Technology-Enhanced Formative Assessment in Mathematics for English Language Learners

A DISSERTATION
SUBMITTED TO THE FACTULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Adam Jens Lekwa

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Jim Ysseldyke, Adviser

December, 2012

Acknowledgments

I wish to acknowledge staff of Renaissance Learning for their generosity and patience. Renaissance Learning granted me access to their data for the quantitative analyses reported in this study, and assisted me in recruiting teachers to interview for the qualitative component of this study. Without their assistance this study would not have been possible.

I also wish to acknowledge my adviser, Professor Jim Ysseldyke, who provided valuable advice, instruction, and feedback throughout my graduate education. Without his encouragement this work might not have been attempted.

Abstract

This paper reports the results of a descriptive study on the use of a technology-enhanced formative assessment system called Accelerated Math (AM) for ELLs and their native-English-speaking (NES) peers. It was comprised of analyses of an extant database of 18,549 students, including 2,057 ELLs, from grades 1 through 8 across 30 U.S. states. These analyses were followed by semi-formal interviews with three teachers in California, Texas, and Minnesota, who use AM in mathematics instruction with ELLs and NESs.

Without taking classroom-level information into account, ELLs, as a group, appeared to receive slightly greater implementation of this formative assessment system than did their native English speaking peers. Yet the results of five sets of generalized linear mixed-effects regression models indicated that, after accounting for classroom membership (or teacher effects), ELLs and their NES peers received roughly equal levels of implementation of the formative assessment system, suggesting that ELLs in this sample tended to be found in the classrooms of teachers who made somewhat greater use of their formative assessment system.

The results of multilevel growth models suggested that ELLs tended to begin the school year with lower levels of mathematics skills, on average, than their NES peers. After accounting for variability associated with classroom membership, there were no significant differences between average growth in STAR Math scores between ELLs and NESs. Implementation of AM was a strong positive predictor of growth for both groups of students, yet there was a significant, but small, negative three-way interaction between ELL status, growth in mathematics, and implementation of AM.

Teacher input, obtained in semi-formal interviews, added perspective to the results of quantitative analyses of implementation and growth. Their responses to questions about patterns of implementation or observations of growth in STAR Math scores provided illustrations of trends encountered in statistical models, and suggested issues for consideration in future research on the topic of formative assessment practices in mathematics for ELLs.

Contents

List of Tables

List of Figures

List of Equations

Chapter 1: Introduction

For many years there has been an increasing emphasis on methods of formative assessment, or the use of data to guide and evaluate instruction. Research on formative assessment has led to recommendations about ways to match the difficulty of instruction to a student’s current skills and knowledge, identification of specific instructional needs, and evaluation of the effects of instruction. Such strategies have been researched and implemented in a variety of contexts, including applications for younger learners developing early literacy skills, or applications for older learners developing content knowledge in subjects such as science. A considerable body of research has developed around the use of assessment to inform instruction in mathematics. Based upon their review of extant research literature, the National Mathematics Panel (2008) recommended the use of formative assessment tools for mathematics instruction—especially when technology can automate data collection and management to provide teachers with information for whole-group and individual student needs, and when teachers are provided guidance on connecting instructional activities to assessment data.

Concurrent with the increased popularity of formative data use is the trend of increasing diversity in cultural backgrounds and educational needs of the population of students attending U.S. schools. Of particular interest is the sub-population of students who come to school without proficient English language. Students who are English language learners (ELLs) may face a variety of unique challenges, including, but not limited to, linguistic barriers in instruction and assessment. These students also come to school with varying levels of English proficiency: some of these students may only have very basic English, others may be conversationally or socially fluent, though still lacking academic fluency.

There has been considerable research devoted to the quality of education for ELLs. Understandably, a great deal of this research is focused on literacy, on disentangling normal second language development from abnormal, or on limiting bias in assessments for accountability or special educational eligibility. Many authors and researchers advocate formative methods of assessment for use with ELLs (good examples include Baker, Plasencia-Peinado, & Lezcano-Lytle, 1998; Deno, 2003). Yet for the most part, such papers are primarily concerned with topics of reading assessment, or are more generally focused literature reviews or editorials (Stiggins & Chapuis, 2005).

Consequently, limited research has been conducted on formative assessment in mathematics for ELLs. Besides instruction and assessment in English, the mathematics learning of ELLs might also be affected by factors such as low family socioeconomic status, early numeracy skills deficits, or lack of exposure to formal education. It is widely known that as a group, ELLs experience the most difficulty in academic skills and subjects relying upon English language and literacy. Perhaps less obvious are issues that arise in learning mathematics.

Although mathematics computation and reasoning do not implicitly require language skills, mathematics instruction in the U.S. is generally provided in English, as are the assessments used to ensure educational accountability. Moreover, students learning mathematics in a second language may have fewer opportunities to learn (via characteristics of the instructional environment such as teacher expectations, and scheduling or placement), or may not experience equal benefit from the same amount of instruction (instructional match variations due to student instructional history, difficulty with academic vocabulary or syntax). Because of potential limitations on opportunities to learn over time, ELLs may require greater implementation of formative assessment practices than students in the general population.

The trend in recent years toward formative use of assessment data shows promise as a strategy to help educators improve math instruction for ELLs, but the instruments and methods used in such assessment could have many of the same limitations as those that arise in large scale summative assessments. Past research on the effects of technology-enhanced formative assessment included ELLs, but effects for ELLs were either not disaggregated in analyses, or the studies were comprised of samples too small to support statistical inferences. Several important questions have not been addressed.

Study Rationale

This study addressed several questions around the use and effectiveness of formative assessment methods for ELLs in mathematics. First, this study examined ways in which one particular mathematics assessment system, Accelerated Math (AM), has been implemented for ELLs in the U.S. Given what is known about the experiences of many ELLs in U.S. classrooms, one may expect to find relatively limited implementation of an instructional resource such as AM (Gándara, Rumberger, Maxwell-Jolly, & Callahan, 2003). Or, considering the pressures educators may feel to have their students make adequate yearly progress, one may expect to find greater levels of implementation of AM for ELLs. Finally, knowledge of ways in which features of this system are used for a large number of students across schools in many states can add perspective to comparative analyses of growth.

This study examined the range of mathematics skill levels with which ELLs and their native English speaking peers start a school year, and differences in how levels of skill change over one school year between these two groups. The analyses conducted within this study were designed to explore overall rates of growth in math, as well as the extent to which implementation of AM produced different levels of growth for both groups of students.

Notes on Terminology

Several of the terms and acronyms used throughout this paper require definition and clarification. First, there are a number contrasts drawn between the two major sub-groups within this study, which were ELLs, and students in the general population or students who are native English speakers (NESs). The term “students in the general population” and the acronym NESs are used interchangeably throughout the paper, although the acronym NESs is favored in figures and tables for efficiency.

Second, the term “formative assessment” is intentionally used in a general sense throughout the paper. There are a range of assessment strategies and products that are associated with this term, such as progress monitoring, curriculum based measurement, or curriculum based assessment. Each of these concepts is an example of formative use of assessment data for various purposes. The technology-enhanced assessment system that produced the data analyzed in this study is referred to as a formative assessment system (in the general sense) because it promotes different types of data use, such as assessment to identify a student’s instructional level, assessment to identify more specific instructional needs, and in many cases, assessment of a student’s skill growth in response to instruction over time.

Chapter 2: A Review of Literature on English Language Learners and the Relationship between Assessment and Instruction

The purpose of this review is to discuss issues in the use of formative classroom assessment to address educational disparities experienced by a specific group of students in U.S. public schools. In the first section of this paper I introduce and describe the group of students considered to be English language learners (ELLs). In the second section I discuss current trends in math achievement for this group. I then compare two paradigms of assessment used to increase achievement, describe ways in which formative classroom assessment in math can be facilitated through the use of educational technology, and discuss unresolved issues related to assessment of English language learners in the classroom.

English Language Learners

Public schools in the U.S. essentially are equipped to provide standardized services to students whose range of needs and characteristics vary widely; such a one-size-fits-all system would function best if all students were alike, had comparable educational needs, responded to instruction in similar ways, and benefitted from the same instruction over the same number of days in a school year. It is generally recognized that all students are not alike and do not benefit equitably from educational services available. The diversity of students in classrooms in the U.S. is an issue of increasing importance both as policies impose ambitious goals for bridging achievement gaps and as specific sub-populations of students attend schools in increasing numbers. School personnel have expended much effort in making systemic improvements at all levels focused on improving educational outcomes for English language learners.

Defining English language learner status. Students with limited English proficiency (LEP) are commonly referred to as English language learners (ELL). As with many categories created by and used for educational policy, there is no universal definition of an English language learner. Although there are no standard criteria that define the nature and extent of a student's English proficiency, the federal government provides at least basic criteria. In general, a student with LEP is a student whose primary language is not English, and who struggles with oral and written English. Under the No Child Left Behind Act (NCLB, 2002), Title IX, section 25, a student can be considered eligible to receive services for limited English proficiency if the student is between the ages of 3 and 21, is either currently in attendance or planning to attend school in the US, has a first language that is not English, and does not speak, write, or read English well enough to benefit equitably from instruction provided in English.

Beyond these general guidelines, there is no universal cut point between adequate or limited language proficiency and deciding where to draw the line can be difficult. Use of available methods such as home language surveys and standardized language tests is complicated by a number of technical issues (Abedi, 2008). Statistics and figures about ELLs in schools, or their levels of performance will vary depending on how the definition of limited English proficiency is applied. National reports, such as the annual Condition of Education report published by the National Center for Education Statistics, tend to make use of parental reports of home languages and English proficiencies, whereas school districts use more involved, individual language and literacy testing.

Consequently, accurate data on the enrollment, involvement, and achievement of these students are difficult to find. Recent estimates state that as of the year 2009 there were about 11,204,000 students attending schools in the US who speak a language other than English at home. The size of this sub-population has essentially doubled since 1979, a growth which is not expected to abate in the near future. About 20% of the total number of students (Kindergarten through 12th grade) enrolled in a public school in the U.S. speak a language other than English at home. Of those students, about 2,654,000, or around 5% of students attending schools at that time in the U.S., are said to have had limited proficiency with English (Aud et al., 2011).

Although their numbers are increasing in every state, ELLs are relatively concentrated in five states: California, Texas, Arizona, New Mexico, and New York. Amongst those five states, California serves the largest population of ELLs. Although students counted within this group come from various countries and cultures, the majority speak Spanish as a first language. Because so many ELLs are from Latino families, the geographic distribution of ELLs is affected; many of those students are concentrated in southern and southwestern regions of the U.S, and also in certain areas of the Midwest and Northeast (Aud et al., 2011). However, it should be noted that not all of these students arrived in the U.S. from neighboring or distant countries: a substantial portion—nearly 60%—was born in the U.S. (Batalova, Fix, & Murray, 2007). Unfortunately these students are often found in lower SES communities, and attend lower performing schools (Fry, 2008; Rumberger & Gándara, 2004; Zehler et al., 2003).

English language learners and academic achievement. The disparity in educational attainment in the U.S. between students belonging to various racial, ethnic, or linguistic groups, referred to generally as “the achievement gap”, is an important issue for educators, students, and researchers. The singular term ‘achievement gap’ is somewhat misleading in that it is a global reference to a complex pattern in which students from racial and ethnic minorities consistently obtain lower test scores than students from the “cultural majority”, who tend to include Caucasians from lower, middle, and upper socioeconomic backgrounds. Rather than a single pattern of inequality between academic performance of students who are from the cultural and linguistic majority and students from various minorities, the “achievement gap” is actually a set of inequities which vary over time, across academic subjects, between measurement instruments, and demographic categories.

Participation in systems of accountability. For the reasons stated above, capturing the complexity of patterns in academic achievement is challenging, and it is a difficult issue to study without misrepresentation or oversimplification. Information on the gap in reading and mathematics achievement between students who are ELL and students from the general population in U.S. public schools has been, for a variety of reasons, difficult to obtain and interpret. As a result, detailed descriptions of the performance of this group can be sparse. This is due in large part to the history of this group’s participation in the state and national assessment programs which are used to monitor the progress of groups of students, and to hold accountable the educational systems that served them.

For example, consider the assessments administered by the National Assessment of Educational Progress (NAEP), the test used by the National Center for Education Statistics to report information to governmental agencies and the general public on long term educational trends in core skill and content areas. Due to the technical barrier of English proficiency, many students were not granted access to the test. Currently, this is an issue that can be addressed through assessment accommodations (methods used to adjust instruments to meet student needs). Until 1996, no accommodations or modifications were allowed on NAEP tests for students with limited English proficiency. As a result, many ELLs were excluded from testing when it was determined by their schools that they would not be able to participate meaningfully (Olson & Goldstein, 1997). Similar trends occurred in state assessment systems as well. Even by 2002, information about the participation and scores of ELLs in reading and mathematics was limited. At that time most states did not report inclusion or exemption rates of ELLs from their large-scale assessments in either core subject (Albus, Thurlow, & Liu, 2002).

More recently, the participation of ELLs in annual testing has been increasing as a result of assessment accommodations. An accommodation is an adjustment to test procedures or materials that does not affect the measurement of a focal variable (such as reading comprehension or mathematical proficiency). Although there can be a great deal of difficulty in development of valid and effective methods of assessment accommodation, requirements stemming from educational testing policy underscored their importance and encouraged their use. In the short time after the publication of the survey of 1999-2000 participation data by Albus, Thurlow, and Liu, there were reports that nearly one fourth of all ELLs tested used accommodations (Zehler et al. 2003). States have made progress in assessment policy for this group of students (Shafer Willner, Rivera, and Acosta, 2008), but comprehensive data on the participation and performance of this diverse group of students in state tests have been difficult to obtain.

