, Tennessee, 2007-2011

7th/8th Grade TAG Language Arts/Social Studies Teacher, Walla Walla Public School, Walla Walla, Washington, 2005-2006

3rd Grade Bilingual (Spanish) Teacher, Walla Walla Public Schools, Walla Walla, Washington, 2003-2005

5th Grade Bilingual (Spanish) Teacher, Woodburn Public Schools, Woodburn, Oregon, 2001-2003

ACKNOWLEDGMENTS

I wish to express appreciation to Dr. Conley and Dr. Hollenbeck for their assistance in the preparation of this manuscript. In addition, I would like to thank Dr. Kamata for his statistical assistance and deep understanding of the ACCESS for ELLs. Dr. Lucero provided valuable insight into the needs of English language learners and second language acquisition. I also wish to thank the Department of Research and Development for the participating district for their generous access to such quality data.

As the author of this study, I wish to declare that I am a current employee of NWEA, working as a content specialist in Early Learning with projects directly related to the MAP and MAP for Primary Grades. This study has been conducted outside the organization and data was received directly from the participating school district. The findings from this study have no bearing on my employment within NWEA. To my colleagues, their support and encouragement was and is greatly appreciated.

To my wife and daughters who provided support, patience, and encouragement through this process. And to my parents who instilled in me the value of education.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

In the United States, student academic achievement is measured using large-scale standardized assessments. The resulting scores indicate individual performance as a function of student growth and are assumed to provide an accurate measure of teacher effectiveness. One of the main purposes of standardized assessments is to hold teachers and schools accountable for student performance. Although such endeavors are exemplary in theory, they are infinitely complex in implementation and practice. Considering the unique issues faced by certain populations (e.g., students with disabilities, minority students, students from low socioeconomic backgrounds, and English language learners) a paralleled relationship between student achievement scores and student growth and/or teacher effectiveness may be difficult to support. As Bailey, Butler and Sato (2007) point out, the continuum of English language proficiency alone "poses challenges for evaluating linkages between the ELD standards and the academic content standards that are not organized on a developmental continuum" (p. 74).

In 2002, No Child Left Behind (NCLB) drastically changed the nature of how English language proficiency was perceived by tying proficiency to academic language (CALP) (Albers, Kenyon, & Boals, 2009). Statistically, English language learners are the fastest growing educational subgroup in the nation (Brooks, Adams, & Morita-Mullaney, 2010; Sullivan, 2011; Young, Cho, Ling, Cline, Steinberg, & Stone, 2008). From 1994 to 2004, the general student population grew by approximately 2%, but the ELL population during that same time grew approximately 60% (Wolf, Kao, Griffin, Herman, Bachman, Chang, & Farnsworth, 2008). In 2009, the number of school age

students whose home language was something other than English was 11.2 million, up from 4.7 million in 1980 (Aud, Hussar, Kena, Bianco, Frohlich, Kemp, & Tahan, 2011; Aud, Wilkinson-Flicker, Kristapovich, Rathbun, Wang, & Zhang, 2013). Twenty-one percent of the total school-age population speaks a language other than English in the home (Aud et al., 2011; Aud et al., 2013). Although ELLs represent more than 400 different languages (Wolf, Kao, et al., 2008), recent demographic data indicate that these students are overwhelmingly Spanish-speaking (72%) (Aud et al., 2011; Aud et al., 2013). According to Albers and colleagues (2009), an estimated 40% of the public education population will speak English as a second language by the year 2030 (Albers et al., 2009). Although ELP assessments are designed to help educators identify students' levels of English language proficiency (WIDA, 2013), academic achievement tests do not, nor can they, account for varied levels of English language proficiency (NWEA, 2011). In fact, giving ELL students academic achievement tests violates one of the primary assumptions of standardized assessments: that the test-takers have no linguistic barriers that might impede their ability to perform on the test (Chen, 2010). Without a direct link between ELP assessments and academic achievement outcomes, a designation of *proficient* in English that is not also associated with higher achievement scores limits the predictive validity of the ELP assessment.

As the proportion of English language learners has grown, the achievement gap between ELLs and non-ELLs has changed little over the last decade (Hemphill, & Vanneman, 2011). The 2011 results for the National Assessment of Educational Progress (NAEP), for example, indicate that the mean reading scores for ELL students were 188, while the mean scores for non-ELLs were 225 (USDOE, IES, NCES, 2012). In terms of

proficiency, 37% of non-ELLs scored *At* or *Above Proficient* compared to only 7% of ELLs. From 2003 to 2011, non-ELL students gained 4% in the number of students rated *At* or *Above Proficient* for reading compared to no change for non-ELLs. (Hemphill & Vanneman, 2011). Although the scores for ELLs and non-ELL alike show increases over time, the gap remains persistent and stable (Hemphill & Vanneman, 2011).

Based on the current classification decisions and reporting criteria for accountability purposes, ELLs are often treated as a static group, ignoring both the diversity of the language proficiency continuum and the dynamic change in proficiency within an academic year. This dichotomous split between ELL and non-ELL fails to take ELLs' English proficiency into consideration, which can result in the misinterpretation of scores on academic achievement scores.

In response, this study explores the relationship between ELP and academic achievement, paying specific attention to the application of standardized assessments scores in determining student reading achievement and within-year growth. Findings provide evidence for the utility of ELP scores in determining readiness for content instruction in English and how the consideration of the level of ELP might impact the interpretation of achievement and growth scores. Topics covered in the following literature review will include (a) the process of language development, (b) the varied definitions of ELLs as well as gaps in achievement unique to this population, (c) the link between ELP and accurately interpreting academic achievement, and (d) validity and assessment concerns with certain standardized exams (e.g., The *ACCESS for ELLs* and the *MAP*). By providing such a comprehensive foundation I hope that the mission laid out by Sireci, Han, and Wells (2008) when they said that educational researchers are

"obligated to use [their] statistical tools to help educational assessment policymakers understand the degree to which the inferences derived from ELLs' test scores are valid . . . and to promote fair and accurate assessment practices for ELLs" can be fulfilled (p. 128).

CHAPTER II

LITERATURE REVIEW

The following section provides an exploration of the role of English language proficiency in academic achievement.

Stages of Language Development

The acquisition of a second language is widely regarded as a progression during which, for example, English language learners move from basic English language proficiency to advanced oral and academic language proficiency (Cummins, 2000). Although several valid attempts to describe the phases of language learning exist, for the purpose of this study the World-Class Instructional Design and Assessment (WIDA) Consortium were reviewed because they are generally regarded as the most comprehensive and direct. The WIDA Consortium has identified six levels of English language proficiency (WIDA, 2013). The levels begin with (PL1) Entering, and progress through (PL2) Beginning, (PL3) Developing, (PL4) Expanding, (PL5) Bridging, and (PL6) Reaching (Source). Three criteria have been used to form the parameters of each level. Each level is based on the students' increasing

> (1) comprehension and use of the technical language of the content areas; (2) linguistic complexity of oral interaction or writing; and (3) development of phonologic, syntactic and semantic understanding or usage as they move through the second language acquisition continuum. (WIDA, 2013)

These levels are identified through the administration of their English language development (ELD) standards-driven assessment called the Assessing Comprehension and Communication in English State to State for English Language Learners (*ACCESS*

for ELLs). By using the *ACCESS for ELLs*, teachers and administrators can better individualize curriculum materials and build more coherent instructional plans (WIDA, 2014).

Although the *ACCESS for ELLs* and other ELP assessments provide valuable information regarding language acquisition, it merely places students on a continuum of proficiency and suggests that a progression occurs. They do not show how and, by extension, how long the process to reach English proficiency will take. According to Hakuta (2000), "even in districts that are considered the most successful in teaching English to English language learners, oral proficiency takes three to five years to develop and academic English can take four to seven years" (p. 13). Cook, Boals, and Lundberg (2011) suggest that English learners grow at different rates, and that "these growth rates are mediated by many factors; clearly, one is students' initial proficiency level" (p. 69). For example, a study conducted by Cook and Zhao (2011) showed that 67% of students with an initial score of PL4 attained proficiency within 5 years and less than 40% of those students were proficient in 1 year (Cook & Zhao, 2011). Fewer than 40% of students with an initial score of L3 were proficient in 5 years and about 10% of those in 1 year (Cook & Zhao, 2011). When considering students with an initial proficiency of L1, only 10% of students achieved proficiency within 5 years (Cook & Zhao, 2011). English language learners are not an homogeneous group and each level of proficiency has its own set of needs and rates of language acquisition (Cook & Zhao, 2011).

Defining English Language Learners

Students for whom English is not their native language are referred to by many names, and are classified by a variety of procedures. The terms limited English proficient

(LEP) and English language learner (ELL) are the most prevalent terms used by the Department of Education and most U.S. states (Wolf, Kao, et al., 2008). Both refer to the proficiency level of the speaker when compared to that of native English speakers. Wolf, Kao, et al. (2008) and others (August & Shanahan, 2008; Carlo et al., 2004) suggest that the terms are interchangeable, though ELL is preferable to LEP due to the negative connotation of the word *limited* which is suggestive of a deficit (Abedi, 2008a; Bailey & Kelly, 2010). ELL is an official designation that depends on the language spoken in the home and the level of English proficiency as demonstrated by an English language proficiency (ELP) assessment or other observational method (Bailey & Kelly, 2010). Non-native speakers who do not qualify as ELLs are referred to as L2 speakers, language-minorities, or emergent bilinguals (Baker, 2003). The WIDA Consortium defines English language proficiency as "the point at which students' English language proficiency becomes less related to academic achievement. Beyond this point, ELL's performance on content assessments is more related to content knowledge than to language proficiency" (Cook et al., 2011, p. 68). Students from non-English-speaking homes who are fluent in English at the time of school entry are labeled *Initially Fluent English Proficient* (IFEP) (Abedi, 2008a). Those students who demonstrate English proficiency during schooling and progress out of the ELL category are labeled *Reclassified Fluent English Proficient* (RFEP) or reclassified as non-ELL (Abedi, 2008a). Classification decisions are made based on performance on an ELP measure; however, unique issues arise when attempts are made to better understand the diversity and subtleties of the language acquisition process and its relationship to academic achievement.

English language proficiency. Classifying English language learners is not a straightforward process. Title VII of the ESEA of 1968, also known as the Bilingual Education Act, provided the first LEP designation (Bunch, 2011). The ESEA recognized that English learners have specific educational needs in order to derive equal benefit from educational opportunities. *Lau v. Nichols*, a class action suit in 1974 brought against the San Francisco school district, paved the way for the first amendments to the Bilingual Education Act (*Lau* v. *Nichols*, 1974). In their decision, the U.S. Supreme Court found that a lack of linguistically appropriate accommodations denied Chinese students equitable access to educational opportunities. The resulting amendments to the Bilingual Education Act required schools take proactive steps to increase English proficiency rather than merely providing access to books and teachers (Stewner-Manzanares, 1988). Additionally, the low-income requirement of the ESEA was removed, meaning that language needs were assessed independent of income level, thus eliminating the assumption that only low-income ELLs needed language support services (Stewner-Manzanares, 1988). Changes continued from the 1970s through the 1990s and each decade saw an increased focus on improving opportunities for ELL populations.

Before the reauthorization of ESEA, language proficiency was based largely on language acquisition theory, which focused primarily on social language (e.g., the ability to use English in social situations) and the ability to communicate with teachers and peers (Albers et al., 2009). Early ELL standards were focused on the same content (and held students to similar expectations for performance) as the English Language Arts (ELA) standards without attending to issues of academic language and the complexity of content-area language demands (Llosa, 2011). Cummins (1980a, 1980b) recognized the

need for more sophisticated language related to academic content areas. Cummins (1980b) divided language proficiency into two dichotomous categories consisting of Basic Interpersonal Communicative Skills (BICS) and Cognitive/Academic Language Proficiency (CALP). The social aspects of BICS include the features of language such as accent, oral fluency, and sociolinguistic competence (Cummins, 1980b). CALP is characterized by the more cognitively demanding aspects of language use such as comparing, critiquing, and synthesizing, as well as the domain-specific vocabulary needed to access academic content (Cummins, 1980a). Zwiers (2008) suggested that the social elements (BICS) are less complex and less abstract than the language demands involved in learning academic content (CALP). BICS are generally acquired in 2-3 years, compared to CALP, which takes between 5-7 years to develop (Cummins, 2000; Young et al., 2008). Such distinctions have significant implications regarding the relative accuracy of measuring student achievement given the time it takes CALP to develop.

In 2002, No Child Left Behind (NCLB) drastically changed the nature of how English language proficiency was perceived by tying proficiency to academic language (CALP) (Albers et al., 2009). According to Abedi (2008a),

> a valid [language proficiency] classification system should be based on the theory of second language acquisition and should clearly identify the level of academic language proficiency that is needed for ELL students to function in academic environments where both instruction and assessment are offered only in English. (p. 29)

Given the link between academic language and subsequent access to content, the shift in philosophy from determining proficiency based on BICS to one based on CALP was an

integral component of increasing accountability for the achievement of ELLs (Albers et al., 2009). Gaining a more intentional focus on academic language is important, however, there is currently no consensus on what constitutes academic language or at which levels a student could be considered proficient (Wolf, Farnsworth, & Herman, 2008). Without a direct link between ELP assessments and academic achievement outcomes, a designation of proficient in English limits the predictive validity and therefore accuracy of achievement scores.

Academic English. Considerable focus on what is meant by *academic language* began with the introduction of BICS and CALP (Bailey & Heritage, 2008; Cummins, 2000; Luke, 2000; Scarcella, 2008). Variations in specific terminology include *Academic Language Proficiency* (ALP), *Academic Proficiency* (AP), and more commonly, *Academic English* (AE) (Anstrom et al., 2010; Bailey & Heritage, 2008; Chen, 2010). Anstrom et al. (2010) note considerable variability in the operationalized definitions of *academic language*. Bailey and Heritage (2008), for example, further divided AE into School Navigational Language (SNL) and Curriculum Content Language (CCL), noting that some academic language demands cover multiple content areas while others are domain specific. SNL, for example, is characterized as the broad skills used to communicate with teachers and peers, while CCL relates to the process of teaching and learning specific academic content (Anstrom et al., 2010).