What is known about the performance of ELLs relative to students in the general population in annual state assessments is limited by several issues. First, although their scores are grouped together and disaggregated such that they can be compared to those of students receiving general educational services, ELLs are not a homogenous group, and research suggests considerable within-group variability in large scale test scores (Stevens, Butler, & Castellon-Wellington, 2000). Abedi (2004) summarized some of the technical issues involved in the academic assessment of ELLs as a targeted group under NCLB. The numbers of ELLs in the U.S. are highly concentrated within a few states, and baseline scores for schools serving larger proportions of ELLs may be lowered, resulting in greater difficulty attaining Adequate Yearly Progress (AYP). Overall, the accuracy of AYP determinations for ELLs is hindered by a lack of consistency in criteria for classifications of ELLs across districts and states, and due to low statistical reliability and validity of state tests designed for use by native English speakers. In addition, it is important to note that as ELLs make linguistic and academic progress, they may be exited from the federal LEP category.

ELL achievement in math. Due to trends in ELL participation in annual testing and states' generally limited tendency to report scores, accountability data have been for the most part unavailable. Albus, Thurlow, and Liu (2002) reviewed data on test score gaps in mathematics on criterion referenced tests reported by 17 states. In all states reviewed, ELLs scored lower on average than students in the general population. The range of the gap across states varied greatly. Differences between percentages of students who tested at levels of proficient or above proficient ranged from nearly zero to over

50%. Only four states reported the amount of information required to determine actual rates of participation, and therefore produced numbers suitable for statistical analysis.

Within-state comparisons of the mathematics achievement of ELLs and students in the general population may not portray important aspects of trends in achievement. As noted by Abedi (2004), the high concentration of ELLs within specific regions complicates AYP determinations. More useful information about trends in the performance of this sub-group can be obtained through the study of differences across individual schools serving larger numbers of ELLs. In a review of school-level data from five states serving high numbers of ELLs, Fry (2008) found that much of the achievement gap observed between ELLs and students in the general population can be explained by school-level characteristics, as ELLs tend to be enrolled in schools with lower overall levels of achievement. By considering school-level characteristics associated with disparities in achievement, methods for improving the achievement of ELLs are more readily identifiable.

Social factors that impact learning. There are a variety of issues that affect the learning and performance of ELLs in schools. Some are issues unique to students learning content and skills in a second language and some are more widespread issues impacting educational outcomes for all learners, regardless of educational classification. Variables such as acculturation, time in the U.S., family socio-economic status, and attributes of schools, teachers, and learners have been associated to varying degrees with educational outcomes for ELLs. Given such a wide array of potentially important variables to address, the problem of improving the overall achievement of this group of students can appear to be a complex and prohibitively difficult task. Key areas of concern are identified and described below.

Language. Limited language proficiency is the one characteristic all ELLs share in common, although there is still variable linguistic proficiency between individuals. For ELLs, most of the emphasis in practice and in research is therefore on language acquisition and literacy. The extent to which a student possesses English proficiency adequate for learning from instruction provided in English is not necessarily based on a dichotomous notion of proficiency (proficient, or non-proficient English.). Cummins (1979, 1981) proposed a distinction between two basic continua of language proficiency, which would become known as Basic Interpersonal Communication Skills and Cognitive/Academic Language Proficiency. Although controversial, this perspective has been adopted widely in practice and in policy.

Students without adequate academic English proficiency have lower performance in language-centered skills such as reading, but because verbal language is a primary means of accessing instruction, lack of linguistic proficiency can reduce a student’s access to instruction or meaning (Bradby, 1992; Abedi & Herman, 2010). One example of the complex relationship between language proficiencies and mathematics achievement was provided by Beal, Adams, and Cohen (2010) in their research on mathematics performance and self-concept in a California high school with a 47% ELL student population. They concluded that measures of conversational proficiency in English were not significantly related to math achievement, and that proficiency in comprehension of English text could be used to predict math self-concept as well as math test scores.

Acculturation and time in country. Acculturation refers to the extent to which an individual has learned the norms, customs, and social language of the culture within which that individual lives. Acculturation has been studied in a variety of ways, and based upon differing perspectives and models (for a concise review, see López, Ehly, & Garcia-Vázquez, 2002). Effective integration into an English-speaking culture would, understandably, be challenging for students who are not native English speakers. Yet currently there is not a well developed literature base on the link between acculturation and academic achievement. The extent to which acculturation can account for variability in academic achievement—especially achievement in specific areas such as mathematics—is unclear. This could depend in large part upon the variety of definitions and frameworks of acculturation used. Some authors suggest that as the integration between cultures increases, culturally and linguistically diverse students experience less social difficulty within the educational systems of the cultural majority (López, Ehly, & Garcia-Vázquez, 2002; Lee, 2002).

Acculturation is a complex construct; understanding its influence may be simplified through identification variables that can be observed directly, instead of defining abstract social constructs. The simplest and easiest indicator of acculturation (as described above) could be the length of time for which a student or student's family has lived within a given cultural majority. In general, the number of opportunities to acquire culturally specific knowledge and language increases over time as individuals integrate within their communities. Although it might be expected that the similarity of the family's primary language to English would be more important to the development of English literacy, the length of time a student's family has lived in the U.S. has been shown to be a stronger predictor of English literacy than the language spoken by the family (Betts, Bolt, Decker, Muyskens, & Marston, 2009). Time to linguistic fluency is an important consideration for educational and policy decision making, as argued by Hakuta, Goto-Butler, and Witt (2000) in a study presenting evidence that on average, ELLs can be expected to attain social fluency in English within three to five years, whereas full academic fluency can require anywhere from four to seven years. There is some evidence that this effect may extend across generations within families.

Socioeconomic status. Socioeconomic status is an abstract reference to the extent to which an individual or family has access to social or economic resources; it has played a prominent role in current understandings of the development of academic problems (White, 1982), but is conceptually diffuse (Tate, 1997). There is evidence in educational research literature that family SES is positively associated with the development of numeracy in children. Common explanations for this pattern include parental educational attainment, exposure to quantities, numbers, or other concepts in early numeracy in childhood, or school and community variables affecting school quality or the quantity of instruction (Reyes &Stanic, 1988; Jordan & Levine, 2009).

Bradby (1992) analyzed data on the academic performance of Asian and Hispanic students from the National Educational Longitudinal Survey of 1988 (NELS:88). Mathematics performance of these two broad groups of students was positively related to English language proficiency, although it was observed that low SES was associated with comparatively lower performance regardless of English proficiency. In a re-analysis of data from prior studies on ELL mathematics achievement, Krashen and Brown (2005) provide evidence suggesting that SES can moderate performance on standardized assessments of mathematics achievement. In each of three studies reviewed, students classified as ELLs from middle to higher levels of SES either outperformed or did as well as students considered to be fully proficient in English who came from lower SES backgrounds. Their findings suggest that on average, language proficiency may be associated more weakly with the development of mathematical abilities than effects of low SES.

Instructional factors that impact learning. Although certain aspects of acculturation, family's length of time in the country, and socioeconomic status have been associated with academic functioning of ELLs, none of these variables can realistically be manipulated by educators to improve performance of this group, nor are results from studies on such variables consistent enough to inform practice or policy. Fortunately, the realm of instructional factors is one area in which educators can (and do) attempt to reduce trends of inequity. But unfortunately, ELLs have often been in the classrooms of teachers without much relevant training or qualifications in the provision of instruction to culturally and linguistically diverse students (Gándara, Rumberger, Maxwell-Jolly, & Calahan, 2003).

Expectations. The process of instruction begins with expectations that students can or cannot learn new skills or information; beliefs about students' potentials to learn guide instructional behavior and decision making. Teachers' expectations may affect students within interpersonal contexts, such as the classroom. When teachers communicate high expectations and emphasis on individual potentials, student learning increases (Goddard, Hoy, & Woolfolk-Hoy, 2000). There is evidence to suggest that this effect could be greater for minority students (Good & Nichols, 2001).

Some educators may have lower academic expectations for, and fewer interactions with, minority students (Garibaldi, 1992). Edl, Jones, and Estell (2008) reported evidence of such an effect for students from linguistic minorities in a study on teacher perceptions of academic and interpersonal competence between European American and Latino students, and concluded that students with lower proficiency in English (receiving language services) were more likely to be viewed as less socially and academically competent than Latinos or European Americans in regular classrooms. The effects of expectations on learning can also occur within systemic contexts such as tracking. In a study on the effects of tracking on high school English language learners, Callahan (2005) found evidence that achievement in mathematics could be better predicted by track placement than by language proficiency.

Instructional match. In addition to expectations held by teachers, students' expectations for their own success or failure when presented with academic tasks are associated with the time spent on those tasks. Betts (1946) described three general levels of difficulty, including frustrational, instructional, and independent. When presented with tasks that are too difficult, students are more likely to engage in task irrelevant behavior. Higher levels of engagement are obtained when students are able to experience success on a majority of task trials (Burns & Dean, 2005; Shernoff, Csikzentmihalyi, Schneider, & Shernoff, 2003).

The match between methods of instruction and a student's current skill levels is also understood to affect rates of learning. Successful learning of a new skill is moderated by the extent to which the student is prepared to learn the new information as it is presented. The match between instruction and learned skill is characterized by the set of new information or skills for instruction and the student's existing skill levels along a hierarchy of stages of learning. In a meta-analysis of empirical findings on the match between instruction and specific aspects of students' skill levels, Burns, Codding, Boice, and Lukito (2010) concluded that mathematics instruction was more effective when targeted to specific profiles of skills.

Opportunities to learn, and practice. In addition to level of difficulty and targeting, the effectiveness of instruction is moderated by the amount of time in which the student is engaged in learning. John Carroll (1963) developed a conceptual model that was intended to classify common influences on learning at school. This model included five "classes" of variables that he determined were related to the amount of learning each student could achieve in an amount of time. Carroll's model was unique in that it organized important variables by student characteristics and characteristics of the learning environment (rather than student characteristics alone, or environmental factors alone).

Carroll defined *aptitude* as the time a student requires to be able to master an instructional objective. Another class of student characteristics Carroll identified was the *ability to understand instruction*, which is the extent to which the student has the requisite language proficiency and background knowledge to be able to acquire and retain new information. Carroll also specified ecological classes of variables: *Quality of instruction* refers to teacher behaviors such as provision of feedback; *Perseverance* refers to the length of time a student spends engaged in the task; *Opportunity to learn* is the time provided by the teacher for instruction and learning to occur.

Under this model, the extent to which learning occurs is a function of time spent engaged in a learning task out of the time required for a student to reach mastery. Although this particular model has not been validated empirically, it has been influential in the thinking of educational researchers and practitioners. A considerable amount of work was completed in the following decades in which the importance of providing a high number of opportunities to learn was documented.

In addition to the importance of the opportunity to respond, students require frequent opportunities for practice. Practice not only maintains prior learning, ensuring ease of recall or performance in future tasks, but practice can also facilitate development of unknown information or unlearned skills. As long as a student has engaged in sufficient practice at an appropriate level of difficulty, allowing for a high degree of success across trials, the student is minimizing erroneous learning. By guiding a student's practice, and providing ample opportunity for practice, the correct skills and knowledge are reinforced.

Feedback. Provision of feedback during instruction is highly related to the provision of opportunities for practice of new skills and knowledge. Feedback is most useful when given soon after a student's academic behavior, and when in reference to task performance instead of student characteristics (Kluger & DeNisi, 1997; Rodriguez, 2004). Additional discussion of the important role of feedback in learning is given under the subsection on specific characteristics of formative assessment (pg. 24).

The Role of Assessment in Instruction

Assessment in education is the collection of data to inform decision making (Salvia, Ysseldyke, & Bolt, 2010). It plays an integral role in setting and achieving goals in education. One of the ways in which the educational system in the U.S. is trying to support its ELLs is through use of annual assessment to hold schools and teachers accountable for the achievement of all students, specifically those from backgrounds of poverty, students with disabilities, or ELLs. Yet these efforts have been controversial, and have seen mixed results—both positive and negative. In this section I describe two types of assessment, and discuss issues involved in the use of assessment for ELLs.

Two uses of data. Broadly speaking, there are two applications of data obtained from educational assessment: formative and summative. The traditional form of classroom assessment is a summative process in which the level or status of some skill is documented after instruction has ended. By contrast, formative assessment is a practice that integrates the collection of data on student performance and the delivery of well-informed instruction. Formative assessment in classrooms is neither a new idea nor complex. Rather, it is an application of general assessment practices or individual tests to a different set of goals. In the practice of formative assessment, the purpose is to provide instructors with information that will be used to enhance learning. This distinction between "formative" and "summative" evaluation was first made by Michael Scriven in 1967 on the topic of formative versus summative evaluation of curricula; this concept was soon adopted into the realm of educational assessment (Bloom, Hastings, &Madaus, 1971), and it is beginning to see widespread and systematic application in classrooms.

Large scale assessment programs are intended to help promote improvements in education. From that perspective it would make sense to say that they are formative in nature. However, the primary function of large scale educational assessment in the U.S. is to account for gain in academic achievement, most typically as part of school accountability systems. At this point in time, such tests are not typically designed to be instructionally informative. This review is not intended to argue against the use of summative measures such as annual state tests, but I instead propose that smaller scale assessments used in a formative method are necessary to guide instruction improve learning for traditionally underserved populations such as ELLs.

Summative data use. The U.S. federal government has held schools accountable for achievement since the 1965 passage of the Elementary and Secondary Education Act (ESEA). Initially, the focus of policy was on process instead of performance. It was generally assumed that improved educational outcomes resulted from provision of (and magnitude of) specific types and amounts of services, and that by examining kinds and amounts of services schools could be held accountable for the achievement of their students. During the 1990s the focus changed to demonstration of accountability through documentation of student outcomes, and thus began trends in annual standardized testing in each state. Arguably, a focus on product rather than process is more likely to provide a stronger guarantee of good educational outcomes.

Since passage of the No Child Left Behind act in 2001, the most recent reauthorization of the Elementary and Secondary Education Act, states have been required to report test results for a number of typically underserved sub-groups of students, including students considered to be LEP; any school serving a certain percentage of ELLs is required to report scores of those students separately as a group, rather than hidden within achievement data from the general population. By requiring states to report the performance of targeted groups separately, the U.S. federal government intended to encourage schools to improve performance of traditionally under-served groups of students.

As a strategy to improve performance of targeted groups on a grand scale, large scale summative assessment is limited by the technical adequacy of the tests used by states. For instance, it has been found that linguistic complexity of tests items reduces the reliability of tests for students with limited proficiency in English, thereby invalidating resulting scores in math—particularly math subscales that are highly text based, such as problem solving, rather than those that are primarily digit based, such as computation. Linguistic barriers also limit the number of items attempted and completed by students with limited English proficiency, further detracting from the overall validity of scores obtained (Abedi, Lord, & Plummer, 1997). In general, it is thought that linguistic barriers obscure scores on content assessments for ELLs, perhaps artificially lowering scores, and producing irrelevant variance (Abedi, 2003; Abella, Urrutia, & Shneyderman, 2005).

Because of the technical issues associated with these types of tests, the extent to which educators can demonstrate accountability via gains in math performance for ELLs is limited. Several authors have argued that high stakes tests for accountability as mandated by NCLB have unintended and potentially serious consequences for struggling students from minority groups. In general, the opposition is not to the use of large scale standardized tests of achievement, but rather to the negative consequences contingent on inadequate test scores, thereby making testing a "high stakes" situation for schools and students. It has been argued that assessment practices with outcomes that entail, at best, escape from punishment do not help to increase academic achievement for minority students (Amerein & Berliner, 2002a; Cizek, 2001), and there have been concerns that they potentially can result in such negative outcomes as decreased performance in math and reading, and increased dropout rates (Amerein & Berliner, 2002b).

Information obtained in qualitative studies of student perspectives on high stakes testing lends support to the notion that assigning negative consequences to tests for accountability does not necessarily produce the intended results for learning. Diamond and Spillane (2004) found that lower performing schools which are placed on probation for failing to meet Adequate Yearly Progress (AYP) goals tend to resort to instructional practices geared specifically toward tested skills, the primary intent being to get off probation, whereas schools not under probation were found to use test data to monitor long term trends in achievement and to aid instructional planning. Cleary (2008) reached a similar conclusion using interview data from 120 Native American students: that instruction geared specifically toward rote and test taking skills would suppress, rather than enhance, student motivation for learning and likelihood for subsequent academic improvement.

Formative data use. The top-down paradigm using policy and sanctions to encourage schools to educate all students equitably and effectively cannot help educators produce the desired effects. To realize long term goals in improvement of the academic achievement of typically underachieving sub populations, more localized, “bottom-up” strategies will be required. For example, Krueger and Whitmore (2002) found evidence that smaller class size helps to reduce achievement gaps, teen pregnancies, and crime in addition to substantially increasing the likelihood for students from minority groups to take college entrance exams. These improvements could be attributed to the finding that students in smaller classes receive more individualized instruction (Molnar, Smith, Zahorik, Palmer, Halbach, & Erle, 1999). Modifications of class size or individualization of instruction are good examples of changes which can be enacted locally that may enable teachers to work more effectively, and improve the learning of all students. Federal and state assessment programs may help ensure that educators hold their students to higher expectations of achievement, but the extent to which such testing can directly inform better instruction is limited. As a localized, direct approach to using data to improve instruction, formative classroom assessment may be a promising bottom-up strategy for helping schools bridge achievement gaps.

Reviewing research conducted on the topic of formative classroom assessment can be difficult for a variety of reasons. It is a fairly general concept with an empirical and theoretical basis that spans across fields from educational measurement to special education and psychology. As a result, a variety of terms have been used to describe the very similar ideas, or the number of similar practices which resemble this concept either in theory, practice, or both. Authors have referred to formative assessment in an educational context with such terms as ‘classroom evaluation’ (Crooks, 1998), ‘formative evaluation’ (Fuchs & Fuchs, 1986; Bloom et al, 1971), ‘assessment for learning’ (Stiggins, 2005), or ‘formative assessment’, the term used by the often cited authors Black and Wiliam (1998a).

Black and Wiliam (1998b) provide perhaps the most conservative or generalized definition of formative assessment. In their article entitled “Inside the Black Box”, they describe the “self evident” interactive nature of teaching, and how assessment is an integral part of the interaction. They argue that the assessment that occurs in classrooms becomes *formative* when the information is applied to improve student learning. Stiggins

(2005) elaborated the idea of using assessment data to adjust instruction and added an emphasis on student-teacher partnership in the assessment process, stressing the importance of developing students' affective responses to assessment outcomes. Boston (2002) also approached the topic of formative assessment in a slightly different perspective, noting that central to any formative assessment is an assumption that students' skills do not reside within a pre-determined and static range, but instead are limited only by the quality of instruction delivered.

Distinguishing characteristics of formative assessment. The most common distinction between formative and summative assessment is generally considered to be the way in which the data are used: data from formative assessments are used to improve performance, whereas data from summative assessments are used to describe performance. Yet this distinction is overly simplistic; it could also be argued that tests used for summative purposes are also intended to improve student performance—that is the ultimate goal of current large scale testing programs.

The key difference is in the way in which improvement is sought. Data that are used in the formative fashion are used to highlight the type and scope of instruction needed, and to evaluate the effects of that instruction after it is delivered. In this way the process is both iterative and interactive. Summative assessments are primarily iterative, and the results can only have indirect effects on future achievement (scores are calculated, and not used to fuel a process improving the same score in future measurements of the same skill). Key components of the iterative and interactive process of formative classroom assessment are described below.

Frequency of data collection. In formative classroom assessment, measurements are taken at multiple points during instructional periods instead of at the end of instruction. As a result, this type of classroom assessment is distinguished from its summative counterpart, to an extent, by the frequency of data collection. Frequent testing does not necessarily guarantee improved instruction and learning. The key is that data are collected so that instruction can be adjusted accordingly. The rate at which data are collected is related to the rate at which the effects of instruction can be evaluated, and practices adjusted accordingly.

Detection of change over time. Formative classroom assessment involves the use of regular, repeated measurements of students' progress in some domain of skills. Measures used to monitor progress may not test the exact skill or concept that is being taught, but will have a robust link to instruction; they will reflect performance in a broad area of content instead of specific sub skills (Fuchs & Deno, 1991). For instance, oral reading fluency has a strong relation to overall reading comprehension, though fluency itself is not an end goal, and the probes used to measure fluency can be passages unrelated to curricular materials in the classroom. Measures should be able to adequately differentiate performance at single points in time (allowing static comparison of peers) and at multiple points over time to observe rates of learning (Fuchs, 2004; Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993; Shinn & Bamonto, 1998).

One difficulty associated with progress monitoring in math is the determination of what to monitor. Most, if not all, methods of monitoring progress in math consist of measurements developed for procedural knowledge (computation), or for conceptual knowledge (application of mathematical knowledge to 'real world' items, or quantitative problem solving). Targets for measurement can either be developed from the curriculum in use, or can involve more general indicators of progress not linked to a specific curriculum. Both methods have advantages: developing measurements linked to a particular curriculum can facilitate direct evaluation of students' responses to instruction, as well as problem analysis when instructional changes are deemed necessary. However, progress indicators (such as many commercially available options) that are not linked to a particular curriculum may facilitate measurement of progress across grade levels.

Analysis of skills. Fuchs and Deno (1991) distinguished between "general outcome measures" and "sub-skill mastery measures". Measurement of student progress often relies upon measurements of general outcomes (for example, speed and accuracy of computation). Measurement of general outcomes can inform instruction at a basic level by providing instructors with a valid and timely evaluation of the instruction provided to each student measured. This signals the need for instructional adaptation during the learning process. But for this to occur successfully, assessment data must help identify specific learning needs, thus identifying what or how to teach. This requires a focus on assessment of mastery of important sub-skills. The skills students learn in school consist of complex hierarchies of subskills; successful mastery of basic subskills facilitates performance of higher level skills. In reading, learning the alphabet and letter sounds allows a student to read words. Learning to recognize words rapidly and accurately allows students to read and comprehend sentences; comprehension is also a function of a student's vocabulary. Any deficit in a higher-order mathematical skill (such as mathematical problem solving, requiring both procedural knowledge and fluent retrieval of basic facts) could be better understood by careful examination of mastery in sub-skills.

Feedback. Formative classroom assessment is interactive in two ways. First, students are a source of performance data for teachers such that teachers can modify instruction accordingly. Second, teachers provide students with feedback information. As described above much of this process is based on teacher behaviors: instructors collect and use information related to their students' performance on a regular basis, continuous progress monitoring signals when instructional adjustments are necessary, more targeted assessments of specific sub skills inform how to adjust instruction, and the extent to which performance responds to instruction is evaluated. However, this type of classroom assessment is also interactive because of the role of the student. There are a range of opinions on the roles students should play in formative assessment. Sadler (1989) proposes student-regulated monitoring and self evaluation. Others, such as Wiliam, Lee, Harrison, and Black (2004) or Boston (2002) have focused more on teachers' roles. From either perspective, students are the end-users of the information produced by formative assessment.

Having regular access to performance data for specific instructed skills allows teachers to provide students with adequately challenging opportunities for practice, and to provide immediate corrective feedback. For students, feedback facilitates learning when there is a focus on individual performance and reinforcement of individual progress rather than difficulty of the task, comparison to peers, or letter grades. But not all feedback is equal. In a fairly large meta analysis involving 607 effect sizes, Kluger and DeNisi (1996) concluded that effects of feedback are most effective when oriented toward the task, and potentially counter-productive when oriented toward the learner.

Truly formative assessment would provide task oriented feedback—often through charts of student progress (for an example, see Fuchs, Fuchs, Hamlett, & Whinnery, 1991).

Effects of formative assessment on learning. Given the simplicity or intuitive logical appeal of the concept, it could be assumed that assessment which occurs in most classrooms is at some point formative, whether explicitly intended to be so or not. As for the sorts of assessment that actually occur in classrooms, there is evidence to suggest that teachers may engage in a fairly diverse array of practices in their classrooms; the variety of testing methods, frequency, and intended uses of data collected do not seem to be associated with number of years teaching, educational setting, grade level, or even district policy (Cizek, Fitzgerald, & Rachor, 1995). It seems that these assessment practices also lack sensitivity to specific sub-skills (versus global skills, like reading proficiency), or the data are not reviewed and applied to the purpose of problem solving. Consequently, teachers are generally able to identify when individual students are struggling, but are less able to derive specific explanations for low performance (Bailey & Drummond, 2006).

Though many authors praise methods of formative classroom assessment and encourage their use, it is a concept that does not appear to be generally understood, and has seen limited implementation in actual classrooms (Marsh, 2007; McNair et al., 2003). Instead, when teachers deliberately engage in classroom assessment, they may do so in a mostly summative fashion intended to evaluate student performance and effort with grades at the end of a unit or period of learning.

The effectiveness of formative classroom assessment practices has been studied often over the past few decades, but a detailed picture of the knowledge base in the literature can be difficult to obtain. Many of the papers on practices which could be (or are) considered formative contain differing terminologies for similar concepts, or emphasize certain strategies (such as provision of feedback) over others. However, efforts have been made to overcome these obstacles, and several reviews and at least one meta-analysis are available that lend some empirical support to the conduct of formative assessment.

Black and Wiliam (1998b) conducted a review of the literature reporting research which had been conducted to study or demonstrate the effectiveness of formative assessment techniques. While acknowledging the difficulties inherent in dealing with such a diversely defined topic appearing in various forms across fields and over a number of years, they identified and reviewed a large number of studies in which principles of formative assessment were employed and results were reported. This process, less selective than that which would be necessary for a meta-analysis, did not lend itself to the creation of overall effect sizes, but the authors were able to identify several key components and broad outcomes of formative assessment. They concluded that in general, formative assessment can be expected to result in moderate to strong effects when the essential elements of feedback, student engagement, and data use for instructional modifications are present.