Although some researchers attempt to identify features that are used across all content areas (Scarcella, 2008), others (Beck & McKeown, 2007; Stevens, Butler, & Castellon-Wellington, 2000) focus specifically on the idea of multiple tiers of vocabulary, including high frequency, non-specialized academic, and specialized

domain-specific content words. Consistent among the many definitions of academic language is the understanding that limited proficiency in this area means a student will be less able to benefit from content instruction presented in English (Albers et al., 2009; Cummins, 1980b; 2000). According to Cummins's (1979) Threshold Hypothesis, the positive aspects of bilingualism do not come into effect until the student has reached a minimum threshold of competence in their second language. In a 2000 revision of Cummins's earlier hypothesis, the impact of discriminatory schooling is explained. In this scenario no allowances are made for access to literacy and comprehensible academic language in both the student's native and secondary languages, therefore placing the student at a significant disadvantage (Ardasheva, Tretter, & Kinny, 2012). In this instance, the disadvantage is related to the ability to comprehend academic language. If ELLs have yet to meet the minimum threshold of competence in their second language they may not understand teacher instructions, which can have a drastic impact on learning. Research shows that it takes between 2-5 years for bilingual students to develop mastery of their oral skills (e.g., sound discrimination, vocabulary, listening comprehension, and oral expression, syntactic, morphological, and pragmatic skills) but up to 7 or more years to reach high levels of literacy skills comparable to native English speakers (Ardasheva et al., 2012; Cook & Zhao, 2011).

Current measures of ELP are now framed within the context of academic language and the four primary content areas: language arts, math, science, and social studies (Albers et al., 2009; Llosa, 2011). Albers et al. (2009) mention that stronger correlation between modern ELP measures based on academic language and academic achievement tests, as compared to more socially constructed measures, indicates more

instructional utility and predictive value of ELP scores. Because academic English and specialized vocabulary knowledge are important aspects of academic achievement, it is important to identify how lower proficiency on those language fields might restrict performance on academic achievement measures.

ELLs and the Achievement Gap

As mentioned in Chapter I, the number of non-native English speakers in the student population is growing at an astronomical rate (Brooks, Adams, & Morita-Mullaney, 2010; Sullivan, 2011; Young et al., 2008). According to Ryan (2013) the percent of people speaking a language other than English increased by 158% between 1980 and 2011. For some populations, such as the Vietnamese population, the increase is nearly 599% (Ryan, 2013). From 1980 to 2011 the United States saw an increase of 25.9 million Spanish speakers (Ryan, 2013). The remaining non-native English speaking population included Indo-European (13%), Asian/Pacific Islander (11%), and other dialects (4%) (Aud et al., 2011; Aud et al., 2013; Ryan, 2013). French, German, Russian and French Creole accounted for the majority of the Indo-European group followed by Polish, Persian, Hindi, and other Indic, Slavic, and Indo-European languages (Ryan, 2013). Based on results from the 2011 American Community Survey, this group self-reported the highest level of English language proficiency (Ryan, 2013). The Asian/Pacific Islander groups reported the lowest English proficiency with 9.7% of Chinese and 7.3% Vietnamese speakers reporting no English (Ryan, 2013). Tagalog had a higher self-reported proficiency than the rest of the group which primarily included Chinese, Korean, Vietnamese, and Tagalog followed by Japanese, Cambodian, Hmong, Thai, and other Asian/P.I. dialects (Ryan, 2013). The remaining dialects of note were

primarily Arabic followed by African languages, Navajo/Native American, Hungarian, and Hebrew (Ryan, 2013). Although a categorical organization of language based primarily on geography might be simpler, it likely masks the uniqueness of each language in a particular group.

Given that the proportion of English language learners is growing, the achievement gap between ELLs and non-ELLs has changed little over the last decade (Hemphill, & Vanneman, 2011). The 2011 results for the National Assessment of Educational Progress (NAEP), for example, indicate that the mean reading scores for ELL students were 188, while the mean scores non-ELLs were 225 (USDOE, IES, NCES, 2012). As mentioned earlier, the *At* or *Above Proficiency* rates are at least 30% below those of native English speakers (Hemphill & Vanneman, 2011). Although the scores for ELLs and non-ELLs alike show increases over time, the gap in proficiency remains persistent and stable (Hemphill & Vanneman, 2011). Moreover, because the ELL population is constantly being refreshed with non-proficient students as proficient students exit, calculating rates of progress for ELLs with any sense of certainty is exceedingly complex.

Link Between ELP and Academic Achievement

Due to the diversity and size of the ELL population, as well as the varied definitions of what constitutes *academic language* and *proficient*, using current ELP assessments as evidence of academic achievement warrants further exploration. Alignment between assessments and standards-based content is necessary to accurately demonstrate student progress (Abedi, 2008a; APA, AERA, & NCME, 1999; Cawthon, 2004). This is especially true when using ELP assessments as indicators of academic

readiness in an English-only setting. Not only must ELP assessments demonstrate internal consistency and concurrent validity with other tests that measure the English proficiency construct (Cawthon, 2004), but as Bailey et al. (2007) argued, ELP tests must also align with academic achievement tests if they are to serve as indicators of students' readiness to meaningfully participate in English-only instruction of academic content. As previously stated, students must reach a threshold of proficiency in English in order to benefit from instruction in that language (Cummins, 1979). This alignment is further complicated when ELP is based on a developmental continuum, whereas academic achievement may not be (Bailey et al., 2007).

A strong alignment between ELP and academic content would suggest that students have access to the types of language experiences in the classroom that would support their development of language proficiency while providing access to content (Mohamud & Fleck, 2010; Wolf, Farnsworth, et al., 2008). Some researchers contend that ELP standards and assessments lack specificity to the language requirements of the content area settings and are too generalized to effectively inform instruction (Bailey et al., 2007; Llosa, 2011). In their alignment study of 5th grade science and English Language Development (ELD) standards, for example, Bailey and colleagues (2007) found that only 34% of all science standards were aligned with the knowledge and performance expectations for ELLs in the state's ELD standards. Additionally, confounding interactions between linguistic requirements, academic content, and prior academic experiences are difficult to disentangle (Ferrara, 2008; Wolf, Farnsworth, et al., 2008). In a study of the English Language Development Assessment (ELDA) speaking subdomain, Ferrara (2008) found that at all grade-level clusters students had difficulty on

social studies items because they were not familiar with the academic content. Further, certain language structures, such as using the past tense, innately critical to social studies curricula, were also more difficult. It is important to account for how variations on one could impact the interpretation of results on the other, realizing that no perfect overlap exists between measures of ELP and academic achievement.

Validity and Assessment

Concerns regarding the alignment of ELP and academic assessments call in to question many of the foundational assumptions of standardized assessments, such as avoiding multidimensionality and providing equal access to all test takers (APA, AERA, & NCME, 1999). For example, academic achievement tests cannot account for, nor control, the varied levels of English language proficiency (NWEA, 2011; WIDA, 2013). As mentioned previously, giving ELL students academic achievement tests violates one of the primary assumptions of standardized assessments: that the test-takers have no linguistic barriers that might impede their ability to perform on the test (Chen, 2010). The *Standards for Educational and Psychological Testing*, commonly referred to as *the Standards*, makes this point clear,

> For all test takers, any test that employs language is, in part, a measure of their language skills. This is of particular concern for test takers whose first language is not the language of the test. Test use with individuals who have not sufficiently acquired the language of the test may introduce construct irrelevant components to the testing process. In such instances, test results may not reflect accurately students' qualities and competencies. (APA, AERA, & NCME, 1999, p. 91)

The concerns addressed in the Standards speak to the concept of validity, or the degree to which the evidence (e.g., raw scores) supports the interpretations of scores (e.g., level of achievement) (Messick, 1989). Sireci, Han, and Wells (2008) further explain that language proficiency is a common cause of construct-irrelevant variance, and that the inferences based on scores by ELL students may not be comparable to those of proficient English speakers. Specific to ELL students, Sireci and colleagues (2008) suggest that at least four types of validity evidence are needed to support the use of academic test scores for ELL students; content validity, internal structure, response processes, and consequential validity. Wolf, Farnsworth, et al. (2008) note that many ELP assessments are lacking in reported validity evidence, and that most are limited to test content and internal structure such as inter-item reliability or inter-rater agreement, which do not provide evidence of a connection to academic achievement. Without strong evidence linking ELP and academic achievement, scores derived from academic achievement measures must be interpreted with caution, if not flatly rejected.

In an attempt to standardize ELP assessments, Title III created the Enhanced Assessment Grant program funding four consortia to creating their own ELP assessments. These four common assessments include the *ACCESS for ELLs*, Comprehensive English Language Learning Assessment (CELLA), English Language Development Assessment (ELDA), and the Mountain West Assessment (MWA) (Bunch, 2011). Although each assessment operationalizes and assesses ELP somewhat differently, they generally follow similar formats. As per NCLB requirements, the tests measure speaking, reading, writing, and listening with scores ranging from three- to six-point rating scales (Wolf, Kao, et al., 2008). Relative proficiency varies depending on the ELP assessment used.

Given the nature of the current study, the *ACCESS for ELLs* will be used as the primary focus of evaluation and thus will be discussed in greater detail. For further research on the other consortium assessments see: (Wolf, Kao, et al., 2008; Wolf, Farnsworth, et al., 2008; Ferrara, 2008).

The *ACCESS for ELLs*

Of the four consortia-based ELP assessments, the ACCESS is the most widely used and provides the most evidence of internal validity (Wolf, Farnsworth, et al., 2008). Currently 36 WIDA member states use the ACCESS as their ELP assessment (WIDA, 2014). A multiple-choice format is most commonly employed to assess students' reading comprehension and listening skills (Wolf, Farnsworth, et al., 2008). Speaking components are often administered individually and scored by district personnel as opposed to remote scoring centers (Ferrara, 2008). Constructed response item formats are used for writing, which is scored locally or at test developer sites. Generally all of the ELP sections are untimed. Raw scores are converted to scale scores, and cut scores are applied to determine proficiency levels. Scale scores are often reported for each domain, as well as a composite score (Wolf, Farnsworth, et al., 2008).

The focus of the ACCESS is on the social and instructional purposes of language within the school setting. In this sense, the purpose of the ACCESS is to measure the ability to communicate information, ideas, and concepts necessary for successful participation in an academic setting (WIDA, 2013). The ACCESS defines the constructs of the four language domains (listening, speaking, reading, and writing) in the context of the four academic content areas previously mentioned and a fifth classified as social and instructional language (Wolf, Farnsworth, et al., 2008). The four language domains of

the ACCESS are aligned to the World-Class Instructional Design and Assessment (WIDA) ELP standards. The listening, reading, and writing components consist of 39-78 items per grade band. The speaking portion is individually administered and organizes tasks into thematic folders to help provide context and reduce cognitive complexity when moving from one topic to another (Ferrara, 2008). Scale scores are converted to a 6-point system of proficiency levels including: *Entering*, *Beginning*, *Developing*, *Expanding*, *Bridging*, and *Reaching*. *Reaching* is considered the level of proficiency needed to be successful in an English-only setting without extra support (Ferrara, 2008). Although a composite proficiency rating based on all language domains is often used as the criteria for exiting language support services, Abedi (2008b) argues that students should demonstrate proficiency in all subdomains to truly be classified as English language proficient.

Considering that ELP assessments are used to make eligibility and program placement decisions, an accurate interpretation of scores based on these measures is essential (Sireci, Han, & Wells, 2008). Cummins (1980a; 1980b) suggests that students who exit too early from support services will likely experience difficulty accessing academic content in English-only environments. Conversely, those students with higher English proficiency are likely more prepared to participate in rigorous academic interactions (Reardon & Galindo, 2009), and therefore increase their academic achievement. In a study of the influence of reading ability on math achievement by ELL students, Chen (2010) found that once a certain language proficiency threshold is reached, language as a mediating factor on math scores lessens. Chen (2010) also found that language proficiency impacted students differently at the various levels of

mathematical understanding; average to above average math students were less impacted by language than students who scored below average in math. Other studies have also attempted to disentangle language proficiency from academic achievement scores (Bailey et al., 2007; Sireci, Han, & Wells, 2008). According to Chen (2010), "a math score with the language influence (e.g. written English) directly controlled, produces a different magnitude of math achievement gap than if the impact of language is not controlled" (p. 4).

Researchers who attempt to understand achievement independent from language proficiency continually encounter the issue of confounding ELP and academic achievement. Some (Grissom, 2004; Halle, Hair, Wandner, McNamara, & Chien, 2012; Kim & Herman, 2012; Scott, Flinspach, Miller, Gage-Serio, & Vevea, 2009) explore critical transition periods like reclassification, looking at scores of groups before and after meeting state proficiency standards. Others (Hakuta, 2000; Keiffer, 2008) have used ELL status at kindergarten enrollment to create a dichotomous classification, usually distinguishing between ELL students' whose home language is not English but fluent at enrollment, and native English speakers. Chen (2010) used scores from a summative reading assessment as a proxy for ELP when determining impact on math scores. In short, based on any of several methodologies, failing to account for ELL proficiency results in the misinterpretation of scores and misassignments.

Purpose

The purpose of this study is to examine the relationship between a student's level of English language proficiency and their academic achievement and growth as

represented by performance on one standardized assessment of reading. The research questions that will be addressed include:

1. What is the relationship between the level of English language proficiency as measured by the ACCESS for ELLs and academic achievement on the Measures of Academic Progress in Reading for 3rd-5th grade English language learners in a large Midwestern school district?
2. Which ACCESS domain scores (reading, writing, speaking, or listening) best predict the overall RIT Reading Measures of Academic Progress score for 3rd-5th grade English language learners in a large Midwestern school district?
3. What are the differences between the level of English language proficiency and fall-to-spring growth in Reading on the Measures of Academic Progress for 3rd-5th grade English language learners in a large Midwestern school district?
4. Is fall-to-spring growth on the Measures of Academic Progress in Reading for differing levels of ELP consistent with predicted growth from the published Student Growth Norms for 3rd-5th grade English language learners in a large Midwestern school district?

These questions will be investigated using the *ACCESS for ELLs* as the standardized measure of language proficiency and the Reading subtest of the Measures of Academic Progress (*MAP*) as measures of academic reading achievement. Each of these measures will be described in more detail in the following sections.