Twelve years prior to Black and Wiliam's literature review, Fuchs and Fuchs (1986) conducted a meta-analysis on what they termed "formative evaluation". They described the concept of formative *evaluation* in much the same way as Black and Wiliam describe formative *assessment*: "…formative evaluation focuses on ongoing evaluation and modification of proposed programs", (pg. 200). This description was meant in contrast to an approach to educational evaluation in which assessments were used in attempts to try to identify cognitive or psychological characteristics of students in order to prescribe an instructional plan. Citing a lack of empirical support for such an approach, Fuchs and Fuchs intended to demonstrate the effectiveness of the “inductive, rather than deductive” strategy of formative evaluation. They identified twenty-one published studies that met their inclusion criteria, which specified that a study should have followed a quantitative design including control samples and a focus on formative evaluation and academic skills for students in special education. They observed an overall effect size of 0.7, indicating not only significant, but meaningful evidence of gains in achievement for students in special education.

Both Fuchs and Fuchs (1986) and Black and Wiliam (1998) reviewed studies that reported greater effects for lower achieving students than for higher achieving students, but the extent to which such methods can be expected to reduce achievement gaps for students in specific at-risk sub-populations may be limited. Much of the literature written specifically about formative assessment lacks operational definition and empirical quality. The ideas and terminology are prevalent—the extensive review by Black and Wiliam has been cited by over 1800 other papers—but the scope of available literature may be greater than the technical quality, suggesting that the idea, at this point, could have more logical appeal than empirical basis. Providing a cautionary reminder of the limits of educational research, Dunn and Mulvenon (2009) questioned the extent to which the oft cited conclusions of Black and Wiliam were ‘conclusive’, and highlighted the need for more research with technical rigor.

On the other hand, the lack of conclusive empirical support for the formative style of assessment could be due to its conceptual breadth. Much of the literature on the *idea* of formative assessment does not typically include articles reporting the results of research conducted on the ‘curriculum based’ family of assessment frameworks. There is, perhaps, a greater pool of evidence to be reviewed within this realm of educational research. The ‘curriculum-based’ family of assessment frameworks is formative conceptually, though with more detailed and operational definitions, and an emphasis on technical qualities of reliability and validity.

There is some documentation of instances in which formative classroom assessment has been used effectively for minority students. In a report published by The El Paso Collaborative for Academic Excellence, Blot, Della-Piana, and Turner (1998) demonstrated a reduction in the achievement gap between Caucasian and Latino students in math and science and a growth in achievement overall in response to the use of two evaluation instruments designed to inform instructional decision making. Yet the extent to which the findings reported in this example of formative assessment can be generalized to a broader, national context for minority students may be limited. To date, there is apparently very little research specifically on the topic of formative classroom assessment for students from minority populations, including ELLs. This issue was discussed by Abedi (2009), who noted specific research needs, and provided a helpful review of issues to consider when planning studies in this area.

Use of Technology to Manage Assessment

On a basic level, there have been those who disagree with methods of formative assessment due to differences in philosophies regarding ‘problems’ and ‘disabilities’; this involves differences in explanations for problem etiology (internal vs. external), and differences in assumptions about proper expectations or rates of learning (Shinn & Bamonto, 1998). Aside from conceptual differences, perhaps the greatest barrier to effective implementation of formative methods is logistical feasibility. The process of progress monitoring and assessment of student performance data to inform instruction is time consuming. In fact, one of the contributing factors to the lack of formative classroom assessment practices is a common perception held by teachers that gathering such amounts of information with such frequency would require too much time (Tierney, 2006).

The scope and complexity of data *potentially* used within such frameworks are such that it is not reasonable to expect teachers to collect, manage, and use data to inform instruction. However, given the nature of the data to be collected and the availability of information technology, much of this process can now be automated, giving teachers increased time to integrate instruction with assessment data. It was not long after the emergence of the notion of formative classroom assessment as a distinct paradigm that computers were being used to address problems of logistics (Fuchs, 1998).In an early example, Fuchs, Fuchs, Hamlett and Stecker (1991) documented increased student performance in mathematics as a result of teacher use of an ‘expert system’—a software application designed to manage progress monitoring data, and to allow teachers quick access to reports on individual student progress. There are also examples within a post-secondary context. For example, Buchanan (2000) demonstrated experimentally that use of a computer assisted system for the assessment of student progress (and provision of corrective feedback) was effective for improving levels or performance in undergraduate psychology courses.

One system for formative assessment in mathematics. In classrooms with built-in tools for continuous progress monitoring and management of student data, teachers have more opportunities to individualize instruction for students, and students can receive greater, more meaningful feedback along with more opportunities to respond at instructional level. Accelerated Math is an example of a commercially available computer-based progress monitoring system designed to largely automate the process of formative classroom assessment, and to provide support to instruction within mathematics curricula used by teachers (Renaissance Learning, 1998). The system is comprised of a computer adaptive mathematics assessment, computer-generated math worksheets and objective tests, an optical scanner to upload student responses, and a software package to store and report information on math performance to teachers and students alike.

A fair amount of evidence obtained by researchers over the past 10 years supports the effectiveness of Accelerated Math. Spicuzza and Ysseldyke (1999) reported the results of a study including a small sample of students in elementary and middle school a mandatory summer school session; rapid gains in math achievement were observed in comparison to progress made in the following nine month school year. Ysseldyke, Spicuzza, Kosciolek, Teelucksing, Boys, and Lemkuil (2003) demonstrated the effectiveness of Accelerated Math for students at low, middle, and high levels of achievement, and provided initial indications that the effectiveness of this program is moderated by the extent to which teachers implement it with high degrees of fidelity.

Empirical studies of the benefits of Accelerated Math have also demonstrated gains in math performance for students in Title 1 programs (Ysseldyke, Betts, Thill, & Hannigan, 2004), and for students in talented and gifted programs (Ysseldyke, Tardrew, Betts, Thill & Hannigan, 2004).

Ysseldyke and Tardrew (2007) conducted a large-scale study of Accelerated Math as a Progress monitoring and instructional management tool including 2202 students in the third to the tenth from 24 states in the U. S. One hundred twenty five teachers from 47 schools were assigned to treatment and control groups. The system was implemented for a full semester pending a one-day training session provided by the researchers. Results of analyses of achievement gains paralleled those obtained in prior implementation studies: large gains were observed across grade levels favoring use of Accelerated Math, and gains were consistent across initial levels of mathematics achievement. A review of survey data collected at the conclusion of the study suggested that teachers in the treatment condition who used Accelerated Math to monitor student progress and manage instruction reported allocating more time to individualized instruction, and experiencing greater success in meeting the needs of their students. Students in the treatment group tended to report feeling better about math than prior to their semester with Accelerated Math.

Although information in demographic descriptions of the samples of several studies on Accelerated Math identify numbers of ELLs, only one Accelerated Math study specifically reported effects for ELLs.Teelucksingh, Ysseldyke, Spicuzza, and Ginsburg-Block (2001) examined the effect of Accelerated Math for ELLs in tandem with delivery of academic consultation to teachers in a large Midwestern school district. Gains in achievement were observed for ELLs in a treatment group compared to controls over the course of one semester with Accelerated Math and consultation for instructional planning. Yet the sample sizes obtained for treatment and control groups of ELLs in this study were too small to warrant quantitative analysis (7 and 16, respectively). Similarly, comparisons could not be made between the effects of instruction that incorporated formative assessment for ELLs and those for students from the general population.

Implications for English Language Learners

Students from linguistic minorities are by far the fastest growing demographic in the U.S. educational system, and the need to determine effective strategies for the instruction of English language learners persists. An estimated 2.65 million students are now classified as ELLs, and are concentrated for the most part in southern and western regions of the country. Partially in response to this growing educational need, federal policy mandates their inclusion in annual state testing used to hold schools accountable for student academic progress. Both participation and performance data from such assessments are not highly available or interpretable, but studies conducted independently by researchers and organizations since before and after the passage of NCLB consistently indicate lower performance by ELLs in reading and math.

The vast majority of research conducted on this issue focuses on literacy and on improvement of English language skills. However, current educational policy, in addition to individual students needs, demands increased performance in overall academic achievement. Performance in all areas can be expected to increase as a result of improvement in English proficiency, but academic proficiency takes years to develop. Students cannot wait years before being taught how to work with numbers and solve quantitative problems. Based on what is known about components of effective instruction, formative classroom assessment is a practice that shows promise for helping ELLs acquire skills in math. By collecting data on student performance at all points *during* the learning process, teachers are better able to match instruction to students' instructional levels, provide immediate corrective feedback and reinforcement, allow students greater opportunities to maintain skills through practice, and monitor progress toward curricular objectives. These ideas have received plenty of attention, and the concept obviously has logical appeal, but there are several important issues that should be addressed.

Components of effective instruction for ELLs. Although proficiency in academic English can complicate communication between instructors and students, there is no indication that commonly identified components of effective instruction would be any less important for ELLs. Yet there is also a lack of evidence to the contrary. There is currently a lack of research specifically addressing these elements effective instruction and interventions for students who are ELLs. Much of what research is conducted on this sub-population focuses on issues of special education evaluation and large-scale assessment (Albers, Hoffman, & Lundahl, 2009). Shyyan, Thurlow, and Liu (2008) provided some preliminary evidence in a qualitative study that included a survey of educators serving ELLs with disabilities. Among the results of strategies endorsed by teachers for instructing ELLs in mathematics was the provision of ample practice of newly acquired skills, monitoring of student progress, and the use of charts to give feedback to students on their progress toward objectives. Though the research design and the sample size did not allow for generalization of their findings to other settings and groups, these initial findings are at least in agreement with other literature on instructional methods that integrate data from classroom assessments with instructional decision making.

Linguistic demands in instruction. Many ELLs receive most, if not all, of their instruction in English. This imposes an apparent limitation on the extent to which ELLs have the opportunity to learn new skills and information. Abedi and Herman (2010) reported on the results of a study on the extent to which students in the 8th grade in California received opportunities to learn that varied by language proficiency status. The data they obtained, which consisted, among other things, self reports of opportunity to learn from 602 students in 24 classrooms, suggested that students who are ELL tended to have fewer opportunities to learn per unit of instructional time than their NES peers. Examination of achievement test scores suggested that perceived opportunity to learn was positively associated with learning. In addition to the barriers to learning imposed by lack of proficiency in academic English, some educational scientists have concluded that native language proficiency is strongly associated with subsequent learning (Cummins, 1979; Clarkson, 1992).

Provision and understanding of feedback. Specific research on the comprehensibility of feedback, or on variability in learners' potentials to elicit feedback, was not identified for this review. This seems to be a problem with some logical appeal, but it is unclear, at this time, whether there is evidence to confirm or disconfirm it. Still, there is sufficient reason for consideration of this possibility in research on formative assessment for ELLs. Differences in receptive language may play a role; if the English used in math test items can be problematic, then the English used in instruction and immediate corrective feedback can be problematic. Expressive language could also play a role. For example, consider situations in which a student does not understand a calculation or problem solving procedure, but is not able to formulate questions. Alternatively, a student may not feel confident enough with his or her language skills to ask questions or seek assistance.

Linguistic issues in assessment. First, the adequacy of current technology-enhanced formative assessment systems for students with limited English proficiency should be examined. Assessment of skills in math should not be dependent upon a student's proficiency with written English, or dependent upon culturally specific knowledge. Therefore, many students with deficits in reading or language skills take standardized math tests with a variety of accommodations ranging from extended time to having portions of the test read out loud to them. The same issue will be true in the context of formative classroom assessment: the instruments used can still rely to some extent upon how a student's linguistic skills grant or deny access to probes and items. In mathematics, this could be an issue for assessment of concepts and applications more so than for computation.

For instance, it has been found that linguistic complexity of tests items reduces the reliability of tests for students with limited proficiency in English, thereby invalidating resulting scores in math—particularly math subscales that are highly text based, such as problem solving, rather than those that are primarily digit based, such as computation. Linguistic barriers also limit the number of items attempted and completed by students with limited English proficiency, further detracting from the overall validity of scores obtained (Abedi, Lord, Plummer, 1997). In general, it is thought that linguistic barriers obscure scores on content assessments for ELLs, perhaps artificially lowering scores, and producing irrelevant variance (Abedi, 2003; Abella, Urrutia, & Shneyderman, 2005).

What is not currently known is the extent to which this issue affects the overall usefulness of such systems for ELLs. It might be reasonable to expect that linguistic demands of an instrument used formatively lessen the usefulness of the resulting data. If there is an interaction between the linguistic complexity of an instrument and the linguistic proficiency of a student, then in what ways could the instrument be refined without diminishing the extent to which the data it produces can be used to inform instruction?

Use of formative assessment. Existing conceptualizations of formative classroom assessment generally focus on the effects that the process creates for learners. However, there is at least an implicit assumption that this process can produce change within teacher behaviors as well. By collecting information on student performance teachers modify and individualize instruction, and they provide feedback to students. As students respond with improvement or lack of improvement, teachers' concepts of their own abilities as instructors and their expectations for future performance from individual students could evolve. This introduces an element of variability to the process which occurs between teacher and student. This variability can lead to improved learning, or can fail to provide positive results (from a lack of data, or lack of feedback).

For instance, in a study by Wiliam, Lee, Harrison, and Black (2004), teachers were guided in the development of their own formative classroom assessment plans to be used over the course of a school year. At the end of the study, classroom effect sizes ranged from -0.4 to 1.5 standard deviations. Although this span of results was a consequence of a loosely controlled research design in which teachers in various schools, operating under different contextual and individual parameters, each devised and implemented their own formative classroom assessment plans, the scope of the potential influence of teacher behaviors in classroom assessment was at least partially demonstrated. Providing researcher guidance and rough heuristics for the development of assessment plans does not necessarily guarantee positive results for students in real classrooms. For such practices to become useful in real classrooms, it may be beneficial to examine the extent to which teacher characteristics are associated with implementation fidelity.