This study investigates the relationship between the level of English language proficiency and academic achievement. Further, it investigates the assumption that as student English language proficiency increases, so, too, does reading achievement. The

first question seeks to confirm the relationship between ELP and academic achievement that has been described for other achievement measures (Chen, 2010; Wolf, Farnsworth, et al., 2008). The second question explores the relationship between domain scores on the *ACCESS for ELLs*, which contributes to the validity of the composite score as a classification decision tool. The third question moves beyond the dichotomous distinction between proficient and non-proficient students, exploring the spectrum of English language development. By linking students to their ACCESS and *MAP* scores, an opportunity exists to look not only at the variance explained by the various levels of ELP accounted for in the ACCESS, but also to explore growth in academic scores with simultaneous growth in English proficiency.

CHAPTER III

METHODS

The present study employed non-experimental design components to analyze extant 3rd-5th grade data to examine the relationship between the *ACCESS for ELLs* as an ELP assessment and the reading subtest of the *MAP*. The following sections provide (a) demographic information for the district and sample, (b) measurement tools and empirical evidence of reliability and validity, and (c) an explanation of the statistical methods used to analyze the data.

Sample Population

This study was conducted in a large Midwestern school district that serves approximately 78,000 students. According to a recent fact sheet provided by the district, the district student population is approximately 56% African American, 24% Hispanic, 14% white, 6% Asian, and 1% Native American. Twenty percent of students are on Individualized Education Plans (IEP) and receive Special Education (SPED) services, 9% are ELLs, and 82% qualify for free or reduced lunch (FRL). The district also has a 15% mobility rate, which represents the number of students who do not finish the academic year in the school in which they started.

The participants were 2,006 3rd-5th grade English Language Learners from 65 different schools. Table 1 provides a demographic breakdown of the students included in the study.

Table 1

Demographics of a Sample of 3rd-5th Grade ELL Students (n = 2,006)

Characteristic	N	%
Gender		
Male	1083	54
Female	923	46
Race/Ethnicity		
African American	49	2.4
Asian	349	17.4
Hispanic	1548	77.7
Native American	1	0
White	49	2.4
Free/Reduced Lunch		
No	77	3.8
Yes	1931	96.2
Individualized Education Program (IEP)		
No	1721	85.7
Yes	287	14.3
School Language Program		
Bilingual/Dual Language	1448	72.2
ESL Focus	442	22
Traditional	116	5.8
Grade		
3rd	713	35.5
4th	712	35.6
5th	551	28.9

Of the original 2,319 ELL records, 313 were incomplete, resulting in a final sample size of 2,006 participants. The excluded data represents 13% of the district's ELL

population. This missing data is consistent with the district reported mobility rate of approximately 15%. The missing data represents a limitation of this study because students with incomplete records who might represent unique profiles cannot be included.

The students represented a variety of different ELL program models including Bilingual, Dual-Language Immersion, ESL stand-alone programs, and traditional monolingual English classrooms. Seventy-two percent of ELL students attended a Bilingual or Dual Language school. The bilingual programs follow a transitional model with a decreasing proportion of native language time and an increase in English instruction. The PK4 (pre-kindergarten) class begins with a 90/10 native language to English breakdown. Each year the proportion changes by 10 percent until 4th grade when there is a 40/60 split between native language and English respectively. This ratio is maintained through high school. The two designated dual language programs in the district are two-way bilingual programs with instruction occurring in both native language and English. These programs serve both language minority and language majority students with about 50% of students coming from each language group. Because the bilingual and dual language programs both emphasize bilingualism/biliteracy, they were grouped together for this study. ESL stand-alone programs differ from traditional monolingual programs in that they have higher proportions of ELL students and are able to provide more intentional ESL support and services. ESL stand-alone and traditional programs had 22.2% and 5.8% of the ELL sample respectively. It is important to note that the language program should be seen as a school-level factor that, although likely, may not represent the instructional programming for each ELL student in that school.

Student home language and first language were also collected and analyzed for the ELL population. These two variables were categorized using the same four categorical groups used by the U. S. Census Bureau and described previously by Ryan (2013). Table 2 displays the specific language breakdown within each category.

Table 2

Breakdown of Language by Home and First Language Usage (n = 2, 006)

	Home Language		First Language	
	N	%	*n*	%
English	**125**	**6.2**	**0**	**0**
Spanish	**1490**	**74.3**	**1568**	**78.2**
Asian/Pacific Islander	**313**	**15.6**	**345**	**17.2**
Bahasa Indonesian	1	0.3	1	0.3
Burmese	12	3.8	13	3.8
Chin	9	2.9	8	2.3
Hmong	205	65.5	235	68.1
Karen - S'gaw	68	21.7	68	19.7
Kayah Eastern	2	0.6	2	0.6
Lao	7	2.2	9	2.6
Other Chinese	1	0.3	1	0.3
Tagalog/Pilipino	1	0.3	1	0.3
Vietnamese	7	2.2	7	2
Indo-European	**9**	**0.4**	**12**	**0.6**
French	3	33.3	2	16.7
Gujarati	2	22.2	2	16.7
Punjabi	2	22.2	2	16.7
Serbian	0	0	4	33.3
Urdu	2	22.2	2	16.7
Other	**69**	**3.4**	**81**	**4**

	Home Language		First Language	
	N	%	*n*	%
Arabic	28	40.6	38	46.9
Ganda	1	1.4	1	1.2
Maay	14	20.3	14	17.3
Mangingo	0	0	2	2.5
Masalit	5	7.2	5	6.2
Rundi	2	2.9	2	2.5
Somali	12	17.4	13	16
Swahili	4	5.8	3	3.7
Tigrigna	3	4.3	3	3.7

Measures

ACCESS for ELLs. Developed as one of the four Consortia tests authorized by NCLB for Title III reporting purposes (WIDA, 2013), the *ACCESS for ELLs* is the most widely used English language proficiency measure in the United States (Wolf, Kao et al., 2008). The first administration occurred in the spring of 2005 with 3 Consortium states. In 2008, 17 states used the *ACCESS for ELLs* as the ELP measure for Title III reporting purposes (Ferrara, 2008). Currently, 36 WIDA member states use the *ACCESS for ELLs* for all identified ELLs in each state (WIDA, 2014). The *ACCESS for ELLs* meets NCLB reporting criteria for determining English proficiency level, monitoring progress towards proficiency over time, providing instructional information for teachers, and serving as a tool for program evaluation (Ferrara, 2008).

As mentioned previously, the *ACCESS for ELLs* assesses four domains of English proficiency: Reading, Writing, Speaking, and Listening. Within these domains are five

academic content areas that include English language arts (LA), Math (M), Science (S), Social Studies (SS), and Social and Instructional language (SI). By sampling from instructional content areas, the *ACCESS for ELLs* ensures that English proficiency is measured in the context of the skills that students need to be successful in the classroom. The reading, writing, and listening portions are administered in a group setting using selected-response items, whereas speaking is individually assessed. The three group setting portions include items grouped into three tiered folders that span the proficiency spectrum, allowing the administrator to give students more individualized content at their ability level, thus keeping them engaged (WIDA, 2013). The speaking portion is adapted somewhat to the proficiency of the student. Responses start at the basic level and move to more demanding tasks, allowing the proctor to stop when the student fails to score a two or higher on a four-point rubric (Ferrara, 2008).

The *ACCESS* uses four grade level bands (1-2, 3-5, 6-8, and 9-12). This investigation will focus on only the 3rd-5th grade band to avoid statistical complications due to scale score linking. Reliability of the overall composite score for grades 3-5 was .930 (WIDA, 2013), meaning that subsequent administrations of the *ACCESS for ELLs* would very likely result in similar English proficiency level designation.

Raw scores are converted to scale scores that range from 100-600 with a centering value of 350, which represents the cut score between Level 3 and Level 4 at the 5th grade. Scale scores are established for each of the four domains. The domains are then combined to form four composite scores. Table 3 describes the weighting of the various composite scores. Pearson correlations for the four scale scores K-12 range from $r = .58$ to -.90. For grades 3-5, the correlational range is $r = .45$ to .70. The highest reported

correlations across all grade bands are between reading and writing (r = .90). The lowest are between writing and speaking (r = .58).

In a study examining the relationship between student performance on the *ACCESS for ELLs* and the New England Common Assessment Program (NECAP), standardized regression coefficients ranged from .03 to .38 for reading and .15 to .30 for writing (Parker, Louie, & O'Dwyer, 2009). Across all four language domains, ACCESS reading had the highest correlation followed by writing. The study by Parker and colleagues (2009) supports the hypothesis used by the ACCESS that reading and writing should be given more weight in the Overall Composite scores (WIDA, 2013).

Table 3

Composite Score Weighting

Composite Score	**Weight by Domain (%)**
Comprehension	Reading + Listening (70/30)
Oral Language	Speaking + Listening (50/50)
Literacy	Reading + Writing (50/50)
Overall	Reading + Writing + Speaking + Listening (35/35/15/15)

The *ACCESS for ELLs* was also compared to pre-NCLB language proficiency tests including the Language Assessment Scales (LAS), IDEA Proficiency Test (ITP), Language Proficiency Test Series (LPTS), and Revised Maculaitis II (MAC II). The ACCESS demonstrated moderate to strong correlations across all tests and domains (WIDA, 2013). Given the increased focus on academic language in the *ACCESS for ELLs* and lack of very strong correlations to pre-NCLB measures, which focused more on

social language, the *ACCESS for ELLs* measures language proficiency somewhat differently than the pre-NCLB assessments (WIDA, 2013). This move towards an academic language focus likely increases the instructional utility of ELP scores as advocated by Bailey et al. (2007).

Proficiency levels for the *ACCESS for ELLs* are derived from scaled scores. The proficiency levels (PL) are based on a 6-point scale, with partial points representing locations between specific levels. The six levels consist of *Entering*, *Beginning*, *Developing*, *Expanding*, *Bridging*, and *Reaching*. *Reaching* is the highest proficiency level and is intended to represent proficiency across the entire WIDA English language proficiency continuum (WIDA, 2013). The state department of education for this district reclassifies students when they either have an overall composite score of PL 6 or what is known as the F*ive and Five* rule consisting of an overall composite score of PL 5 and a Literacy Composite Score of PL 5. According to Ferrara (2008), students who are categorized as *Reaching* proficiency demonstrate "specialized or technical language reflective of the content area at grade level" (p. 166). Classification accuracy for all cut scores in grades 3-5 ranges from .87 to .99. Classification accuracy at the cut point between level 5 and 6 (*Bridging* and *Reaching*) on the *ACCESS for ELLs*, the point used by the district in this study for eligibility to receive ESL services, ranges from .93 to .95 (WIDA, 2013).

Results are also reported by English language proficiency standard. For comprehension, which includes the five content areas (SI, LA, M, S, and SS), scores are reported as number correct out of a maximum possible based on the form of test taken. The speaking portion of the test includes the raw number of tasks met or exceeded with a

maximum score of 3 for SI, and a score of 5 for LA/SS and M/S. Writing is comprised of three ratings for each of three tasks. The 0-6 ratings are given for Linguistic Complexity, Vocabulary Usage, and Language Control.

Academic achievement based on the *Measures of Academic Progress* (*MAP*). The *MAP* is a computer adaptive interim assessment appropriate for students in grades K-12 (NWEA, 2011). Based on Item Response Theory (IRT), *MAP* scales are derived from the one-parameter logistic IRT model (1PL) with scores reported in Rasch Units (RIT). This enables the creation of a vertically linked, equal-interval scale. This vertical scaling makes it possible for scores to be compared across grades and over time. The test adapts to the test takers' level of performance. If the first question is answered correctly, the test adjusts accordingly and provides a more difficult follow-up question and vice versa. By doing this, testing time and standard error of measurement are significantly reduced compared to fixed form assessments (NWEA, 2011).

Because of the dynamic item selection process of a computer adaptive test, students do not see identical items, making a traditional conception of test-retest reliability impossible. The second test a student sees is similar in content and structure, though different in items from the same pool. As such, NWEA reports a hybrid between test-retest reliability in the form of a parallel forms reliability, conceptualized as the consistency of covalent measures taken across time (NWEA, 2011). Based on the 2009 norms, test-retest reliability for grades 3-5 for the Spring 2008 to Spring 2009, Wisconsin aligned *MAP* reading test were $r = .79$ to .80. The adaptive nature of the *MAP* also makes internal consistency difficult. NWEA reports the marginal reliability coefficient for internal consistency for the Wisconsin aligned reading grades 3-5 ($r = .94$ to .94).

Using information from state achievement test linking studies, NWEA created simulated proficiency cut scores to demonstrate classification accuracy and consistency. Reading classification accuracy and decision consistency were provided at .99 and .99 respectively. In terms of concurrent validity, NWEA reports Pearson Product-Moment Correlations for state accountability tests. The state-aligned *MAP* test for the students in this sample has correlations for grade 3-5 reading of r = .78 to .82.

Reading goal structure. Within the domain of reading, the *MAP* is broken into four goal areas, each consisting of two to three sub-goals. Each test event is balanced by goal area to ensure broad coverage of all reading content areas. All items are dichotomously scored and follow a selected response format. Following the completion of the test, an overall score is reported in RIT and goal areas are presented as a RIT range. Reporting of a RIT range as opposed to a single value for the goal area results from limited item coverage needed for such specificity. The breakdown of the reading test can be seen in Table 4 below:

Table 4

Goal and Sub-goal Structure of the MAP Reading Test

Goal	**Sub-goal**
Determine Meaning of Words, Phrases in Context	Use Context Clues to Determine Meaning Use Knowledge of Word Structure
Understand Text	Understanding of Literal Meaning: Literary Understanding of Literal Meaning: Informational Understanding of Sequence of Events
Analyze Text	Analyze Literary Text Analyze Informational Text Analyze Author's Use of Language

Goal	Sub-goal
Evaluate and Extend Text	Evaluate and Extend Literary Text Evaluate and Extend Informational Text Evaluate and Extend Author's Use of Language

***MAP* as a growth measure.** The *MAP* reports scores on an equal-interval scale for reading. A change in scores from fall to spring testing can be used to determine growth in each of the three domains (reading, language usage, and math). Growth norms are used to determine an expected change in scores from fall to spring testing (Thum & Hauser, 2012). Student norms (Conditional Growth Index – CGI) allow comparisons of individual growth over time to a nationally representative sample. These norms provide the percentile ranks for each testing session, as well as how the change in scores compares to other students in the same grade, domain, and seasonal RIT score. The *MAP* also gives school norms. These norms provide information similar to the student norms, but are designed to interpret the achievement and growth of groups of individuals. Instead of comparing the results of a single third grader, the group performance of all third graders in the school is available.

Statistical Analyses

Multiple statistical techniques were used to address each of the four previously presented research questions. The following section describes the procedures employed and how they addressed each of the research questions.

The first research question explored the relationship between ELL performance on the *ACCESS for ELLs* and the *MAP*. As such, I computed correlations. For the *MAP*, the correlations describe the relationship between scores on subsequent administrations. Intercorrelations of the domain scores on the *ACCESS for ELLs* examined how this ELP

assessment quantifies the four facets of English language proficiency (reading, writing, speaking, and listening) and their relationship to the study's sample. Finally, direct correlations between the *ACCESS for ELLs* domain and composite fall, winter, and spring RIT scores on the *MAP* serve to describe the relationship between the language domains and reading achievement.

The second question attempted to validate the use of weighted composite scores on the *ACCESS for ELLs* for reclassification purposes. I used a hierarchical regression model to evaluate the relative strength of each domain score and describe the variance explained by those scores in the *MAP* spring RIT score for the ELL students. The regression model accounted for the student's initial status (fall RIT) and the following demographics: gender, race, free and reduced lunch status (FRL), whether the student received special education services (SPED) through an individualized education plan, the student's home language, and the type of language program provided by the school attended. The demographic variables were entered in stepwise fashion to evaluate the fewest variables that make a significant contribution to the overall model summary. The regression model was as follows:

$$Y_i = \beta_0 + \beta_1(\text{fall RIT}) + \beta_2(\text{demographic variables as needed}) \ldots + \beta_3 (\text{Reading scale score}) + \beta_4(\text{Writing scale score}) + \beta_5(\text{Speaking scale score}) + \beta_6(\text{Listening scale score}) + e_i$$

The third and fourth questions explored the relationship between English language proficiency and growth on the *MAP*. I compared mean scores on two growth indicators, a simplified gain score and a conditional growth index score, via a series of ANOVAs. Groups were defined by the six proficiency levels. Post hoc Games-Howell

tests and trend analyses were used to describe the observed relationships. Finally, I compared three generalized language proficiency categories based on post hoc analyses or state reclassification criteria to explore differences in conditional growth index scores.

CHAPTER IV

RESULTS

To investigate whether demographic variables might differ for this sample, a series of chi-square analysis were conducted. In terms of grade level differences, only participation in the three types of language programs offered by the district differed significantly ($\chi^2 = 16.58$, $df = 4$, $N = 2{,}006$, $p = .002$). In fifth grade, more students were enrolled in traditional schools and fewer students were enrolled in bilingual schools compared to 3rd and 4th grade enrollment.

Gender differences only existed in the percent of students with IEPs ($\chi^2 = 37.81$, $df = 1$, $N = 2{,}006$, $p < .001$). Males were more likely than expected to have an IEP.

Several differences based on race should be considered. In terms of Free/Reduced Lunch (FRL) eligibility, African Americans and Hispanic students were more likely to qualify than Asian/Pacific Islanders or White students ($\chi^2 = 14.17$, $df = 4$, $N = 2{,}006$, $p = .007$). Cramer's V was .08, which is a small effect. In terms of the proportion of students on an IEP, African American, Asian, Hispanic, and White students had 10%, 7%, 16%, and 20% IEP services respectively ($\chi^2 = 26.85$, $df = 4$, $N = 2{,}006$, $p < .001$). Cramer's V was .12, which was a small to medium effect. However, participation rates in the three district language programs differed significantly by ethnicity ($\chi^2 = 1554.47$, $df = 8$, $N = 2{,}006$, $p < .001$). This difference (Cramer's V = .62) was expected given that the vast majority of bilingual programs are only available in Spanish. For example, 99% of students enrolled in the bilingual schools were Hispanic, which accounted for 92% of the Hispanic population in this sample. Ninety-two percent of African American, 91% of Asian, 3% Hispanic, and 67% of White students attended ESL

focused schools. The remaining 8%, 9%, 5%, and 33% respectively attended traditional schools.

FRL eligibility, by virtue of race differences, was also significant in terms of home language ($\chi^2 = 22.67$, $df = 4$, $N = 2{,}006$, $p < .001$), first language ($\chi^2 = 21.30$, $df = 4$, $N = 2{,}006$, $p < .001$), and district language program ($\chi^2 = 29.43$, $df = 2$, $N = 2{,}006$, $p < .001$). Because race is related to the language group to which a student belongs and therefore the language program s/he likely attends, it was not surprising that FRL demonstrated a small (Craver's V = .10 - .12), but significant difference on these three variables. A similar trend existed in terms of students on IEPs. For these students, home language and first language differed significantly ($\chi^2 = 20.33$, $df = 4$, $N = 2{,}006$, $p < .001$; $\chi^2 = 22.05$, $df = 4$, $N = 2{,}006$, $p < .001$). These differences also translated to district language program enrollment ($\chi^2 = 16.59$, $df = 2$, $N = 2{,}006$, $p < .001$). Because Hispanic students attend bilingual programs and White students attend traditional programs more than their peers, and both are overrepresented with IEPs, it is no surprise that the proportion of students on IEPs in these three programs differed. Bilingual programs had ELL IEP rates of 16%, ESL focus schools had 9%, and traditional schools had 20%.

Question 1: Relationship Between the *ACCESS for ELLs* and *MAP*

Table 5 presents the intercorrelations between the fall, winter, and spring RIT scores on the reading subtest of the *MAP*, as well as the means and standard deviations for each. Due to the nature of the vertical scaling of *MAP* scores, the means and standard deviations differ by grade level. As such, the table was disaggregated by grade level. Because of the demonstrated group differences based on the demographic variables

previously discussed, those demographics were included as controls. The Pearson Correlation coefficients were adjusted accordingly.

Table 5

Intercorrelations, Means, and Standard Deviations for Three MAP Reading Variables, Controlling for Demographics (n = 2,006)

Variable	Fall RIT	Winter RIT	Spring RIT	*M*	*SD*
Fall RIT	**--**	**0.83**	**0.78**	**184.05**	**16.56**
3 (*n* = 713)	--	0.81	0.76	175.57	15.36
4 (*n* =712)	--	0.78	0.75	185.44	15.03
5 (*n* = 581)	--	0.79	0.70	192.75	14.67
Winter RIT	**--**	**--**	**0.80**	**189.39**	**15.89**
3	--	--	0.79	182.63	14.94
4	--	--	0.76	190.63	15.31
5	--	--	0.77	196.15	14.41
Spring RIT	**--**	**--**	**--**	**194.38**	**15.54**
3	--	--	--	188.54	15.2
4	--	--	--	195.18	15.01
5	--	--	--	200.57	13.95

All correlations significant at $p < .001$

As can been seen in the table, the correlations between all three testing seasons ranged from .70 to .83 and were statistically significant ($p < .001$). These strong positive

correlations, which would be considered very large effect sizes according to Cohen (1988), mean that students who had high RIT scores in the fall or winter testing season were likely to have high RIT scores in subsequent seasons. The strongest correlations emerged between adjacent testing seasons with the fall-to-winter relationship slightly stronger that the winter-spring or fall-to-spring seasons. Such consistently high correlations suggest strong reliability of the *MAP* over the course of the school year.

In a similar exploration of domain scores for the *ACCESS for ELLs*, Table 6 presents the intercorrelations, means, and standard deviations for the scale scores in reading, writing, speaking, and listening. Correlation coefficients were adjusted to account for demographic differences, and results were disaggregated by grade level. All correlations were significant ($p < .001$). The strongest of the six associations was between the reading and listening domains ($r = .58$ to $.71$), which would be considered large to very large effect sizes (Cohen, 1988). The high correlation is likely because these two subtests cover the most similar content, differing primarily by the mode in which the information is received. For reading, the student must pull the information off the page, as opposed to the listening subtest which would present the information orally. Reading is also strongly correlated with the writing domain, while only moderately associated with speaking. The speaking domain had the lowest correlations for all three of the remaining domains. Given the differences in correlations between domains, it should be noted that each subtest assesses a slightly different facet of English proficiency.

Table 6

Intercorrelations, Means, and Standard Deviations for Four ACCESS for ELLs Domain Scores, Controlling for Demographic Variables (n =1,874)

Variable	Reading	Writing	Speaking	Listening	*M*	*SD*
Reading	**--**	**.66**	**.45**	**.71**	**338.95**	**29.18**
3 (*n* = 713)	--	.66	.41	.67	327.87	25.66
4 (*n* = 712)	--	.60	.41	.68	343.10	30.84
5 (*n* = 449)	--	.59	.40	.58	349.95	25.67
Writing	**--**	**--**	**.43**	**.57**	**350.05**	**25.97**
3	--	--	.39	.56	341.94	25.82
4	--	--	.39	.51	352.96	25.15
5	--	--	.39	.46	358.32	23.87
Speaking	**--**	**--**	**--**	**.47**	**365.61**	**34.48**
3	--	--	--	.43	357.59	33.97
4	--	--	--	.42	368.42	34.50
5	--	--	--	.44	373.87	32.63
Listening	**--**	**--**	**--**	**--**	**357.93**	**33.73**
3	--	--	--	--	342.99	26.89
4	--	--	--	--	362.78	35.41
5	--	--	--	--	373.92	31.13

All correlations significant at $p < .001$

The difference in fifth grade sample size between the two tables is a result of how the state designates ELLs as Reclassified Fluent English Proficient (RFEP). Students who score an overall proficiency level of 6.0 on the *ACCESS for ELLs* are automatically reclassified. Starting in fourth grade students can also be reclassified with an overall

composite score of 5.0-5.9 *and* a literacy composite of 5.0 or higher. Based on that state criteria, 123 4th grade students from this sample would be considered RFEP as a result of the *Five and Five* rule. If the previous cohort of students experienced a similar number of student reclassifications in this manner, then the difference of 132 students as seen in this sample would be appropriate. Fifth grade students who are RFEP start the year with an overall composite of 6.0 and do not take the *ACCESS for ELLs* during that year. However, their progress is monitored for an additional two years following reclassification. For this study RFEP students still constituted the ELL population, as dictated by the state.

In order to explore the relationship between the *MAP* and the *ACCESS for ELLs,* a correlation between the three testing seasons on the *MAP* and the four domain scores on the *ACCESS for ELLs* was computed. The results are presented in Table 7, and were adjusted for controlled demographics and disaggregated by grade level. Each of the four domain scores was consistent across *MAP* testing seasons. In all cases the *ACCESS for ELL* scores correlated most strongly with the winter *MAP* testing season because of the close proximity of their test dates. Both the *ACCESS for ELLs* and the winter administration of the *MAP* occurred during the month of February for all students.

Reading and writing domain scores had the highest correlations to *MAP* across all testing seasons, ranging from $r = .57$ to .70 for reading and r = .59 to 72 for writing. These were large to very large effect sizes (Cohen, 1988). The relative strength between reading and writing differ slightly by grade level. For example, in 3rd grade, reading had a correlation to Fall RIT of $r = .64$ ($p < .001$) while writing had a correlation of $r = .72$ ($p < .001$).

Table 7

Correlations Between ACCESS for ELLs Domain Scores and Fall, Winter, and Spring RIT, Controlling for Demographic Variables (n = 1,874)

Variable	Fall RIT	Winter RIT	Spring RIT
Reading	**.69**	**.69**	**.65**
3rd (*n* = 713)	.64	.67	.63
4th (*n* = 712)	.61	.60	.57
5th (*n* = 449)	.70	.70	.68
Writing	**.71**	**.70**	**.67**
3rd	.72	.72	.66
4th	.64	.64	.61
5th	.63	.61	.59
Speaking	**.48**	**.50**	**.47**
3rd	.40	.44	.42
4th	.46	.47	.46
5th	.49	.50	.42
Listening	**.58**	**.57**	**.55**
3rd	.53	.55	.51
4th	.47	.47	.48
5th	.51	.52	.50

All correlations significant at $p < .001$

In 5th grade, the correlations were essentially flipped. Reading had the higher correlation of $r = .70$ ($p < .001$) compared to writing with a correlation of $r = .63$ ($p < .001$). This same reversal generally happened at each testing season for these two domain scores. Speaking scores correlated to *MAP* with a range of $r = .40$ to .50 ($p < .001$)

across all testing seasons. Listening demonstrated slightly higher correlations to *MAP*, ranging from $r = .47$ to .58 ($p < .001$). The lower correlations between these two subdomains and *MAP* should not be entirely surprising because the *MAP* is a reading test with no audio or speaking components. Still, their medium to large effect sizes indicates a significant relationship between the two measures.

Because composite scores are used for reclassification decision purposes, it was important to also correlate these scores to the *MAP*. As mentioned previously, *ACCESS for ELL* composite scores are based on weighted combinations of the four domain scores. The literacy composite score and the oral language composite score are based on an equal weighting of reading and writing or speaking and listening respectively. The comprehension composite score is based on a 70/30 weighting of reading and listening. The overall composite, which is the basis for most reclassifications, is comprised of 35% reading, 35% writing, 15% speaking, and 15% listening. Table 8 displays the correlations and means and standard deviations for the four *ACCESS for ELL* composite scores.

Given the strong correlations between reading and writing and the *MAP*, it should not be surprising that the composite scores that weight those two domains the heaviest would have the highest correlations to *MAP* reading scores. The overall and literacy composite scores demonstrated a range of correlations from $r = .65$ to $r = .77$ ($p < .001$) across all grades and *MAP* testing seasons. These are higher than the correlations for either of the two domain scores individually. The comprehension and oral language composites scores also have stronger correlations to the *MAP* than their individual components. Oral language, which consists of the speaking and listening domain scores,

had the lowest correlation to the *MAP*; however, those correlations (r = .54 - .63, p < .001) are still high, suggesting that students with higher English proficiency in speaking and listening will have higher *MAP* reading scores.