Use of technology during instruction. Another potential issue associated with the use of technology enhanced formative assessment for ELLs is that of technology itself. There has been interest in automating various aspects of instruction long before the days of the personal computer. One of the earliest examples of this interest is the device invented by Sidney Pressey, an educational psychologist active in the first half of the previous century. Pressey developed a machine, resembling a typewriter, which was designed to assist students in the acquisition of facts by matching the presentation of stimuli to student performance (Pressey, 1927). The device never achieved popularity, but Pressey’s interest in the creation of “labor-saving devices” continues to be shared by many.

With the accessibility and flexibility of computer technology we currently enjoy, the popularity of educational technology products has increased substantially—for better or worse. In a large federally funded randomized controlled trial, Dynarski and colleagues (2007) sought to document the extent to which the use of instructional technology products impacted learning. The diverse set of computer-based instructional programs used in this study included programs that were described as supplements to the curriculum or as full curricula. After controlling for a variety of contextual variables, the authors concluded that the presence and use of computer-based educational programs was not significantly related to gains in achievement. Although the time teachers and students spent with the "products" was used as the index of implementation; the authors did not discuss the basis on which use of such products would be expected to improve learning, and they did not examine ways in which these products were used in classrooms. If technology truly is an important issue for instructional or assessment research, then the components of instruction or learning that technological tools were designed to impact should be specified and operationally defined.

Findings from an experiment conducted by Ysseldyke and Bolt (2007) on the use of Accelerated Math suggested that teacher implementation of the system—known to moderate the effectiveness of group outcomes—could be different for students from linguistic minorities. In this second large-scale study, classrooms from eight schools in seven states were randomly assigned to treatment or control conditions. As in previous studies on this system, effects were to be compared between treatment and controls, observing patterns gains across levels of achievement and teachers' fidelity of implementation. At the conclusion of the academic year the researchers found that teachers from the treatment condition had excluded 39.5% of their students from the use of Accelerated Math; no patterns in exclusion from treatment were evident. Before an automated progress monitoring system can be considered effective for students who are

ELLs, the extent to which teachers will implement with fidelity and how ELLs, as a group, respond to such instruction must be determined.

Technical adequacy. To what extent are the data produced by such systems technically adequate for ELLs? This question is related to the issue of linguistic complexity in the instruments used, but also covers other potential issues such as comparative reliability between groups, or the extent to which such data can be used to predict progress toward local benchmark tests or annual state tests. Decision making based on progress monitoring data is most effective when those data are adequately reliable and valid.

Gersten, Keating, and Irven (1995) discuss validity of classroom assessments on the basis of how the data are used. According to findings from multiple researchers, both expected and unexpected findings, there should be strong concerns about validity of classroom assessment data. Tools used in the curriculum-based family of assessment frameworks have been researched fairly thoroughly, and their psychometric properties (as they apply to students in the general population) are known in some detail. However, the same cannot be said for students who are ELLs. This has profound implications for the extent to which existing methodologies can be applied successfully. At this point, we do not know how the psychometric properties of commonly used instruments change or stay the same for ELLs. Current methods in progress monitoring math performance have been shown to have adequate technical qualities (Thurber, Shinn, & Smolkowski, 2002) but, as was shown from the comprehensive work of Abedi and colleagues in the realm of large scale assessment in math, we know that the same instruments cannot be expected to exhibit the same technical qualities for students with limitations in academic English proficiency

Chapter 3: Study Design &Methods

This study focused on the use and effectiveness of technology-enhanced formative assessment in mathematics for English language learners. The goals of this research were twofold: to examine formative assessment practices for ELLs, and to examine differences in growth of mathematics skills for those students versus NES students. By focusing on the sub-population of ELLs, who share only one common characteristic (limited English proficiency), and by focusing on mathematics (a group of skills that are distal tolanguage skills), this study also provided an examination of the potential role of language in the process of instruction, assessment, and learning.

Research Questions

The following research questions were addressed in this study:

1. To what extent is technology-enhanced formative assessment implemented differently for ELLs than for NESs?
2. Do ELLs and NESs in the same grade begin a school year with equivalent levels of mathematics skills?
3. Do ELLs and NESs exhibit similar growth in mathematics skills over time?
4. Do ELLs and NESs respond differently to formative assessment in mathematics?
5. What were teachers' of ELLs perceptions of barriers to implementation, as well as relative strengths and weaknesses of the technology-enhanced formative assessment system?

Design and Participants

This study involved analysis of a longitudinal and cross-sectional database of students who attended schools that had purchased and used Accelerated Math (AM) to facilitate teachers' formative assessment of their students' mathematics skills. Student data were provided by Renaissance Learning, the publisher of Accelerated Math, from the 2009-2010 school year, including records from the Accelerated Math hosted database, plus STAR Math assessment data across the entire school year. The data were selected from schools that listed students as ELL in their demographic data, and from schools in which each student's grade level had been identified. The entire database included Accelerated Math and STAR Math records from 51,130 students from 2867 classrooms, and 149 Schools in 32 states in the U.S. The number of students from each state within each grade included in the sample is displayed in Table 1.

Table 1: Numbers of students by state

	Grade								
State	1	2	3	4	5	6	7	8	Total
California	398	667	603	809	658	146	82	63	3,426
Mississippi	292	384	566	460	610	159	154	198	2,823
Texas	182	189	433	631	323	255	317	161	2,491
Nevada	0	54	117	106	95	405	382	227	1,386
North Carolina	296	75	127	92	170	130	123	102	1,115
Michigan	153	172	161	140	135	117	7	7	892
Louisiana	19	86	176	51	212	52	93	87	776
Arizona	58	80	170	81	161	93	53	44	740
New Mexico	105	61	124	62	72	48	99	74	645
Wisconsin	13	22	50	123	93	82	67	77	527
Arkansas	0	0	0	246	73	0	0	0	319
Kansas	54	145	109	1	0	0	0	0	309
Colorado	0	1	9	2	1	0	136	129	278

North Dakota	35	30	30	31	27	38	38	38	267
Florida	0	0	94	121	49	0	0	0	264
Illinois	0	0	1	190	39	0	0	0	230
Minnesota	1	0	45	54	71	56	0	0	227
Oregon	4	39	56	43	42	43	0	0	227
Indiana	4	0	4	177	38	0	1	2	226
New Jersey	0	43	22	47	1	52	6	52	223
Tennessee	0	0	198	11	10	0	0	0	219
Wyoming	0	53	39	39	64	0	0	0	195
New York	0	0	0	100	62	17	14	1	194
Hawaii	0	0	0	11	4	33	96	46	190
Oklahoma	7	39	0	39	30	45	11	3	174
Washington	3	10	1	10	5	25	31	16	101
Northern Mariana Islands	0	0	36	0	0	24	0	0	60
Iowa	3	0	3	4	2	2	0	0	14
Missouri	0	0	3	2	0	0	0	1	6
Connecticut	0	0	0	0	0	2	0	2	4
Massachusetts	0	0	1	0	0	0	0	0	1
Total	1,627	2,150	3,178	3,683	3,047	1,824	1,710	1,330	18,549

The data provided by Renaissance Learning were delivered in relational format, in which records of Accelerated Math use, individual STAR Math assessments, information about individual students, and information about individual schools were sent in separate files. Each of the files was imported into a Microsoft Access database table; relationships between the tables were defined according to several key variables. Each school identifier (SchoolID) was associated with at least one class identifier (ClassID), which was associated with at least one student identifier (StudentID). Receiving the data in relational format facilitated data cleaning and specification of further criteria for records to be included in the analysis dataset. Renaissance Learning did not provide any information that could be used to identify actual students or schools beyond the names of the states from which the data were collected.

A subset of the entire database was selected according to the following criteria: (1) records from areas outside of the United States were omitted from analyses; (2) data were included from students in grades 1 through 8 because of the limited number of ELLs in grades 9 through 12; (3) students were required to have at least 2 STAR Math assessments during the school year; (4) students were required to have come from a school in which Accelerated Math had been purchased for use—independent of the extent to which implementation of the system actually occurred in classrooms; (5) students were required to be associated with only one classroom unique identifier. The final sample included 18,549 students from 1,401 classrooms within 123 separate schools spanning grades 1 through 8. Of these students, 2,057 were listed as ELLs.

English language learners. For analyses on implementation and gain to be conducted with confidence, it was first necessary to characterize the group of students in the dataset listed as ELLs. The data files provided by the company arrived with two binary indicators, bLEP (binary variable, Limited English Proficiency) and bELL (binary variable, English Language Learner). The response to a query from the author to Renaissance Learning staff indicated that there should be no meaningful differences between these indicators. Furthermore, there did not appear to be substantial overlap in the dataset between students listed under one or the other variable. All students with a “1” rather than a “0” under bLEP and bELL were included as ELLs for the purposes of this study. Data summarizing the numbers of students in the bELL and bLEP group are shown in Table 2.

Table 2: bELL and bLEP

		bLEP		
		No	Yes	Total
bELL	No	16505	545	17050
	Yes	493	1019	1512
	Total	16998	1564	18562

To further characterize this group, several demographic and academic statistics were examined. ELLs in this dataset appear across states and within ethnicity categories at rates similar to those reported in recent comprehensive demographic reports of ELLs in US schools. The percentage of ELLs within each classroom was highly positively skewed: most classrooms in the sample did not include any ELLs. The average percentage of ELLs within classrooms was about 0.09%, and the maximum was 100%. Table 3 displays data on the percent of ELLs recorded in the classrooms in this sample.

Table 3: Clustering of ELLs within Classrooms.

Percent ELL	Classes
0 to 10	1126

10 to 20	48
20 to 30	51
30 to 40	37
40 to 50	31
50 to 60	34
60 to 70	27
70 to 80	17
80 to 90	8
90 to 100	22
Grand Total	1401

Demographic characteristics of the dataset are illustrated in Table 4, which displays the percent of students per demographic group (such as ELL or NES) within each grade level, and in the sample overall.

Table 4: Student Demographics

Demographic Categories		Grade								
		1	2	3	4	5	6	7	8	Average
Language Status	NES	86.61%	83.77%	87.67%	88.53%	86.56%	92.88%	94.33%	97.14%	88.92%
	ELL	13.39%	16.23%	12.33%	11.47%	13.44%	7.12%	5.67%	2.86%	11.08%
Sex	Female	47.33%	48.65%	45.53%	42.33%	45.72%	30.21%	31.93%	35.86%	41.99%
	Male	51.81%	47.30%	48.14%	42.41%	47.26%	31.52%	30.70%	37.22%	43.06%
	Unknown	0.86%	4.05%	6.32%	15.26%	7.02%	38.27%	37.37%	26.92%	14.95%
Ethnicity	American Indian or AK Native	0.80%	0.60%	0.79%	0.38%	0.33%	0.27%	0.18%	0.38%	0.47%
	Asian or Pacific Islander	0.74%	1.16%	2.48%	1.33%	1.54%	2.47%	0.18%	1.58%	1.51%
	Black	28.62%	23.67%	22.61%	17.03%	19.41%	8.11%	14.44%	12.48%	18.72%
	Hispanic	35.63%	31.77%	25.16%	26.17%	26.49%	14.14%	15.84%	16.54%	24.70%
	White	19.96%	24.00%	21.45%	14.15%	19.54%	11.67%	5.55%	17.22%	17.12%
	Unknown	14.25%	18.79%	27.52%	40.94%	32.69%	63.34%	63.82%	51.80%	37.47%
SES	Free Lunch	8.85%	7.91%	5.76%	7.74%	12.96%	11.62%	7.19%	3.46%	8.40%
	Reduced Price	1.11%	1.16%	0.98%	1.00%	1.38%	1.10%	0.47%	0.15%	0.99%

	Full Price	90.04%	90.93%	93.27%	91.26%	85.66%	87.28%	92.34%	96.39%	90.61%
Academic Status	General Education	97.91%	92.70%	93.11%	92.34%	91.01%	97.15%	98.01%	97.52%	94.15%
	Special Education	2.09%	2.28%	2.05%	2.36%	3.28%	1.59%	0.82%	2.03%	2.18%
	Gifted and Talented	0.00%	5.02%	4.85%	5.29%	5.71%	1.26%	1.17%	0.45%	3.67%

Two separate datasets were produced from this final sample of students. The first dataset was created in the "short" format that included one row per student. This dataset consisted of a summary of Accelerated Math records (libraries, problems attempted and correct, the counts of objectives mastered, first and last dates of use of Accelerated Math), as well as demographic information. The second dataset was created in the "long" format, in which a single student may have multiple rows of data, depending on the number of STAR Math tests administered. This longitudinal dataset included information from each individual STAR Math assessment for each student over the school year (such as score, percentile, and date), plus Accelerated Math summary information, which were repeated across rows as necessary.

The second dataset included longitudinal STAR Math records from a 41 week time window. Records were filtered by first and last STAR Math assessment dates for each student; records that occurred before September 1, 2009, and after June 1, 2010, were excluded from the final sample. Time in this dataset was represented by a continuous variable called Week, which was comprised of integers that ranged from 0 to 40, corresponding with the order of the week within the defined range. Visual examination of test administration over time suggested a typical school year, with noticeable breaks at the beginning of a typical school year in the fall, near the end of the calendar year (winter break), and again at the end of the typical nine-month school year in the spring of 2010.