Table 8

Correlations Between ACCESS for ELLs Composite Scores and Fall, Winter, and Spring RIT, Controlling for Demographic Variables (n = 1,874)

Variable	Fall RIT	Winter RIT	Spring RIT	*M*	*SD*
Comprehension	**.70**	**.70**	**.67**	**344.75**	**28.62**
3rd (*n* = 713)	.65	.69	.64	332.57	24.30
4th (*n* = 714)	.61	.60	.58	349.10	30.12
5th (*n* = 449)	.70	.71	.69	357.20	24.98
Literacy	**.77**	**.76**	**.72**	**344.75**	**25.37**
3rd	.74	.76	.71	335.17	23.68
4th	.69	.69	.65	348.26	25.50
5th	.75	.74	.72	354.39	22.56
Oral Language	**.62**	**.63**	**.59**	**362.03**	**29.52**
3rd	.54	.58	.54	350.57	26.15
4th	.55	.56	.56	365.84	29.90
5th	.58	.60	.54	374.18	27.53
Overall	**.77**	**.77**	**.73**	**349.68**	**24.84**
3rd	.73	.76	.71	339.52	22.59
4th	.70	.70	.67	353.30	24.94
5th	.76	.76	.72	360.06	22.11

All correlations significant at p < .001

The previous results described the reliability and internal consistency of each of the measures in question. For the *MAP*, correlations between fall, winter, and spring administrations demonstrated the relationship between scores over subsequent administrations. The intercorrelations of the domain scores on the *ACCESS for ELLs* described the relationship between the four facets of English language proficiency. The correlation between the *MAP* and *ACCESS for ELLs* provided a direct link between the two measures.

Question 2: *ACCESS for ELLs* Domain Scores and the *MAP*

Because the composite scores on the *ACCESS for ELLs* are based on weighted domain scores, it is critical to explore whether those weights are appropriate when considering student performance on the *MAP*. Because fall RIT was highly correlated to spring RIT and the sample demonstrated demographic differences, a hierarchical multiple regression was conducted to investigate the best *ACCESS for ELLs* domain scores as predictors of spring RIT while controlling for initial status (fall RIT) and demographics. The demographic variables were entered in Step 2 of the analysis in stepwise fashion to see if any meaningfully improved the model. According to Keppel and Zedeck (1989), the goal of the stepwise strategy is to find the smallest subset of variables that explains a significant amount of variance in the dependent variable. This effectively removes any variables that do not meaningfully contribute to the model. The domain scores were entered in a simultaneous manner so that all four would be required components of the final model. Table 9 shows the unstandardized regression coefficients (*B*), the standard error of *B*, and the standardized regression coefficients (β).

The complete model, as presented in Step 3, consists of fall RIT, SPED status, and the four *ACCESS for ELL* domain scores. This model was statistically significant, Adjusted R^2 = .61; $F(6, 1868) = 486.09$, $p < .001$. Taken together, the model explains 61% of the variance in spring RIT. The four domain scores account for 6% of the variance over and above fall RIT and student demographics. Of note, only reading, writing, and speaking significantly contributed to the model. Semipartial correlation coefficients (*sr*) for each of the domain scores were as follows: Reading = .10, Writing = .11, Speaking = .06, and Listening = .03, meaning each domain score accounts for up to 1% unique variance explained after controlling for other factors. Follow-up analyses to examine interaction effects between SPED status and the four ACCESS for ELLs domain scores were not statistically significant.

The same hierarchical regression analysis was conducted with each of the three grade levels to see if the relative weights of the domain scores differed by grade. The results of these analyses are presented in Tables 10-12. Similar to the aggregated results, the model that includes fall RIT, various demographics, and the four domain scores was significant in each grade. In each case, fall RIT explains most of the variance, followed by the addition of domain scores, and finally one or more demographic variables. SPED status was the only demographic variable that contributed significantly to the model in fourth and fifth grades (4^{th} grade: Adjusted R^2 = .70; $F(6, 706) = 280.71$, $p < .001$; 5^{th} grade: Adjusted R^2 = .65; $F(6, 442) = 136.69$, $p < .001$). In third grade, SPED status and the language program the student attended also added significantly to the model (Adjusted R^2 = .67; $F(7, 705) = 211.58$, $p < .001$).

Table 9

Variance in Dependent Variables

		B	*SE B*	*β*
Step 1				
	Constant	197.74	0.26	
	Fall RIT	0.75	0.02	.74*
Step 2				
	Constant	198.25	0.27	
	Fall RIT	0.7	0.02	.69*
	SPED	-5.12	0.72	-.12*
Step 3				
	Constant	196.07	0.28	
	Fall RIT	0.42	0.02	.41*
	SPED	-2.43	0.70	-.06*
	Reading	0.09	0.01	.16*
	Writing	0.12	0.01	.19*
	Speaking	0.04	0.01	.08*
	Listening	0.02	0.01	.03

Note. R^2 = .54 for Step 1: ΔR^2 = .01 for Step 2: ΔR^2 = .06 for Step 3 (*p*s < .001). * *p* < .001.

In terms of the unique variance explained by each of the *ACCESS for ELLs* domain scores, some differences exist between grade levels as to which scores contributed to the model, and the relative strengths of those contributions. In 3rd grade, Reading significantly contributed to the model and had a semipartial correlation coefficient (*sr*) of .11, meaning approximately 1% of the unique variance was explained by the addition of this variable to the regression equation. In 4th grade, Reading did not significantly contribute to the model (*p* = .110) after accounting for fall RIT, SPED

status, and the other domain scores. In 5th grade, Reading contributed the most of the four domain scores with a semipartial correlation of .19, representing approximately 4% of the unique variance explained. Writing was a significant contributor in each grade level, accounting for approximately 1% of the unique variance explained. Speaking significantly contributed in 3rd and 4th grades with less than 1% variance explained, but was not significant in 5th grade (p = .126). Listening was significant only in 4th grade and contributed less than 1% unique variance explained in spring RIT.

Table 10

Variance in Dependent Variable by Grade Level (3rd Grade n = 713)

		B	*SE B*	*β*	*sr*
Step 1					
	Constant	192.42	0.37		
	Fall RIT	0.78	0.02	.79	.79**
Step 2					
	Constant	192.91	0.4		
	Fall RIT	0.74	0.02	.75	.71**
	SPED	-5.44	1.11	-.12	-.11**
Step 3					
	Constant	196.28	0.95		
	Fall RIT	0.75	0.02	.76	.71**
	SPED	-5.17	1.1	-.11	-.11**
	Language Program	-2.59	0.67	-.09	-.09**
Step 4					
	Constant	194.55	0.91		
	Fall RIT	0.51	0.03	.76	.32**
	SPED	-2.95	1.09	-.06	-.06*

		B	*SE B*	*β*	*sr*
	Language				
	Program	-2.48	.63	-.09	-.08**
	Reading	0.1	0.02	.18	.11**
	Writing	0.08	0.02	.14	.08**
	Speaking	0.04	0.01	.08	.07*
	Listening	0	0.02	.00	.00

Note. R^2 = .62 for Step 1: ΔR^2 = .01 for Step 2: ΔR^2 = .01 for Step 3: ΔR^2 = .05 (*p*s < .001). * $p < .01$, ** $p < .001$.

Table 11

Variance in Dependent Variable by Grade Level (4th Grade n = 712)

		B	*SE B*	*β*	*sr*
Step 1					
	Constant	199.36	0.36		
	Fall RIT	0.80	0.02	.80	.80**
Step 2					
	Constant	200.03	0.36		
	Fall RIT	0.73	0.02	.73	.66**
	SPED	-7.57	1.03	-.18	-.16**
Step 3					
	Constant	198.47	0.38		
	Fall RIT	0.51	0.03	.51	.31**
	SPED	-5.38	1.04	-.13	-.11**
	Reading	0.04	0.02	.07	.04*
	Writing	0.09	0.02	.16	.10**
	Speaking	0.03	0.01	.08	.07**
	Listening	0.03	0.01	.07	.05*

Note. R^2 = .64 for Step 1: ΔR^2 = .03 for Step 2: ΔR^2 = .04 for Step 3 (*p*s < .001). * $p < .05$, ** $p < .001$.

Table 12

Variance in Dependent Variable by Grade Level (5th Grade n = 449)

		B	*SE B*	*β*	*sr*
Step 1					
	Constant	204.09	0.53		
	Fall RIT	0.72	0.03	.74	.74**
Step 2					
	Constant	204.73	0.53		
	Fall RIT	.67.03	0.69	.69	.64**
	SPED	-5.13	1.16	-.15	-.14**
Step 3					
	Constant	200.95	0.62		
	Fall RIT	0.31	0.05	.32	.19**
	SPED	-1.96	1.11	-.06	-.05
	Reading	0.18	0.03	.32	.19**
	Writing	0.09	0.03	.15	.10**
	Speaking	0.02	0.02	.05	.04
	Listening	0.03	0.02	.05	.04

Note. R^2 = .55 for Step 1: ΔR^2 = .01 for Step 2: ΔR^2 = .08 for Step 3 (*p*s < .001).

The previous results controlled for initial status (fall RIT) and various demographic considerations. The study of the ACCESS for ELLs and the New England Common Assessment Program (NECAP) conducted by Parker and colleagues (2009) does not appear to have done so. For comparison purposes, a multiple regression analysis consisting of only the four ACCESS for ELLs domain scores and spring reading RIT on the MAP was conducted. This model was statistically significant, Adjusted R^2 = .55; $F(4, 1869) = 563.19$, $p < .001$. In this view, the four domain scores account for 55% of the variance in spring RIT. Semipartial correlation coefficients (*sr*) for each of the domain scores were as follows: Reading = .19, Writing = .27, Speaking = .10, and Listening =

.05. Table 13 shows the unstandardized regression coefficients (*B*), the standard error of *B*, and the standardized regression coefficients (β) for the uncontrolled model.

Table 13

Variance in Dependent Variables in an Uncontrolled Model

	B	*SE B*	*β*
Constant	193.01	0.24	
Reading	0.17	0.01	.30**
Writing	0.23	0.01	.38**
Speaking	0.06	0.01	.12**
Listening	0.04	0.01	.07*

** $p < .001$ * $p < .01$.

Question 3: English Language Proficiency and Growth

The previous analyses focused on the relationship between the *ACCESS for ELLs* and student *achievement* on the *MAP*. In order to explore how English language proficiency is related to fall-to-spring *growth* on the *MAP*, the sample was categorized according to their overall proficiency level and differences in mean gain (spring RIT – fall RIT) were compared. The six proficiency levels represent an increase in English proficiency as students move through the following stages of development: *Entering*, *Emerging*, *Developing*, *Expanding*, *Bridging,* and *Reaching*. Table 14 shows the means, standard deviations, and norm-referenced percentiles for both fall and spring RIT, as well as the mean gain for each proficiency level. The *MAP* does not provide percentiles for within-year gain for reasons that will be explained below. For reference, results for the

district's non-ELLs were also included. Figure 1 displays the means plot for fall-to-spring gain by proficiency level.

For each proficiency level, mean RIT increased from the fall to the spring testing season. The gain, represented by change in RIT score from fall to spring, varied by proficiency level with students at the *Entering* level of English proficiency growing on average approximately 2 RIT points between testing seasons. The *Developing* students averaged approximately 12 RIT points over that same time period. The remaining proficiency levels gained between 8 and 11 RIT points over the year. A Welch ANOVA, which accounts for unequal group variances, was conducted and determined that the six proficiency levels differed significantly by fall-to-spring RIT gain, $F(5, 129.11) = 13.28$, $p < .001$. A trend analysis indicated that the data were better fit by a quadratic trend, $F(1, 2000) = 18.95$, $p < .001$, than a linear trend, $F(1, 2000) = 4.87$, $p = .027$.

Question 4: Conditional Growth on the *MAP*

The primary limitation of using a basic gain score as a measure of student growth is that it fails to consider the student's initial achievement status. A student's fall RIT is related to the rate of change in scores from one testing season to the next (NWEA Growth Norms, 2011). For this reason, percentiles are not available for a gain score, because gain is relative to starting achievement level. As such, the conditional growth norms are used to compute a conditional growth index (CGI), which measures the normative gain made by a typical student in a similar grade, with a similar starting RIT, and roughly the same instructional weeks between testing sessions. The CGI is reported in standard deviation units. To demonstrate that this was also the case for this study, the correlation between fall RIT and within-year gain was $r(2{,}006) = -.40$, $p < .001$, which is a medium

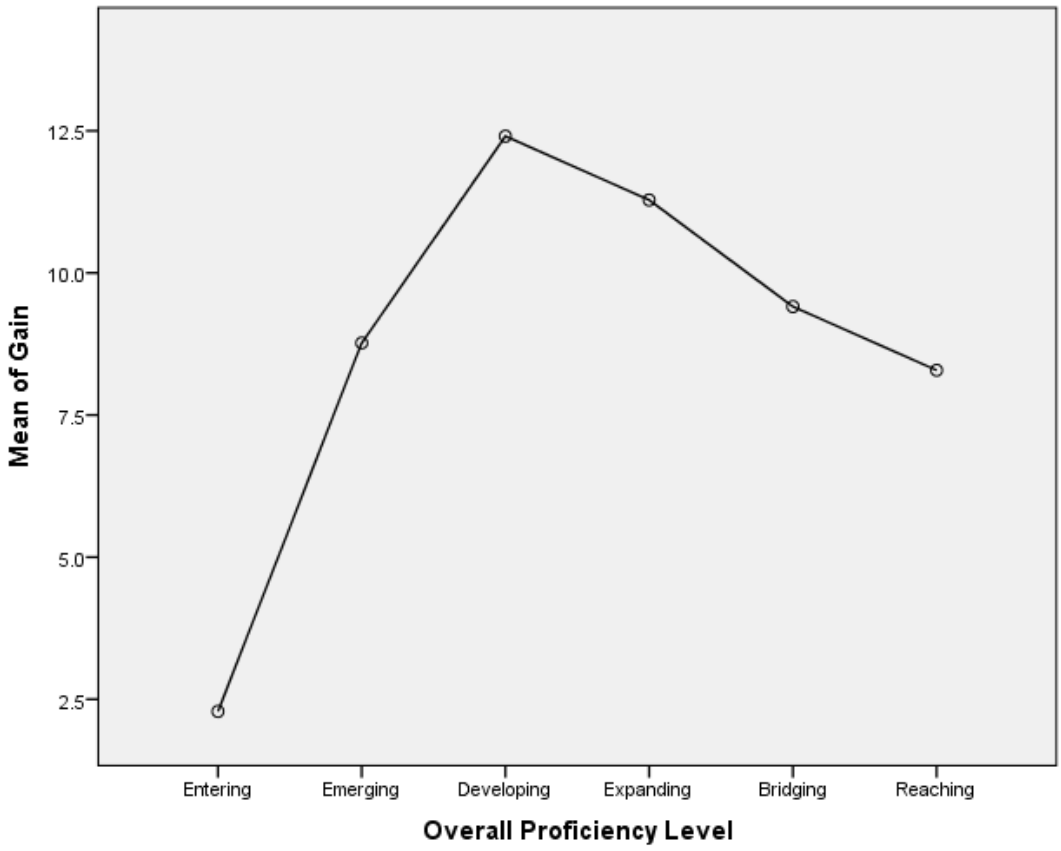
Table 14

Means, Standard Deviations, and Reading RIT Gain by English Proficiency Level

		Fall			Spring			Gain	
Variable	*n*	*M*	*SD*	Median Percentile	*M*	*SD*	Median Percentile	*M*	*SD*
Entering	14	152.93	6.32	1	155.21	6.78	1	2.29	5.89
Emerging	86	154.48	10.84	1	163.24	12.18	1	8.77	9.49
Developing	382	169.04	12.39	3	181.45	12.54	5	12.41	11.19
Expanding	660	182.38	12.06	16	193.66	10.85	21	11.28	10.24
Bridging	544	192.14	10.26	32	201.55	9.48	39	9.41	9.21
Reaching	320	200.88	8.86	50	209.16	7.59	56	8.29	8.83
ELL	2006	184.03	16.56	20	194.37	15.55	25	10.34	10.01
Non-ELL	12692	190.49	17.91	32	198.48	16.72	35	7.99	9.43

Figure 1

Mean Reading RIT Gain by Overall English Proficiency Level



to large effect (Cohen, 1988). This means that students with higher fall RIT scores are more likely to demonstrate lower RIT gain from fall to spring. Table 15 presents the means and standard deviations of CGI for each of the six proficiency levels. The mean for Non-ELL students has been included for reference.