Technology-Enhanced Formative Assessment System

Accelerated Math. Accelerated Math (Renaissance Learning, 2009) is the formative assessment system that produced the data analyzed in this study. The logistical requirements of formative assessment can be prohibitive without the use of computers to automate data collection and management. Although is not the only product available to educators at this time, AM was selected because it has been researched thoroughly, includes features that allow for the examination of multiple components of formative evaluation, and is used commonly enough that analyses including sufficient numbers of ELLs can be conducted.

Use of AM is a process of repeated assessments and instructional decisions. Ideally, the process begins with an administration of STAR Math, which is intended to help teachers determine their students' instructional levels, and to help evaluate improvement in mathematics skills over time. AM is based on libraries of mathematics assessment items. These items are assigned as practice items, test items, exercise items, or diagnostic items. Teachers are intended to use information from their students' STAR Math assessments to assign each student to an AM library that would provide an appropriate level of challenge. Within each library, students are assigned sets of practice items that belong to a series of objectives. Practice assignments (containing the practice items) are assigned by teachers—not assigned automatically by the software. Practices are printed on to paper worksheets that include a series of multiple-choice practice items. Once a student completed specific practice items within one objective, he or she transcribes responses onto a sheet readable by an optical scanner connected to a computer. Immediate feedback can be produced immediately and reviewed by instructors and students.

If a student responds to at least 75% of the practice items for an objective correctly, that student may be flagged (for the instructor) as "ready to test". The teacher assigns and prints out a test for the student to take. To pass the test, the student must respond to at least 85% of the items correctly. After passing the test the student has mastered the objective, and may be assigned new activities from the next objective in the library. When a student does not pass the test, the teacher may choose to assign the student ‘exercise’ problems—intended as additional opportunities to respond and to receive corrective feedback—or diagnostic items, which can be used to examine specific sub-skill deficits to identify more precisely the student’s instructional needs. Once a student passes exercise or diagnostic problems, they may take the test again.

The final result of this series of activities in AM is the mastery of an objective. Mastery of an objective is evidence that some, but perhaps not all, components of formative evaluation occurred during instruction. It is an indication that a teacher administered, at a minimum, test items to the student, and engaged in more than summative use of assessment data. Similarly, the number of libraries completed by students can be seen as products of implementation, although on a larger scale and without as much variation. Most students work within a single library during the school year; a small group completes objectives in multiple libraries. Earlier studies conducted on the use of AM report that in a given school year, and with normal implementation of the system, students may be expected to complete around 190 objectives within a school year, which would be about five objectives per week in a school year of nine months.

STAR Math. Data generated by administrations of STAR Math 4.0 (Renaissance Learning, 2009) were used in this study as a response variable in models constructed to evaluate effects of implementation of AM. STAR Math was designed to allow educators to identify students’ instructional levels reliably and accurately through the use of computer adaptive testing. Computer adaptive testing is a method of assessment in which items are selected on the basis of correct or incorrect prior responses such that items administered are matched with the skill level of the student, resulting in an assessment with increased sensitivity to individual students' instructional levels within math curricula.

This test was designed to be administered to students in the first through twelfth grades, and can be administered individually or to whole groups. Because the version of STAR Math that produced the data used in this study was based on a bank of items that had been vertically scaled, individuals' or groups of students' scores can be compared across grade levels and time. Each administration of STAR Math consists of 24 items selected from a pool of 1974 items developed under the framework of item response theory, and requires roughly 15 minutes for completion. Scaled scores may range from 1 to 1400. Students who participated in this study took STAR Math at least twice: once near the beginning of the school year (set to September 1 for this study) and once again at a later time.

STAR Math 4.0 still used the normative data from the older 2.0 version. Norms for that version were obtained in 1998 from 7517 students randomly selected from 399 schools in various regions of the U.S. Estimates of the split half reliability for STAR Math 4.0 across the first through twelfth grades ranged from 0.777 to 0.882, and 0.944 overall. Estimates of alternate forms reliability ranged from 0.721 to 0.799 across grades, and 0.908 overall. Evidence of the validity of STAR Math was obtained in a series of studies correlating student performance on a variety of tests including, but not limited to, the Iowa Test of Basic Skills, the Stanford Achievement Test, and the Terra Nova. Correlations observed consistently clustered around a range of .60 and .80.

Quantitative Analyses

Mixed effects regression. Research questions 1 through 4, on differences between implementation of AM and growth in mathematics between groups, were addressed through the use of mixed effects regression. Mixed effects regression (MER) allows for the development of models based on clustered data, such as students within classrooms, or test scores within students over time. MER is also useful for constructing statistical models based on unbalanced data, such as unequal numbers of students within clusters (classrooms), unequal numbers of measurements for individual students over time, or unequal numbers of students in different groups (such as grades or language status).

MER is based on a set of assumptions similar to those used in traditional, single-level regression, but modified to accommodate more complex data structures. Individual observations may be assumed to depend on a higher order clustering variable, such as a variable that identifies classrooms or teachers. Response variables were assumed to follow a specific distributional form (for these analyses, normal or Poisson distributions), and were assumed to have homogenous and normally distributed random effects and residuals.

MER differs from ordinary regression in two important ways. First, the model parameters are often estimated with likelihood functions rather than the traditional least-squares method; for applied research this distinction is important only because it has implications for how individual models can be evaluated and interpreted. The second, and more salient difference, is in the random or *stochastic* portion of the regression formula, which is expanded to include random effects in addition to the overall error term. Random effects can be understood as residual error terms that have been reserved for specific groups of data included in the analysis. For example, in educational research using mixed-effects regression the variable used to identify classrooms often receives a random effect term. The use of such an effect allows models to account for variability due to dependence (clustering), which can help improve the precision of parameter estimates.

Fixed effects can be understood as the usual set of predictors included in a multiple regression equation. Fixed effects in MER can be baseline parameter estimates that represent a referent category, and parameter estimates that are differences from that referent category. These fixed effect parameters can be used to describe, or even plot, slopes and intercepts for groups. For example, a fixed effect b_0 might describe the intercept for students in the first grade (if grade = 1 is used as a referent category), fixed effect $b_1$would be the difference between second grade and first grade intercepts, and fixed effect $b_2$would be the difference between third and first grade intercepts.

Equations 1 and 2 demonstrate a basic mixed-effects model in which students are nested within classrooms. The same model is specified in two formats to increase clarity. Equation 1 demonstrates the hierarchal aspect of mixed effects modeling by specifying sets of parameters at two different levels. This illustrates how higher-order parameters can be viewed as outcomes of their own regression equations. The predicted value of y for student i in classroom j is a function of an intercept term b_0 and a slope term b_1, and the residual error term that represents the difference between the predicted and observed score.

Equation 1

$$\widehat{y_{ij}} = b_{0i} + b_{1i}(Slope) + e_{ij}$$

$$b_{0i} = \gamma_{00} + \gamma_{01}(ELL) + \zeta_{0j}$$

$$b_{1i} = \gamma_{10} + \gamma_{11}(ELL) + \zeta_{1j}$$

Equation 2 is the same model, but written in a composite format. Parentheses are used to indicate fixed effects, and brackets to indicate the stochastic portion, including random effects and the overall residual error term.

Equation 2

$$\widehat{y_{ij}} = (\gamma_{00} + \gamma_{01}(ELL)) + Slope(\gamma_{10} + \gamma_{11}(ELL)) + [\zeta_{0j} + \zeta_{1j} + e_{ij}]$$

All analyses were conducted in the software package called R, version 2.12.1 (The R Foundation for Statistical Computing, 2010). Formulation and evaluation of all models was conducted using the “lmer” function within the lme4 package (Bates, 2011). The title “lmer” is an acronym that stands for “linear mixed effects modeling”. The lmer function was designed to fit linear and non-linear models to data with a variety of distributional forms.

Analyzing implementation of formative assessment between ELLs and NESs. It was assumed, for the purpose of this study, that an item attempted by a student could also be understood to be an item assigned by a teacher. Evidence in support of this assumption was found in the way classrooms appeared to use types of items. The extent to which each major component of AM (practice items, test items, exercise items, diagnostic items, and objectives) was used for each student was analyzed. Descriptions of the frequency at which the components of AM were used for ELLs and NESs were calculated and are summarized in Chapter 4. Statistical analyses on differences between groups were conducted on each component separately.

Generalized linear mixed-effects models. Each of the five variables listed above were modeled as a function of duration of use, grade level, and ELL status. These variables were assumed to represent discrete *counts* of events. Count data are understood to follow a Poisson distribution. Unlike the normal distribution, which is defined by a mean and a standard deviation, the Poisson distribution is defined only by the mean. This is because the lower tail of the Poisson distribution is set to zero, and dispersion, or variability, depends on the distance of the mean of the count from zero. Because of this the variance and the mean of the Poisson distribution are expected to be equal.

Variables that follow non-normal distributions can be modeled with *generalized* linear modeling. A generalized linear model is similar to a normal linear model in that it includes a response variable, predictors, and an error term. Generalized linear models also include a link function that serves to transform the set of predictors to match the assumed distribution of the response variable. The natural log link was specified for the models on Poisson distributed count data in this study.

Earlier studies on the effects of technology-enhanced formative assessment in classrooms consistently find that much of the variation in results can be explained by ways in which the system was or was not implemented (Ysseldyke & Bolt, 2007; Bolt, Ysseldyke, & Patterson, 2010). To account for the probability that the counts of items attempted might be highly correlated with classroom membership, a random effect for

ClassID intercept was added to each model. These random effects were assumed to have means equal to zero, and normally distributed variances.

Depending on the distribution of the response variable, and the link function selected, generalized linear mixed-effects models can use a range of methods of parameter estimation, aside from the maximum likelihood function (described under “Growth Models”). The five models of implementation analyzed here were fit with the Laplace approximation. The Laplace approximation, similar to maximum likelihood estimation, is an iterative method of parameter estimation useful in analyses of Poisson-distributed data (Bolker et al., 2008).

In social sciences it is not uncommon for count data to exhibit variance greater than the mean (in other words, greater variation than would be expected, assuming Poisson-distributed data). This is known as overdispersion. Unless it is addressed in a model, overdispersion can artificially decrease standard error estimates for predictors, which increases the chance of Type 1 error. In single-level generalized linear models, overdispersion may be addressed through the calculation of an overdispersion factor; each predictor’s standard error may be multiplied by this factor to correct for overdispersion and reduce Type 1 error (Gelman & Hill, 2007).

However, the same method is not advised for mixed effects models. Following the recommendations of several authors, each implementation model was fit with a data-level random effect (a random effect for StudentID). This method, though somewhat controversial, is currently understood to be an appropriate way of accounting for overdispersion in nested count data (Elston, Moss, Boulinier, Arrowsmith, & Lambin, 2001; Gelman & Hill, 2007). To examine the suitability of this method of correction, models of implementation were fit with and without a data-level random effect, and then compared for differences in standard error. Because the results of the models fit with a data-level random effect demonstrated more conservative parameter estimates (larger standard errors for predictors), only those results are reported in Chapter 4.

Predictors of implementation. Each index of implementation was modeled on the same set of four predictors. The first predictor, called Weeks, was included to control for length of implementation, which could reflect many factors irrelevant to the study (student moved into or out of district, etc.). It was calculated by subtracting the date of the first recorded use of AM for each student from the date of the last recorded use.

Past research provided information to suggest that implementation of this product might vary substantially by grade level, with lesser implementation for later grade levels. Given this expectation, a variable for grade level was entered immediately following Weeks. Initial model fits produced non-trivial colinearity between Grade and other predictors. To address this issue, Grade was centered on the grand mean grade level, and relabeled as cGrade.

A student’s linguistic group status was indicated by a binary variable in which NES was coded with “0”, and ELL was coded with “1”. An interaction between ELL and Weeks was added to each of the five implementation models to help evaluate potential differences in *rate* of implementation in addition to overall amount. Variations in the rate of implementation could be taken as indications of consistency or fidelity of AM use (or a lack thereof).Equation 3 displays the model fit for each of the five implementation indices.

Equation 3

$$\widehat{y_{ij}} = \gamma_{00} + \gamma_{01}(Weeks) + \gamma_{02}(cGrade) + \gamma_{03}(ELL) + \gamma_{04}(ELL * Weeks) + \zeta_{0i} + \zeta_{0j} + \varepsilon_{ij}$$

Parameter estimation. It should be noted that the overall fit of each model of implementation could not be compared to a standard point of reference because these models were fit by the Laplace approximation, which does not allow for the types of comparisons commonly used in MER. This is also true because the focus of the analyses of implementation was on differences between ELLs and NESs for each index of implementation after controlling for other predictors. The same set of predictors was used for each model, so only five models are reported. The goal was not to evaluate competing sets of predictors for a single response variable, but to evaluate the equivalence of implementation across response variables between ELLs and NESs with the same set of predictors.

Analyses of growth over time. Growth models were built according to a stepwise forward selection process (Peugh, 2010; Singer & Willett, 2003). In this process, individual predictors are added to a model incrementally. This produces sets of nested models, ranging from a fully unconditional model with no predictors to the full model, including all predictors relevant to the research question. The inclusion of predictors, and the order in which each was introduced to a model, was guided by the specific research questions specified.

Model parameters were estimated by full maximum likelihood estimation (ML). Maximum likelihood estimation is a method in which potential values are created for each parameter specified in a model, and then compared against the observed data across a series of iterations. Iterations of the maximum likelihood function stop when a specified criterion is reached, and estimates of deviance—the extent to which the proposed values deviate from observed data—exhibit minimal decreases in size. The step-wise forward selection method for model development has been recommended for analyses based on ML parameter estimation.

A set of eight mixed effects models were constructed to examine change in STAR Math scaled scores over the course of the 2009-2010 school year. The metric used for time in the longitudinal analyses was numeric variable called Week, which has a value of 0 for the week of September 1, 2009, through Week 40, at the end of May 2010.This metric—the number of week between 0 and 40—was selected to maximize sensitivity to changes over time in STAR Math scaled scores. Students were tested a varying number of times, and on a varying set of weeks. On average, there were just over two measurement records per student out of the span of 40 weeks in which students could have been assessed.

The lack of correspondence between the number and timing of measurements is known as imbalance—a issue that can be problematic for longitudinal studies. Mixed effects modeling takes advantage of the hierarchical structure of the dataset to account for this imbalance via “partial pooling” (Gelman& Hill, 2007). The primary consequence of this imbalance over time was that most students contributed little information to the overall model. The relatively large size of the sample included in this study helped mitigate the extent to which imbalance was a barrier to analysis.

The functional form of the growth curve was assumed to be linear. This was determined by visual analysis of the growth patterns within each grade level, and by comparisons of preliminary models specifying linear or quadratic growth. This decision was also influenced by the observation that the students and teachers in this sample did not use AM consistently across the school year. Based on these observations, linearity was selected as the most conservative form of growth.

Unconditional models. The first model fit was an unconditional means model that included random effects for student and classroom, and no fixed effects. After the fully unconditional model, an unconditional growth model was specified that contained a fixed effect for Week, and random effects for within-student intercept and slope (growth over Weeks). The unconditional means model provided a baseline against which the unconditional growth model could be compared. This helped establish that growth was significantly greater than zero in this dataset, and could therefore be modeled. The unconditional growth model served as a baseline to which other models could be compared to evaluate model fits.

Grade. Singer and Willett (2003) recommend that predictors that are directly related to the research questions be added first, and that predictors, or covariates, included as methods of statistical control should be added to later model fits. After Week was added to the unconditional growth model, the first predictor to be added was cGrade (grade level, centered). The grand mean of grade in the ‘short’ dataset was subtracted from each value of Grade in the ‘long’ format dataset. This was done to reduce colinearity present between the Grade variable and STAR Math scaled scores.

Grade level was not related directly to the research questions of this study, but because of the nature of STAR Math scaled score, and because grade level is known to be associated with differences in implementation of AM and skill growth, cGrade was added as the first covariate. Two parameters were added: grade as a predictor of the Level 1 intercept, and grade as a predictor of Level 1 slope. This facilitated interpretation of subsequent estimates: failure to account for grade level would produce results that would be difficult to interpret.

ELL. ELL status was added after cGrade, with an estimate of difference from the Level 1 intercept, and an estimate of difference from the Level 1 slope, or time in weeks. ELL status was coded as “0” or “1”, where “0” = False, or native-English speaking student, and “1” = True, or ELL. Parameter estimates for the fixed effects associated with ELL represent the *difference* in intercept and growth from NESs.

Implementation. The number of objectives mastered by each student in AM was used as the index of implementation in this study. Using this variable as an index of implementation assumes that teachers assigned, at a minimum, test problems; if students and assigned tasks were matched appropriately, and if results were used by teachers to provide feedback or adjust instruction, then students should master more objectives. This is also problematic: while a high number of objectives mastered suggests a greater use of formative assessment by teachers and students, it also represents gains in mathematical skill, which is what STAR Math was intended to assess over time. This calls attention to the importance of independent examination of other indices of implementation for the interpretation of the results of growth models.

Because the raw count of objectives mastered by each student was distributed non-normally, it was not appropriate to include in models without transformation. The number of objectives mastered by each student was split into an ordinal variable with nine groups, dummy-coded 0 through 8. The numbers of students included within each level of this variable are displayed in Table 5.

Table 5: Partitioning of Objectives Mastered

	Objectives Mastered		
Group	Minimum	Maximum	Students in Group
0	0	0	4,760
1	1	3	988
2	4	8	1,816
3	9	15	1,750
4	16	25	1,859
5	26	39	1,842
6	40	57	1,842
7	58	89	1,856
8	90	424	1,836
Total			18,549

The full model fit included intercept and growth estimates for cGrade, ELL, and Objectives, plus a three-way interaction between ELL, Objectives, and Week, and each relevant two-way interaction that was included in the previous model (ELL by Week, Objectives by Week, and the ELL by Objectives intercept term). Equation 4 displays the full model in the sub-models format.

Equation 4

$$\widehat{SS_{ijk}} = b_0 + b_1(Week) + \zeta_{2k} + \varepsilon_{ijk}$$

$$b_0 = \gamma_0 + \gamma_1(cGrade) + \gamma_2(ELL) + \gamma_3(Objectives) + \gamma_4(ELL * Objectives) + \zeta_{0j}$$

$$b_1 = \gamma_5 + \gamma_6(cGrade) + \gamma_7(ELL) + \gamma_8(Objectives) + \gamma_9(ELL * Objectives) + \zeta_{1j}$$

Model comparison. Models were evaluated through comparisons of the Akaike Information Criterion (AIC; Akaike, 1973), which is equal to 2*K* + *Deviance*, where *K* is the number of fixed and random effects included in the model. Multiple AIC values can be compared when evaluating predictive quality. Smaller AIC values indicate better representations of the original dataset, as long as the models are based on the same response variable, include the same participants, and were produced under the same method of parameter estimation. The AIC is not a standardized value, meaning that a single AIC cannot be interpreted in isolation—it is not bounded by any set values, such as the usual R^2, which is bounded by 0 and 1. The difference between two AIC values, ΔAIC, is used to evaluate the relative strengths of a set of models. The model with the best fit to the observed data has a ΔAIC equal to 0, others may have ΔAIC values greater than or equal to 0.

Comparison of the differences between model likelihoods can be used to order models from best to worst fit, but is less useful in communicating the extents to which individual models or predictors improved model qualities. Because models fit were nested, the significance of each predictor was evaluated with the Likelihood Ratio Test (LRT). The LRT is based on a chi-squared statistic of the difference between the deviance of a simpler model (a reduced model), and a more complex model (a full model). The resulting statistics and *p*-values were used to evaluate the significance of the contributions of individual predictors included in the full model.

Teacher Perceptions of Implementation and Effects of Technology-Enhanced Formative Assessment

To help contextualize quantitative results and to provide some guidance for further research, several teachers who use Accelerated math were interviewed by the researcher. The interview was semi-formal, and was conducted via telephone. Actual teacher and school names are not reported. Teachers and schools were instead given pseudonyms.

The specific set of questions covered in this interview is displayed in Table 6. The first, second, and fourth questions were intended to gather information about each teacher's perception and use of AM. The third and fifth questions seek information about each teacher's experience in providing instruction to ELLs, and about his or her use of AM with ELLs, plus any perceived differences in the outcomes of the use of AM. The sixth question was intended to gather input on teacher perceptions of the accessibility of AM for students in their classrooms. The seventh and eighth questions were intended to promote reflection about ways that use of AM has or has not impacted the way each respondent teaches.

Table 6: Questions for Teachers who Use Accelerated Math for English Language Learners

1. How long have you used Accelerated Math?
2. What is your impression of Accelerated Math as a way to assess the math skills your students have learned?
3. What is your experience in providing math instruction to students classified as English language learners?
4. The basic features of Accelerated Math include practice problems, test problems,

exercise problems, diagnostic problems and libraries of math objectives.

4a. How often do you use each of these components in your classroom?

4b. What are some ways in which you believe Accelerated Math should be used differently for English language learners?

5. Compared to your students who are native English speakers,

5a. what are some ways in which the effects of using Accelerated Math are the *same* for your students who are ELLs as your students who are native English speakers?

5b. what are some ways in which the effects of using Accelerated Math are *different* for your students who are ELLs ?

6. To what extent is the content (such as practice or test items) in Accelerated Math accessible for your students who are ELLs? In other words, to what extent does the content present linguistic or cultural barriers to comprehension?

7. What is your opinion on the effectiveness of Accelerated Math for the identification of instructional needs in math for English language learners?

8. In your experience, how has the use of Accelerated Math impacted the amount and frequency of corrective feedback (or reinforcement) you provide your students who are English language learners?

Chapter 4: Results

Research Questions

In this chapter I report the results of the analyses performed for the five research questions. The research questions were:

1. To what extent is technology enhanced formative assessment implemented differently for ELLs than for NESs?
2. Do ELLs and NESs begin a school year with equivalent levels of mathematics skills?
3. Do ELLs and NESs exhibit similar growth in mathematics skill over time?
4. Do ELLs and NESs respond differently to formative assessment in mathematics?
5. What were teachers' of ELLs perceptions of barriers to implementation, as well as relative strengths and weaknesses of the technology-enhanced formative assessment system?

Results are reported under three sections. Descriptive statistics and model fits for used to address the first research question are reported under the header of "Implementation of Accelerated Math". The results of longitudinal models used to address research questions 2, 3 and 4 are reported under the section called "Longitudinal Analyses". Results of the brief, informal interviews with users of AM are summarized by interview question under the section "Teacher Interviews".

Implementation of Accelerated Math

Descriptive Statistics. Means, medians, standard deviations, and minimum and maximum values for AM items attempted between ELLs and NESs are displayed in Table 7.

Table 7: Descriptive Statistics for AM Components

		Item Type					
Group	Statistic	Practice	Test	Diagnostic	Exercise	Objectives	Libraries
NES	Mean	423.36	207.97	159.70	132.13	41.69	1.69
	Standard Deviation	506.12	219.19	174.37	206.38	41.92	1.78
	Minimum	1.00	5.00	5.00	1.00	1.00	1.00
	Median	228.00	125.00	100.00	54.00	28.00	1.00
	Maximum	5,456.00	1,735.00	1,690.00	2,636.00	424.00	19.00
ELL	Mean	561.29	259.98	176.60	233.33	46.95	1.48
	Standard Deviation	597.78	243.59	197.11	273.28	41.43	1.00
	Minimum	6.00	4.00	5.00	2.00	1.00	1.00
	Median	339.00	190.00	120.00	132.00	36.00	1.00
	Maximum	3,289.00	2,335.00	1,455.00	1,827.00	251.00	7.00

The difference between the level of the minimum library worked and the initial STAR Math grade level recommendation was calculated for each student. Values greater than 0 indicated that a student might have been working above his or her skill level; values less than 0 indicated that a student might have been working below his or her skill level. The percent of each group at each level of difference is listed in Table 8. In this dataset the greatest differences were either four library grades above the student’s grade level, or up to eight library grade levels below a student’s grade level. Between 80% and 90% of students in either language group had a minimum library level that was either at their own grade level, or one level above their own. For example, this suggests that most fifth graders worked on material that, at a minimum, tended to be fifth grade level material, or perhaps fourth grade level material.

Table 8: Differences between Minimum AM Library and Grade Level

Level Difference	NES	ELL
-4	0.01%	0.00%
-3	0.01%	0.00%
-2	0.11%	0.05%
-1	1.70%	0.49%
0	63.79%	77.49%
1	21.73%	12.01%
2	7.59%	6.81%
3	3.01%	2.63%
4	1.49%	0.49%
5	0.41%	0.05%
6	0.07%	0.00%
7	0.04%	0.00%
8	0.03%	0.00%

Generalized linear mixed-effects models. To help evaluate the importance of classroom-level effects on the implementation of AM between groups of students, five models were fit that included random effects for StudentID (a data-level random effect) and ClassID (a classroom level random effect), following the recommendation of Gelman and Hill (2007). The estimates for the mixed-effects models of the 5 features of implementation studied are listed in Table 9; values for each predictor are listed in columns, according to features of AM listed in rows. Exponentiation of each fixed effect parameter estimate gives the value of the proportional change in the response variable predicted by a one-unit change in the predictor.

The intercept estimates for each response variable suggested implementation of each feature that was significantly greater than 0. Estimates for Weeks, which was the total number of weeks each student used AM, were significant for all five features. For each additional week of AM use, students tended to receive about 5.13% more practices, 4.81% more test items, 3.15% more Diagnostic items, 3.98% more Exercise items, and 3.56% more objectives mastered. Grand-mean centered grade level, cGrade, was a significant predictor of practice item use and diagnostic item use alone, predicting an increase of about 6.72% practice items and about 10.52% diagnostic items per grade level. Students who were ELLs did not appear to receive significantly lesser or greater implementation of AM overall, except for the case of diagnostic items, in which ELLs appeared to receive about 13% less after controlling for classroom level variability. Although ELLs appeared to have diagnostic items assigned at a significantly greater number per week of AM use the size of this difference was negligible.