A similar pattern as shown in gain scores by proficiency level emerges; however, the use of the CGI provides more confidence that the observed growth is not just a function of initial status. Starting with the lowest proficiency level, Entering, the average CGI was -1.27. This means that students at this proficiency level demonstrated growth on average 1.27 standard deviations below students with similar starting RIT from the

Table 15

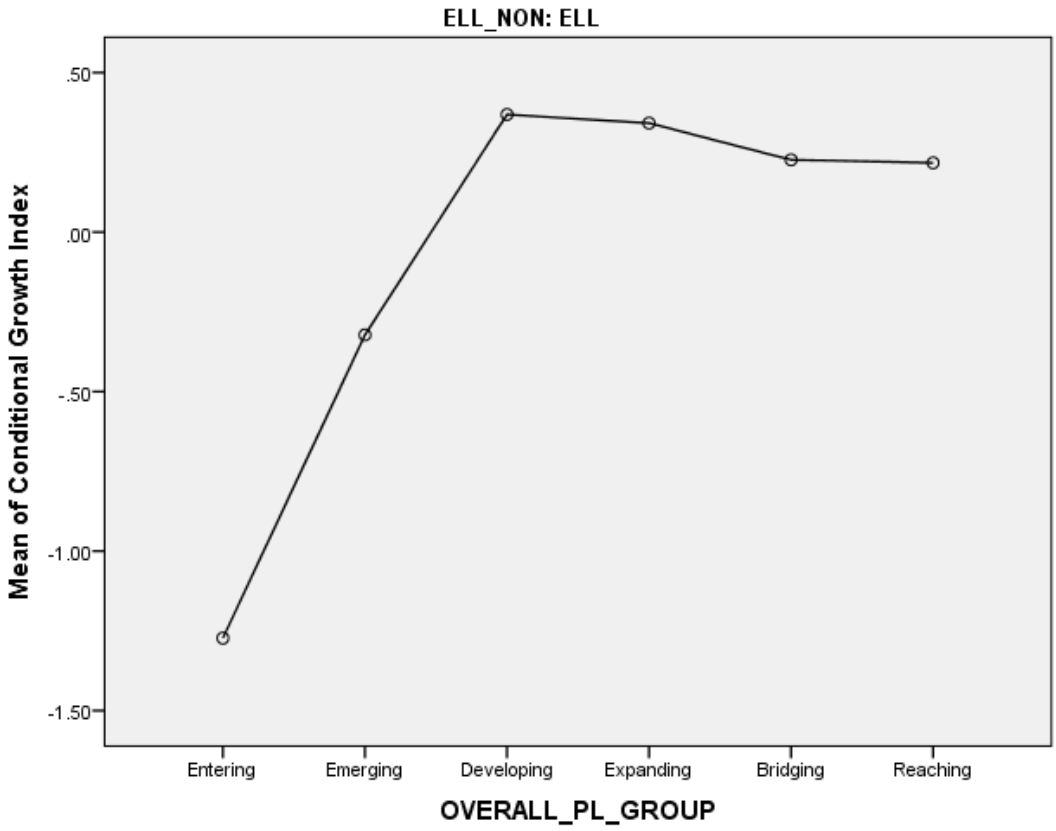
Means, Standard Deviations, and Percentiles for CGI by English Proficiency Level (n = 2,006)

Variable	*n*	*M*	*SD*	Median Percentile
Entering	14	-1.27	0.94	9
Emerging	86	-0.32	1.37	40
Developing	382	0.37	1.60	60
Expanding	660	0.34	1.42	60
Bridging	544	0.23	1.29	60
Reaching	320	0.22	1.22	58
ELL	2006	0.26	1.40	54
Non-ELL	12692	-0.01	1.32	50

nationally representative mean growth of the *MAP* norming sample. One sample t-tests determined that the mean CGI for each level of proficiency was different from zero. A Welch ANOVA was conducted and showed a statistically significant difference in CGI between different levels of English proficiency, $F(5, 127.77) = 10.26$, $p < .001$. A trend analysis indicated that, unlike the gain score, the CGI data were not better fit by a quadratic trend, $F(1, 2000) = 17.123$, $p < .001$, than a linear trend, $F(1, 2000) = 21.38$, $p < .001$. Figure 2 displays the mean plots for CGI by each level of proficiency.

Figure 2

Mean Conditional Growth Index by Overall English Proficiency Level



Post hoc Games-Howell tests from the previous ANOVA indicated that the two lowest levels of English proficiency, *Entering* and *Emerging*, were significantly lower than the remaining four levels ($p < .05$ for all comparisons). The effect size, *d*, for comparisons involving the *Entering* level of proficiency ranged from 1.03 to 1.22. These are much larger than typical effect sizes (Cohen, 1988). For *Emerging* level students, d = .42 to .47. These are considered small to medium effect sizes (Cohen, 1988). The top four proficiency levels were not significantly different from one another.

Using the information that the lowest two proficiency levels differed from the rest of the groups in terms of CGI, and applying the state reclassification decisions using the

overall proficiency level of 6 or the *Five and 5* rule (PL5 in both overall composite and literacy composite), three general language ability groups were created: *Beginning*, *Intermediate*, and *Proficient*. The means, standard deviations, and percentiles for CGI for these three groups is presented in Table 16. *Beginning* English speakers demonstrated an average growth of -.46 standard deviations compared to the average gain by students with similar starting RIT and grade from the *MAP* norming sample. *Intermediate* and *Proficient* speakers were .33 and .21 standard deviations above the mean respectively. The mean CGI for both the *Intermediate* and *Proficient* students was above the district non-ELL average. A Welch's ANOVA found that the three proficiency groups differed on CGI, $F(2, 269.61) = 15.73$, $p < .001$. Planned contrasts revealed that *Beginning* English speakers demonstrated lower CGI scores than *Intermediate* or *Proficient* speakers ($d = .54$ for both comparisons, $p < .001$). *Intermediate* speakers did not differ from *Proficient* speakers ($p = .064$).

Table 16

Means and Standard Deviations for Conditional Growth Index and Percentile by Language Proficiency Group (n = 2,006)

		Conditional Growth Index		Conditional Growth Percentile	
Variable	*n*	*M*	*SD*	*M*	*SD*
Beginning	100	-0.46	1.36	37.81	32.47
Intermediate	1,307	0.33	1.46	57.10	32.95
Proficient	599	0.21	1.23	53.74	31.23
Non-ELL	12,692	-0.01	1.32	49.02	31.76

The preceding information attempted to address the three main components related to English language proficiency and academic achievement. The first component had to do with the integrity of the assessments in question and whether they performed the same for this sample population as they were described in their respective technical reports. The second component dealt with the validation of the use of the overall and literacy composite scores as decision-making tools for reclassification purposes. The final component explored the variation in English proficiency group profiles with regards to reading achievement and growth.

For Research Question 1, MAP fall, winter, and spring scores demonstrated correlations of r = .70 to .83. Strongest correlations were between adjacent testing seasons, and 3rd grade correlations were consistently stronger than 4th or 5th. Intercorrelations of ACCESS for ELLs domain scores ranged from r = .39 to .71. The strongest relationship was between the reading and listening domains, decreasing in strength from 3rd to 5th grade. The speaking domain had the lowest correlations of r = .43, .45, and .47 for writing, reading, and listening respectively. When considering ACCESS for ELLs domains scores and MAP reading RIT scores, the strongest correlations were for reading (r = .65 to .69) and writing (r = .67 to .71) to each of the MAP administrations. Listening scores demonstrated correlations of r = .55 to .58 and speaking scores ranged from r = .47 to .50.

Because decisions about qualification and placement of ELLs is based on weighted composite scores, Research Question 2 explored the weights of each language domain score on spring reading score on the MAP. A multiple regression model that

accounted for initial status (fall RIT) and various demographic variables accounted for 61% of the variance in spring RIT on the MAP. After controlling for fall RIT and demographics, ACCESS for ELLs domain scores accounted for an R^2 change of .04, .05, and .08 for grades 3-5 respectively. Standardized regression coefficients were larger for writing and reading (.16 and .19) than speaking and listening (.08 and .03). Special education eligibility also contributed significantly to the model. In an uncontrolled model, standardized regression coefficients reading, writing, speaking, and listening were .30, .38, .12, and .07 respectively.

Research Questions 3 and 4 looked beyond status level achievement, focusing instead on within-year growth. Question 3 looked specifically at a basic gain score consisting of the difference between spring RIT and fall RIT. Question 4 employed a conditional growth index that compares an individual student's RIT gain in the context of the MAP growth norms which takes into account initial status (fall RIT), grade, and number of instructional weeks between administrations. This process provides a standardized measure of student growth. In both cases, significant differences in within-year growth on the MAP were found for different levels of English proficiency. Whereas each level of increased proficiency demonstrated higher mean fall and spring RIT than the previous, the growth rates did not follow the same pattern. The two lowest levels of proficiency demonstrated significantly lower growth, 1.25 and .32 standard deviations below the mean for the norming sample, compared to the highest four levels of proficiency (.37, .34, .23, and .22 standard deviations above the mean).

CHAPTER V

DISCUSSION

For the native English speaker, the entire school year is dedicated to learning the adopted curriculum. While many native English-speaking students will succeed in this endeavor, other non-native English speakers may struggle to keep up. English language learners (ELL) have the dual task of learning the same content as their native English peers in the same time frame, while also mastering the complexities of a new language. For some ELLs, English may not just be a second language, but a third or more. From a readiness for English instruction standpoint, English language proficiency assessments have the complex task of identifying when students have enough ability in English to meaningfully participate in the English-only classroom, and discerning when their performance is on par with their native English counterparts. Academic achievement and growth measures designed to assess student understanding of a given subject for fluent English speakers may not accurately reflect the understanding of ELLs due to their limited language abilities. The current study explored the relationship between two widely uses measures: the *ACCESS for ELLs* for English language proficiency, and the reading subtest of the *MAP* for academic achievement and growth.

Relationship Between *ACCESS for ELLs* and *MAP*

The first research question attempted to establish the link between these two assessments. If the assessment of English language proficiency is to be used to measure of readiness for academic achievement, it must relate to the academic achievement assessments used by all students. Without this link, a designation of ready for English instruction or English fluency is not useful if it does not also relate to higher academic

achievement. This was one of the shortfalls of pre-NCLB English proficiency assessments identified previously (Albers et al., 2009; Llosa, 2011).

The first step in establishing this link was to look at the consistency and reliability of ELL performance across testing seasons (fall, winter, and spring) on the *MAP*. The reported test-retest reliability for the state aligned version of the test used in this sample was r = .79 to .80 (NWEA, 2012). This strong relationship means that high scores in one season correlate to high scores in subsequent administrations. The ELL students in this sample demonstrated almost identical correlations between seasons, r = .78 to .83, providing evidence that the *MAP* behaves similarly over time for ELLs and native English speakers.

With regards to the *ACCESS for ELLs*, intercorrelations between the four language domains of reading, writing, speaking, and listening attempted to show both divergent and convergent connections between the various facets of English language proficiency. A very strong correlation between two domains would indicate that they are likely measuring similar constructs. Weak correlations means that a high score in one does not necessarily correspond to an equally high score in the other. This would suggest that each domain is measuring a slightly different aspect of English. According to the technical report for the *ACCESS for ELLs*, correlations between domain scores for 3rd through 5th grades range from r = .43 to .68, with stronger relationships between reading and writing, and weaker relationships between reading and speaking (WIDA, 2012). A strong link between reading and writing makes sense in that both require comprehension of written English, with reading being more of the receptive function, while writing is more expressive. The lower correlation between reading and speaking is likely because

these two facets differ not only in receptive and expressive language, but also in mode of communication.

The students in my study demonstrated similar domain associations, with all 3rd through 5th grade correlations ranging from r = .39 to .71. As in the technical report, reading was strongly correlated to writing (r = .59 to .66) across all grades. Reading was more strongly correlated to listening for this group of students than was reported in the technical report. Both reading and writing were less correlated to speaking, with a range of r = .39 to .45. Given the similarity in intercorrelations between the current study and those reported in the technical manual, support that the decisions based on performance on the *ACCESS for ELLs* by this group of students would be very similar to those presented by the assessment company. In essence, both the *ACCESS for ELLs* and the *MAP* behaved as expected with this sample population.

By exploring the direct correlation between the *ACCESS for ELL* domain and composite scores and the overall RIT on the *MAP*, we gain confidence in the use of one test (the *ACCESS for ELLs*) to make a reasonable estimation of performance on the other (the *MAP*). This is important when reclassification and program placement decisions are based almost exclusively on the results of the ELP assessment. Scores for the reading and writing domains demonstrated the strongest correlations to the *MAP* across all grades and testing seasons (r = .57 to .72). These large to very large effect sizes are meaningful in that the English language requirements identified by the *ACCESS for ELLs* for proficiency show commensurate gains on an achievement measure that is assessing the same construct of interest, namely comprehension of the written language. The lower correlations (r = .40 to .58) between the speaking and listening domains was not

surprising, given that the *MAP* does not assess speaking, nor does it provide audio on its items. The effect sizes associated with listening and speaking are still moderate to strong, suggesting that increased English fluency in these domains is associated with higher reading achievement scores.

In looking at the relationship between the weighted composite scores on the *ACCESS for ELLs* and the *MAP*, the overall composite and the literacy composite scores had the highest correlations, r = .72 to .77, across all testing seasons. These correlations were higher than any of their component domain scores, making them the best candidates for decision making purposes for this sample of students. These results support the decision to use the overall composite and literacy composite scores as the primary measure for reclassification to fluent English proficient, at least in terms of a reading achievement measure. Though not a direct contradiction to Abedi's (2008b) recommendation that reclassification decisions should be based on all domain scores, the lack of contribution of listening in some respects, and speaking to a greater extent, on this achievement measure supports the use of the overall or literacy composite score.

Because the composite scores are based on weighted contributions from the domain scores, the second research question addressed whether those weights were supported in this sample population. In their study comparing the *ACCESS for ELLs* to the New England Common Assessment Program (NECAP) and used by WIDA (2013) to support their weighted composite scores, Parker, Louis, and O'Dwyer (2009) found standardized regression coefficients of .03 to .38 for reading and .15 to .30 for writing and consistently high correlations between these two domains and the NECAP. As previously mentioned, reading and writing did correlate most strongly with the *MAP* in

the current study, though writing was stronger than reading in both 3rd and 4th grade. The current study calculated standardized regression coefficients for each of the domain scores, but unlike the study by Parker et al. (2009), this regression model controlled for initial RIT and student demographics. The standardized regression coefficients for reading and writing were .04 to .19 and .08 to .10 respectively ($p < .001$ for each). The speaking domain score contributed significantly in 3rd and 4th grades with a standardized regression coefficient of .07 in each, but did not contribute to the model in 5th grade. Listening did not contribute to the regression model in any of the three grades. A model that did not control for initial RIT or demographics, similar to the study by Parker et al. (2009) found standardized regression coefficients that ranged from .07 to .38. These results are consistent with those from the other study.

On the surface, the results from my study seem to mirror those found by Parker et al. (2009), however, by controlling for initial status and demographics, reading and writing each explain only 1% of the variance in spring RIT on the *MAP*. While these domain scores are statistically related to the *MAP* and contribute more than the speaking and listening domains, the practical significance may be small. In fact, of the approximately 70% variance explained by the full regression model, less than 5% could be attributed to the domain scores on the *ACCESS for ELLs* after accounting for controlling factors. A low percentage of variance explained by the *ACCESS for ELLs* may not be bad. If it were too high, then the *MAP* would be more of a language test than one of reading achievement. Although the relative weights between reading and writing differ slightly by grade level, the equal weighting in the literacy composite and equal but

greater weighting in the overall composite seems to be supported for use with this sample population.

One possible misinterpretation of these results would be to suggest that speaking and listening are not important components to English proficiency or academic achievement. On the contrary, both listening and speaking contribute significantly to the overall model, albeit in lesser amounts than reading and writing. A decision to focus instruction on reading and writing at the expense of speaking and listening could have detrimental effects on the classification of ELLs. The compensatory approach to a weighted composite score could result in increased proficiency rates with lower scores in two of the four language domains. This realization does support Abedi's (2008b) concern that a composite score might not capture all the important nuances of English language proficiency if instructional decisions such as the one mentioned here were put into place.

Demographic Control

One interesting result of using a stepwise regression to control for student demographics was the difficulty of disentangling those variables. The study included student level information including gender, race, free and reduced lunch eligibility, IEP/SPED status, home language and first language. There was also a school level demographic identifying which language program was employed at the enrolled school. Though tolerance and variance inflation factors suggested no multicollinearity, some variables were inextricably related. One clear example is how the relationship between race and free and reduced lunch eligibility, impacts home language, first language, and school language program. A chi-square analysis indicated group differences between

race and free and reduced lunch eligibility (χ^2 = 14.17, *df* = 4, *N* = 2,006, *p* = .007), with a greater proportion of African American and Hispanic students eligible compared to Asian or White students. Race was directly related to both home and first language. Hispanic ELL students were more likely to speak Spanish as their first (and probably home) language. Because the district predominantly offers bilingual education as a language program in Spanish, Hispanic students are more likely to be enrolled in those programs. For this sample, Hispanic students accounted for 99% of ELL students enrolled in bilingual programs. Therefore, by association, FRL status is linked to all three language variables via its association with race for this sample. As a result, first language was removed from the stepwise regression. After controlling for initial status, only IEP/SPED status for all three grades and language program for 3rd grade significantly contributed to R^2 change.

Growth and English Language Proficiency

The third and fourth research questions moved beyond status level achievement and focused instead on within-year growth. Given the current trend in accountability frameworks, 30 states and D.C. have state legislation requiring the use of growth as a component for evaluation (Collins & Amrein-Beardsley, 2013). Despite considerable research around the achievement gap between ELLs and non-ELLs and their changes over time, there is a paucity of literature that looks specifically at within-year growth on standardized achievement measures by ELLs. My study attempted to fill that gap by comparing the mean fall-to-spring RIT gain and conditional growth index (CGI) on the *MAP* for each of the six levels of language proficiency as measured by the *ACCESS for ELLs*.

A gain score is a simplified measure of growth that subtracts the fall RIT from the spring RIT. The difference in RIT is conceptualized as growth in reading achievement over the course of the school year. One of the primary limitations of the use of a gain score as an indicator of growth is that is fails to account for a student's initial status, in this case the fall RIT score. According to NWEA's published Conditional Growth Norms, a student's fall RIT was related to the rate of change in scores from one test administration to the next. In the case of this study, the correlation between fall RIT and fall-to-spring growth was $r(2{,}006) = -.40$, $p < .001$, which is a medium to large effect (Cohen, 1988). This means students with higher fall RIT are more likely to experience less RIT gain over the course of the year. To account for this limitation, the Conditional Growth Norms provided a means of calculating a conditional growth index that accounts for an individual's grade, initial status, and educational opportunity. The CGI is reported as a standard deviation unit and represents the difference between the observed score and mean growth of students in the same grade with similar fall RIT and the same number of instructional weeks between administrations. Although Table 1 and 2 appear to present different results (observed gain compared to conditional growth), those differences are not functionally different, but rather a function of the RIT scale. For this reasons, gain scores should be interpreted with caution.

The mean fall and spring RIT for each level of proficiency were stratified, such that the lowest level of proficiency, *Entering*, had the lowest fall and spring RIT. The next level, *Emerging*, was higher in both fall and spring RIT, but lower than proficiency level three students. This pattern continued through the highest proficiency level of *Reaching*. The mean score for the six proficiency levels represented a median percentile

of 1, 1, 3, 16, 32, and 50 for the fall respectively. Spring median percentiles in order of proficiency were 1, 1, 5, 21, 39, and 56. The lowest proficiency group gained on average 2.29 RIT between administrations. Subsequent mean gain for increasing levels of English proficiency was 8.77, 12.41, 11.28, 9.41, and 8.29 RIT respectively.

The results of an adjusted Welch ANOVA that accounted for unequal variances in groups showed that the average RIT gain by proficiency level was significantly different, $F(5, 129.11) = 13.28$, $p < .001$. A trend analysis indicated that the data were best fit by a quadratic trend, $F(1, 2000) = 18.95$, $p < .001$, suggesting curvilinear growth exemplified by low RIT gain for low English proficient students, followed by a peak of high RIT gain for middle proficiency students, and tapering off for high proficiency students.

The mean gain score for the lowest proficiency group translated to a CGI of -1.25. Practically speaking, this means that the average RIT gain for an *Entering* level proficient student is 1.25 standard deviations below the mean of norming sample students with the same fall RIT and number of weeks of instruction between administrations. The CGI for the remaining five proficiency levels was -.32, .37, .34, .23, and .22 respectively. The median conditional growth percentiles for the six proficiency levels were 9, 40, 60, 60, 60, and 58. The average fall and spring RIT for the four lowest proficiency levels were below the district average in both testing seasons, while the top two proficiency levels we above. In terms of CGI, only the bottom two proficiency levels were below the district mean CGI. This suggests that although the level 3 and 4 proficiency students (*Developing* and *Expanding*) started and ended the year below the district mean RIT, they narrowed the gap on average. The top two groups, *Bridging* and *Reaching*, started and finished the year above the district mean RIT and grew more than the district average.

A Welch's ANOVA, $F(5, 127.77) = 10.26$, $p < .001$, and post hoc Games-Howell tests ($p < .05$) comparing the CGI for each level of proficiency confirmed that the two lowest proficiency levels were significantly different from the other four. The four highest proficiency groups demonstrated no significant difference. A trend analysis indicated that the CGI data were not better fit by a quadratic trend, $F(1, 2000) = 17.123$, $p < .001$, than a linear trend, $F(1, 2000) = 21.38$, $p < .001$. These results are somewhat similar to the differential learning rates explored by Chen (2010). Chen (2010) found that students with different levels of math achievement differed in language proficiency, and that higher levels of math reduced the influence of language on math scores. The present study explored the first half of her analysis, but without more sophisticated statistical models was unable to partition the specific influences of English proficiency and content understanding.

Using the results from the Games-Howell tests and the state reclassification criteria, *Beginning*, *Intermediate*, and *Proficient* language ability groups were created. A Welch's ANOVA found that the three proficiency groups significantly differed on CGI, $F(2, 269.61) = 15.73$, $p < .001$. Planned contrasts revealed that Beginning English speakers demonstrated lower CGI scores than *Intermediate* or *Proficient* speakers ($d = .54$ for both comparisons, $p < .001$). *Intermediate* speakers did not significantly differ from *Proficient* speakers ($p = .064$).

Implications

There are a several important takeaways from this exploration of reading achievement and growth. For starters, for this robust sample of ELLs, it was clear that the level of English language proficiency and not just ELL status is an important

consideration. Each proficiency level differed by fall RIT, spring RIT, and growth/CGI, but the differences in achievement levels did not match the profile of differences in growth. That is to say that although *Beginning* English proficiency was related to low RIT *and* low growth, that was not the case for *Intermediate* and *Proficient* English speakers. In the case of the higher proficiency groups, *Intermediate* speakers demonstrated lower initial and ending achievement, but slightly higher growth than their *Proficient* peers. Additionally, both of the higher proficiency groups beat the district mean conditional growth index. Practically speaking, this means that the two lowest proficiency groups lost ground compared to their English speaking peers while the four highest levels of proficiency demonstrated a narrowing of the achievement gap.

From an instructional standpoint, these differences in achievement and growth profiles could help teachers better plan for and address the learning needs of ELLs. A teacher could combine conditional growth norms and instructional objectives that would occur at the target RIT. If a student's initial RIT and the projected growth for a student is known, objectives in that growth target range that will likely match the instructional needs of that student can be found. Knowing now that the English proficiency level of students might impact projected growth, teachers could adjust their expectations, and fine tune their approach in an attempt to better differentiate instruction for their ELLs. Failure to account for these differences in proficiency levels could have negative consequences in terms of instructional planning. For *Beginning* English students, a teacher would likely have overestimated the growth targets and identify learning objectives that were still too difficult for the student to grasp. It was potentially worse for *Intermediate* or *Proficient* students who demonstrated above average growth. In this case, a teacher would likely

have underestimated their growth target and provided instructional objectives that were not challenging enough. In the interest of closing the achievement gap, not accounting for this differential growth could be a missed opportunity.

This study also helped to contextualize student growth by comparing a basic gain score to a standardized measure of growth. Although the use of a gain score demonstrated differences by proficiency level, it did not explain how that gain compared to other students with similar scores. A mean gain of 2.29 RIT for the students in the lowest proficiency group was less than 6 points off the growth of non-ELLs for this district. The greatest RIT gain with a mean of 12.41 by the third proficiency group was just over 4 points higher than non-ELLs. The conditional growth index (CGI) showed that the low gain was actually 1.25 standard deviations below the mean for growth while the high gain was .37 standard deviations above the mean. Without this standardized view, it is difficult to determine what a difference of 4 or 6 points really means. The CGI also helps to account for characteristics of the RIT scale.

From an accountability standpoint, these differences in growth by proficiency level could impact program or teacher evaluations in important ways. For example, if a school wants to evaluate the effectiveness of a new reading or language development program used with all of its ELLs, results may depend on the proficiency levels of the students involved. If the school uses student growth data from the *MAP* to measure effectiveness, then student growth scores from *Beginning* English speakers compared to the published growth norms could have underestimated the effectiveness of the program. The opposite would have been true if comparing growth from *Intermediate* or *Proficient* students. The above-average growth demonstrated in this study could make a program

look more effective than it was. In either case, significant cost in terms of purchasing new curricular materials and the time invested in professional development would be wasted, especially if the program being evaluated were actually effective.

In terms of teacher evaluation, classroom composition both in the number of ELLs and differing levels of proficiency can vary widely within a district, school, grade, or even classroom. If teachers are evaluated on the basis of student achievement and growth scores, such as is the case with value-added models (Chudowsky, Koenig, & Braun, 2010), then a teacher effectiveness rating that only accounts for ELL status and not the specific level of language proficiency may provide an incorrect and inappropriate estimate. Teachers who have large proportions of *Intermediate* English speakers, a situation not uncommon in urban classrooms, may receive an artificial boost to their effectiveness estimate because those students typically demonstrate above-average growth. Conversely, a teacher with many *Beginning* speakers could receive a negative estimate of effectiveness even though the performance of those ELLs met expectations given the proficiency level.

Based on the results from this study, there appears to be a policy issue related to how English language learners are labeled. The current dichotomous system that differentiates English proficient students from those who are not is likely a holdover from accountability systems more focused on achievement than growth. As noted previously, the students in this study with the highest levels of English proficiency demonstrated achievement rates at or above the 50^{th} percentile. In this respect, the dichotomous label seems to meaningfully separate two groups of English speakers. The same cannot be said when considering within-year growth. Classified ELL students with intermediate levels

of English proficiency demonstrated growth more similar to their reclassified English proficient peers than current ELLs of lower proficiency. In this case, a dichotomous label fails to account for the interaction between language status and learning status including variation within the classified ELL category. Where the distinction between groups is more apparent in terms of status achievement, the relationship between ELL classification and growth is cloudy at best. A more appropriate method of labeling could be based on achievement and growth, or even more simply, the status categories of beginning, intermediate, and proficient.

Limitations

These initial findings warrant further exploration into the role of English language proficiency in academic achievement. Although the sample came from a large population with ample ELLs, specific programmatic, instructional, or demographic differences could render them substantially different from neighboring districts or distant states, thus limiting the generalizability of this study. Additionally, this sample represents one cohort of students over one academic year assessed with one English language proficiency assessment and one reading interim assessment. Additional grades, extended periods of observation, or other measures could shed light into how growth changes by age of student and/or over the language acquisition process.

Another limitation of this study deals with the level of specificity of the data. As an outside researcher limited to quantitative extant data, details regarding instructional alignment to English language development standards or academic content standards were not available. Also, the implementation fidelity of the instruction occurring in schools or classrooms could not be evaluated.

Future Research

This study confirmed the use of weighted ACCESS for ELLs composite scores for decision-making purposes when using MAP reading RIT. An important follow-up study would be to confirm if a similar relationship exists between ACCESS for ELLs domain scores and other achievement measures such as the other subtests of the *MAP* including language usage, math, or science. Although decisions based heavily on the reading and writing domains with regards to reading achievement makes sense from a content standpoint, similar decisions may not necessarily be appropriate for other subjects.

Though this study focused on one measure of English language proficiency, it would also be worthwhile to explore results with alternative measures of ELP such as the state-specific Washington Language Proficiency Test (WLPT), the English Language Proficiency Assessment (ELPA), or the upcoming English Language Proficiency Assessment for the 21st Century (ELPA 21) that is currently being developed by a 10-state consortium. Depending on how ELP is defined, different measures might produce different evaluations of a student's level of English proficiency. For example, does the increased correspondence to college and career readiness standards as one of the Looking at achievement and growth rates of other ELP measures could corroborate the findings from the present study.

Based on the limited control of program offerings or demographic variables available, more specific research plans could attempt to better control for those differences and begin to work towards a causal explanation linking English proficiency, language programs, and student-level demographic variables to academic achievement

and/or growth. One piece of student information that was not considered in this study was the date of entry to the United States compared to the student's level of English proficiency. It would be interesting to see how students with similar levels of proficiency, but different amounts of time in the country compare in terms of achievement and growth. Similarly, a more detailed treatment of language program (bilingual versus traditional ESL) could provide important information regarding the effects of bilingualism not just in terms of academic achievement, but also in growth rates. Given that a majority of students in this study attended bilingual programs, would similar results exist if all students were in traditional ESL programs?

Conclusion

In light of the increased use of standardized achievement measures with English language learners, this study explored the performance of a group of 3rd-5th grade ELLs on two widely used measures. Correlations for the three MAP administrations were consistent with those reported in the tech manual, suggesting the test performed for these ELLs similarly to the norming sample. Intercorrelations of domain scores on the ACCESS for ELLs was similar for this sample as reported in the tech manual, providing generalizability of this study. Correlation and standardized regression coefficients between ACCESS for ELLs domain scores and MAP reading RIT were similar to other studies using other measures of reading achievement. With more variance explained by reading and writing scores, this study provides support for the use of the weighted composite scores for decision-making purposes.

These results also suggest that the current policy of labeling ELLs into dichotomous categories (proficient versus not proficient) fails to capture important

differences in student performance at differing levels of English proficiency. Within-year growth is not consistent with the current ELL classification system. Low proficiency students demonstrated low achievement and below average growth. Intermediate speakers demonstrated lower than average achievement, but higher than average growth, and proficient speakers demonstrated higher than average achievement and growth. This study revealed that intermediate speakers perform more like lower proficiency students in terms of achievement, but more like proficient speakers when considering growth. A labeling system that does not account for the variation within the ELL group is unlikely to provide the precision of knowledge necessary to support the overall goal of closing the achievement gap.

REFERENCES CITED

Abedi, J. (2008a). Classification system for English language learners: Issues and recommendations. *Educational Measurement: Issues and Practice, 27*(3), 17-31.

Abedi, J. (2008b). Measuring students' level of English proficiency: Educational significance and assessment requirements. *Educational Assessment, 13*, 193-214. doi:10.1080/10627190802394404

Albers, C. A., Kenyon, D. M., & Boals, T. J. (2009). Measures for determining English language proficiency and the resulting implications for instructional provision and intervention. *Assessment for Effective Intervention, 34*, 74-85.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (APA, AERA, & NCME) (1999). *The Standards for Educational and Psychological Testing.* Washington, D.C.: American Educational Research Association.

Anstrom, K., DiCerbo, P., Butler, F., Katz, A., Millet, J., & Rivera, C. (2010). *A Review of the literature on academic English: Implications for K-12 English Language Learners*. Arlington, VA: The George Washington University Center for Equity and Excellence.

Ardasheva, Y., Tretter, T. R., & Kinny, M. (2012). English language learners and academic achievement: Revisiting the threshold hypothesis. *Language Learning, 62*(3), 769-812.

Aud, S., Hussar, W., Kena, G., Bianco, K., Frohlich, L., Kemp, J., & Tahan, K. (2011). *The condition of education 2011* (NCES 2011-033). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Aud, S., Wilkinson-Flicker, S., Kristapovich, P., Rathbun, A., Wang, X., & Zhang, J. (2013). *The condition of education 2013* (NCES 2013-037). U.S. Department of Education, National Center for Education Statistics. Washington, DC. Retrieved May 27, 2013 from http://nces.ed.gov/pubsearch.

August, D., & Shanahan, T. (2008). *Developing reading and writing in second-language learners: Lessons from the report of the National Literacy Panel on Language-Minority Children and Youth*. New York, NY: Taylor & Francis.

Bailey, A. L., Butler, F. A., & Sato, E. (2007). Standards-to-standards linkage under Title III: Exploring common language demands in ELD and science standards. *Applied Measurement In Education, 20*(1), 53-78.

Bailey, A. L., & Heritage, H. M. (2008). *Formative assessment for literacy, grades K-6: Building reading and academic language skills across the curriculum.* Thousand Oaks, CA: Corwin Press.

Bailey, A. L., & Kelly. (2010). The use and validity of home language surveys in state English language proficiency assessment systems: A review and issues perspective. CA: EVEA.

Baker, C. (2003). 6. Education as a site of language contact. *Annual Review of Applied Linguistics, 23*, 95-112.

Beck, I. L., & McKeown, M. G. (2007). Increasing young low-income children's oral vocabulary repertoire through rich and focused instruction. *The Elementary School Journal, 107* (3), 251-271.

Brooks, K., Adams, S. R., & Morita-Mullaney, T. (2010). Creating inclusive learning communities for ELL students: Transforming school principals' perspectives. *Theory Into Practice*, *49*(2), 145-151.

Bunch, M. B. (2011). Testing English language learners under No Child Left Behind. *Language Testing, 28*(3), 323-341.

Carlo, M. S., August, D., McLaughlin, B., Snow, C. E., Dressler, C., Lippman, D. N., Lively, T. J. & White, C. E. (2004). Closing the gap: Addressing the vocabulary needs of English-language learners in bilingual and mainstream classrooms. *Reading Research Quarterly, 39*, 188–215. doi: 10.1598/RRQ.39.2.3

Cawthon, S. (2004). How will No Child Left Behind improve student achievement? The necessity of classroom-based research in accountability reform. *Essays in Education, 11*, 11.

Chen, F. (2010). Differential language influence on math achievement (Doctoral dissertation). Retrieved from ProQuest. (3434131).

Chudowsky, N., Koenig, J., & Braun, H. (Eds.). (2010). *Getting value out of value-added: Report of a workshop*. National Academies Press.

Cohen, J. (1988). *Statistical power and analysis for the behavioral sciences (4th ed.)*. Thousand Oaks, CA: Sage.

Collins, C & Amrein-Beardsley, A. (2013). Putting growth and value-added models on the map: A national overview. *Teachers College Record, 116*(1), 1-32.

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.

Cook, H. G., & Zhao, Y. G. (2011). How English language proficiency assessments manifest growth. In *annual meeting of the American Educational Research Association, New Orleans, LA*.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of educational research, 49*(2), 222-251.

Cummins, J. (1980a). The entry and exit fallacy in bilingual education. *NABE: The Journal for the National Association for Bilingual Education, 4*(3), 25-59.

Cummins, J. (1980b). The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue. *TESOL Quarterly, 14*(2), 175-187.

Cummins, J. (2000). BICS and CALP. In M. Byram (Ed.), *Encyclopedia of language teaching and learning* (pp.76-79). London: Routledge.

Ferrara, S. (2008). Design and psychometric considerations for assessments of speaking proficiency: The English Language Development Assessment (ELDA) as illustration. *Educational Assessment, 13*, 132-169.

Grissom, J. B. (2004, July 30). Reclassification of English learners, *Education Policy Analysis Archives, 12*(36). Retrieved June 6, 2013 from http://epaa.asu.edu/epaa/v12n36/.

Hakuta, Kenji. (2000). *How long does it take English learners to attain proficiency*? (Report No. 2000-1) UC Berkeley: University of California Linguistic Minority Research Institute. Retrieved from: http://escholarship.org/uc/item/13w7m06g

Halle, T., Hair, E., Wandner, L., McNamara, M., & Chien, N. (2012). Predictors and outcomes of early versus later English language proficiency among English language learners. *Early Childhood Research Quarterly*, *27*(1), 1-20.

Hemphill, F. C., & Vanneman, A. (2011). *Achievement gaps: How Hispanic and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress*. Statistical Analysis Report. NCES 2011-459. National Center for Education Statistics.

Keppel, G. & Zedeck, S. (1989). Data analysis for research designs: Analysis of variance and multiple regression/correlation approaches. New York, NY: W. H. Freeman

Kieffer, M. J. (2008). Catching up or falling behind? Initial English proficiency, concentrated poverty, and the reading growth of language minority learners in the United States. *Journal of Educational Psychology*, *100*(4), 851-868. doi:10.1037/0022-0663.100.4.851.

Kim, J., & Herman, J. L. (2012). *Understanding patterns and precursors of ELL success subsequent to reclassification* (CRESST Report 818). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Lau v. Nichols, 414 U.S. 563 (1974).

Llosa, L. (2011). Standards-based classroom assessments of English proficiency: A review of issues, current developments, and future directions for research. *Language Testing, 28*(3), 367-382.

Luke, A. (2000). Critical literacy in Australia: A matter of context and standpoint. *Journal of Adolescent & Adult Literacy, 43*(5), 448-461.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Mohamud, A., & Fleck, D. (2010): Alignment of standards, assessment and instruction: Implications for English language learners in Ohio, *Theory Into Practice, 49*(2), 129-136.

Northwest Evaluation Association (NWEA). (2011). *Technical Manual for Measures of Academic Progress (MAP) and Measures of Academic Progress for Primary Grades (MPG).* Portland, OR: Northwest Evaluation Association.

Parker, C. E., Louie, J., and O'Dwyer, L. (2009). *New measures of English language proficiency and their relationship to performance on large-scale content assessments* (Issues & Answers Report, REL 2009-No. 066). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. Retrieved from http://ies.ed.gov/ncee/edlabs.

Reardon, S. F., & Galindo, C. (2009). The Hispanic-White achievement gap in math and reading in the elementary grades. *American Educational Research Journal, 46*(3), 853-891.

Ryan, C. (2013). Language use in the United States: 2011. *PDF). American Community Survey. US Census Bureau. Retrieved*, 08-11.

Scarcella, R. (2008). Academic language: Clarifying terms. *AccELLerate! the Quarterly Newsletter of the National Clearinghouse for English Language Acquisition (NCELA), 1*(1), 5-6.

Scott, J.A., Flinspach, S.L., Miller, T.F., Gage Serio, O., Vevea, J.L. (2009). An analysis of reclassified English learners, English learners, and native English fourth graders on assessments of receptive and productive vocabulary. In K.M. Leander, D.W. Rowe, D.K. Dickinson, M.K. Hundley, R.T. Jimenez, V.J. Risko (Eds.), *58th Yearbook of the National Reading Conference* (pp. 312-329). Oak Creek, WI: National Reading Conference.

Sireci, S. G., Han, K. T., & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment, 13*, 108-131.

Stevens, R. A., Butler, F. A., & Castellon-Wellington, M. (2000). *Academic language and content assessment: Measuring the progress of ELLs.* (No. CSE Tech. Rep. No. 552). Los Angeles: University of California: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Stewner-Manzanares, G. (1988). *The bilingual education act: Twenty years later*. National Clearinghouse for Bilingual Education.

Sullivan, A. L. (2011). Disproportionality in special education identification and placement of English language learners. *Exceptional Children*, *77*(3), 317-334.

Thum, Y.M. & Hauser, C. (2012). RIT scale norms: For use with Measures of Academic Progress (MAP®) and MAP® for Primary Grades. Portland, OR: Northwest Evaluation Association.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics (USDOE, IES, NCES). (2012). The Nation's Report Card. NCES 2012-457.Retrieved from http://nces.ed.gov/nationsreportcard/pdf/main2011/2012457.pdf

WIDA (2014). World-class instructional design and development. Wisconsin Center for Education Research at the School of Education, University of Wisconsin-Madison. Retrieved from http://www.wida.us/membership/states/

Wolf, M. K., Farnsworth, T., & Herman, J. (2008). Validity issues in assessing English language learners' language proficiency. *Educational Assessment*, *13*, 80-107.

Wolf, M. K., Kao, J. C., Griffin, N., Herman, J. L., Bachman, P. L., Chang, S. M., & Farnsworth, T. (2008). *Issues in assessing English language learners: English language proficiency measures and accommodation use* (CRESST Report 732). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

World-Class Instructional Design and Assessment (WIDA). (2014). ACCESS for ELLs interpretive guide for score reports. Wisconsin: WIDA Consortium.

World-Class Instructional Design and Assessment (WIDA). (2013). Annual technical report for *ACCESS for ELLs* English language proficiency test, Series 203, 2011-2012 Administration. (Annual Technical Report No. 8). Wisconsin: WIDA Consortium.

Young, J. W., Cho, Y., Ling, G., Cline, F., Steinberg, J., & Stone, E. (2008). Validity and fairness of state standards-based assessments for English language learners. *Educational Assessment, 13,* 170-192.

Zwiers, J. (2008). Language for academic thinking. Building academic language: Essential practices for content classrooms (pp. 19-39). San Francisco, CA: Jossey-Bass.