Table 9: Models of Implementation

Parameter		Model									
		Practice		Test		Diagnostic		Exercise		Objectives	
Intercept	B	3.972	***	3.263	***	3.403	***	2.821	***	1.802	***
	S.E.	0.033		0.038		0.050		0.043		0.035	
Weeks	B	0.050	***	0.047	***	0.031	***	0.039	***	0.045	***
	S.E.	0.001		0.001		0.001		0.001		<0.001	
cGrade	B	0.065	***	0.004		0.100	***	0.006		0.015	
	S.E.	0.012		0.013		0.018		0.016		0.012	
ELL	B	-0.031		0.013		-0.136	**	0.032		-0.075	
	S.E.	0.043		0.065		0.048		0.055		0.055	

ELL*Weeks	B	0.001	0.001	0.004	**	< -0.001	0.002
	S.E.	0.001	0.002	0.001		0.001	0.002

Longitudinal Analyses

Variance decomposition. To determine the extent to which variability in STAR Math scaled scores occurred within students, between students, and between classrooms, a fully unconditional model with random effects for StudentID (to indicate the between-student level) and ClassroomID (to indicate the between-classroom level) was fit on the full set of STAR Math scaled scores, unconditional on time in weeks or any other predictors in the set. The standard deviations associated with random effect estimates for the residuals, StudentID, and ClassID represent the amount of variation observed at each potential level of analysis, including variation between individual measurement, between students, and between classrooms, respectively.

A two-level unconditional means model, with STAR Math assessments nested within students suggested that about 16.6% of the variability in scaled scores was observed within students, whereas about 83.3% of the variability was observed between students (without controlling for grade level). Because a central question of this study focused on the effects of instructional behaviors—specifically those behaviors associated with formative assessment—a second unconditional model was fit that included random effects for ClassID (or teachers) in addition to the random effects for students and the model residuals.

It was assumed, for the purpose of this study, that a classroom identifier was strongly associated with instructional behaviors, if not individual instructors. Results of the three-level unconditional model indicated that about 16.3% of the variance in STAR

Math scores was observed within students, about 15.9% of the variance was observed between students, and the remainder of the variance, about 67.8%, was observed between classrooms. A likelihood ratio test between the two unconditional models suggested that the three-level model produced a significantly better fit to the data than the two-level model (χ^2=20909.78,df = 1, p< 0.001).

Growth models. A set of seven growth models comprised of three levels (models A through G), and one two-level model (H) was constructed for STAR Math scaled scores, including predictors Week, Grade, ELL, and Objectives Mastered. Information about the model fits and parameter estimates is displayed in Table 10. Each model was based on 53,583 individual measurements from 18,549 individual students in 1,401 classrooms. Following the recommendations of Singer and Willet (2003), models were fit in a step-wise forward selection method starting with an unconditional growth model.

An unconditional means model (model A) included random effects for intercepts (cluster means) at the student level and the classroom level; this provided baseline against which more complex models could be compared. An unconditional growth model (model B) with random effects for within student growth, between student intercept, within class growth, and between class intercepts, was fit to provide a baseline against which subsequent models, including substantive predictors, could be compared. Parameter estimates are displayed in Table 10. The intercept estimate for mean STAR Math score at Week 0, and the estimate for Week were both significantly different from zero. The random effects estimate for between-student growth was almost perfectly correlated with the random effect for between student intercept, in addition to accounting for a minimal amount of the variance; this random effect was excluded from subsequent models. A likelihood ratio test between the three-level unconditional means model and the three-level unconditional growth model indicated a better fit for the unconditional growth model (χ^2= 22,908, df = 3,*p*< 0.001).

AIC values displayed at the bottom of Table 10 demonstrate the level of deviance associated with each model. The largest ΔAIC estimate of 5,215 was observed between model G and the model B, the unconditional growth model. The ΔAIC between model E and model B was equal to 5,175; models D and C had ΔAIC values of 1,329 and 1,319, respectively. Because these values were obtained from set of nested models, ΔAIC values can be used to summarize the contributions of individual parameters. Of the predictors added as fixed effects to these models, Objectives Mastered was associated with the largest overall decrease in deviance.

Model C included grand-mean centered grade level as a predictor of Level 1 intercept and growth. This variable, labeled as cGrade, was produced by subtracting the average grade level from each row of the person, or short format, dataset. This was done to reduce colinearity with other predictors. The value for the level one intercept, representing the average score for all groups at Week 0, lowered slightly, compared to the model B—the unconditional growth model. The estimate of growth in scaled score over time was relatively unchanged. The significant positive estimate for cGrade in the level one intercept indicates the extent to which average scores differed at week 0 between grades. There was a significant negative interaction between cGrade and Week, indicating that there was less growth in STAR Math scores at higher grade levels. The exclusion of a random effect for Week at level two resulted in larger residuals—note that evaluations of the significance of random effects and residuals are not provided because of a lack of consensus on the interpretability of current methods of comparison. Despite a larger residual, model C had an AIC much smaller than model B, the unconditional growth model (χ^2= 1318.63, df = 1,*p*< 0.001).

Model D included the binary predictor called ELL (0 = NES, 1 = ELL), which was added to the level two sub-models for the level one intercept and slope parameters. There was a significant difference between intercept estimates for ELLs and NESs, in which ELLs obtained scores that tended to be lower at the beginning of the year. The 95% confidence interval for the difference between the growth estimates between the two groups contained 0, suggesting no systematic differences in growth in math skills over time between the 2 groups. It may be of interest to note here that the same model, excluding classroom-level random effects, and including a within-student random effect for Week, produced a small but significant difference between groups, in which ELLs saw less growth over time. This is discussed further in chapter five. Model D demonstrated a better fit to the data than model C, suggesting that the addition of ELL improved overall fit (χ^2= 13.845, df = 2, *p* = 0.001).

Model E replaced the level two sub-model predictors ELL and ELL*Week with Objectives, the index of implementation of AM. This model, though not technically “nested” with model D, provided a way to compare the relative contributions of predictors for ELL intercept and growth, and implementation group intercept and growth. Like model D, inclusion of implementation in model E did not produce parameter estimates meaningfully different from those produced in model C, which included only cGrade and cGrade*Week as predictors in the Level 2 sub-models. The parameter estimate for cObjectives suggested that students who received greater implementation of

AM also started at Week 0 with greater math skills. The parameter estimate for cObjectives*Week indicated that students who received greater implementation also exhibited greater growth over time. Model E demonstrated a significantly better fit than model C (χ^2= 3,860.7, df = 2, *p* = 0.0001), as well as Model D (χ^2= 3,846.9, df = 0, *p* = 0.0001).

Model F included the parameters of model E, plus ELL and its interaction with Week. The difference between intercepts for the general population and ELLs still appeared to be significant and negative, and the estimate for the difference between growth for ELLs and the general population did not differ meaningfully from 0. Residual variance appeared to be equivalent in both. Model F demonstrated a significantly better fit to the data than Model E (χ^2= 41.012, df = 2, *p* = 0.001), suggesting that inclusion of both predictors for ELL status and implementation produces a better fit to the data than can be achieved by inclusion of either predictor alone.

A final model, G, included all the fixed and random effects of Model F, plus a three-way interaction term estimating the difference in the effect of implementation over time for ELLs and NESs (cObjectives*Week*ELL). Inclusion of this interaction term did not meaningfully change the estimates for the other parameters. The small, but significant, three-way interaction suggested that ELLs in classrooms in which students received greater implementation of AM received smaller growth estimates than NESs who received comparable levels of implementation, yet still demonstrated greater growth in mathematics skills over time than ELLs who received low, or no implementation of AM. Model G demonstrated a better fit than model F, providing further evidence of the value of including the three-way interaction (χ^2= 6.5, df = 2, *p* = 0.039).

Table 10: STAR Math Growth Models

	Fitted Models											
	B		C		D		E		F		G	
Parameters	Est.	S. E.	Est.	S. E.	Est.	S. E.	Est.	S. E.	Est.	S. E.	Est.	S. E.
Fixed Effects												
Intercept	573.57	4.12	562.04	2.28	562.85	2.30	564.12	2.72	565.27	2.73	565.29	2.72
cGrade			58.86	1.08	58.77	1.08	57.46	1.25	57.33	1.25	57.33	1.25
ELL					-8.25	2.83			-12.00	2.63	-14.67	2.84
Objectives							25.59	0.48	25.68	0.48	25.37	0.50
Objectives:ELL											2.37	0.97
Week	2.11	0.05	2.10	0.04	2.10	0.04	2.09	0.04	2.10	0.04	2.10	0.04
Week:cGrade			-0.40	0.02	-0.40	0.02	-0.39	0.02	-0.39	0.02	-0.39	0.02
Week:ELL					-0.02	0.08			-0.06	0.08	0.01	0.08
Week:Objectives							0.13	0.01	0.13	0.01	0.14	0.01
Week:Objectives:ELL											-0.06	0.03
Random Effects												
L2 Intercept	3,902.20		4,695.24		4,691.58		3,403.91		3,395.51		3,395.89	
L2 Slope	0.08											
L3 Intercept	21,777.00		5,664.74		5,657.18		8,747.82		8,719.22		8,678.51	
L3 Slope	1.61		1.03		1.03		0.97		0.96		0.96	
Residual	2,956.90		2,970.09		2,970.22		2,967.21		2,967.53		2,967.24	
AIC	617,145		615,826		615,816		611,970		611,933		611,930	

Teacher Interviews

Three teachers were interviewed between January and May in 2012. Teacher A was a third grade general education teacher at an elementary school in central California. Teacher B was a seventh grade math teacher at a junior high school in southern Texas, near the national border. Teacher C was a third grade Teacher of English Learners at a school district in Minnesota (with a substantial population of ELLs), who was teaching math to a mixed group of ELL and NES third graders.

Each interview was conducted via telephone. The questions posed to each interviewee were intended to solicit input on issues surrounding the use of AM, as well as perceived barriers, limitations, or benefits for ELLs. The same question wording and order was used in each interview, although questions were clarified upon request, and additional comments or observations were encouraged. Teachers' responses to each of eight questions are summarized below.

Question 1: How long have you used Accelerated Math? The three teachers interviewed represented a wide range of experience in teaching, and in the use of AM. Teacher A reported having taught for 6 years, and having used AM consistently throughout that time. Teacher B reported that, at the time of the interview, she was in her third year of teaching, and she was nearing two years of experience with AM. Teacher C reported having been a teacher for the past 22 years. She was not certain of the exact time when her district purchased AM, but estimated having used AM in her classroom for more than 10 years.

Question 2: What is your impression of Accelerated Math as a way to assess the math skills your students have learned? Each of the three teachers interviewed perceived AM as a tool that promotes assessment of mastery of specific skills, and by extension, individual instructional needs. Teachers B and C went slightly beyond the question of AM as an assessment of what students have learned, and noted that the information they receive from this system provides their students with higher amounts of individualized practice; Teacher C also commented on the flexibility of AM as an assessment, and explained how AM allows her to tailor assessments of specific skills to her curriculum or state standards. Teacher A expressed a broader perspective, and described AM as a useful way to evaluate growth in overall math skills over time (she considered STAR Math interim assessments as part of the AM process).

Question 3: What is your experience in providing math instruction to students classified as English Language Learners? These three teachers reported a wide range of experience in providing math instruction to ELLs. Teacher A reported that she has been teaching math to classes including ELLs periodically throughout her six years of teaching; at the time of her interview, she reported having four ELLs in her classroom. Teacher B, a relatively new teacher in southern Texas, reported that she has always had ELLs in her classroom throughout her three years of teaching. Teacher C reported having been an ELL teacher for 22 years, and that in addition to her role as an ELL teacher, she provided math instruction to a mixed group of students (ELL and NES alike) at the time of her interview.

Question 4a: How often do you use each of these components in your classroom? Questions 4a and 4b were intended to solicit input on overall patterns (or preferences) regarding the use of AM, as well as opinions on the extent to which such a system should be used differently for students who are ELLs.

Teacher A reported that she makes frequent use of each of the major components of AM, and described the process she follows: She assigns objectives based on STAR Math performance, assigns diagnostic items to determine specific instructional needs, and assigns practices and tests accordingly. She reported that her students may master around 75 objectives in one school year. She also noted that a new math curriculum was limiting the extent to which she has been able to use AM at the time of the interview. Teacher C reported using each feature on a regular basis, but did indicate that this is an individual preference, and not necessarily a norm throughout the building. Teacher B did not specify the extent to which she uses each component of AM, but instead described her expectation that her students complete, on average, four objectives per week, or 13 to 16 objectives per month (a relatively high rate of objective mastery, compared to records analyzed in this study).

Question 4b: What are some ways in which you believe Accelerated Math should be used differently for English language learners? When asked about ways in which AM should be used differently for ELLs, each teacher expressed the opinion that AM should not necessarily be used differently for ELLs because of its role as a tool to help them differentiate instruction for small groups and individuals. Teachers A and B both reported that their students who are ELLs encounter greater challenges with AM items that are highly dependent on language and text, and that they encourage group work for this reason; by pairing students with similar language backgrounds, but differing proficiency in English, peers can help each other build math language skills. Teacher C also noted this trend, but explained that her own approach focused more upon careful selection of objectives that appear less biased, and on provision of accommodation (read aloud) on exercises. Teacher C also noted that she finds AM's diagnostic items particularly helpful for her students who are Ells, but that this usefulness has more to do with math skills than language proficiency (students with lower skills, or who experience more difficulty acquiring new math skills).

Question 5a: Compared to your students who are native English speakers, what are some ways in which the effects of using Accelerated Math are the same for ELLs? Questions 5a and 5b were intended to prompt discussion on the ways in which the effects of AM are the same or different for their students who are ELLs or NESs. All three teachers perceived AM as being similarly beneficial for all of their students, regardless of proficiency in English. Teacher A reported having observed similar benefits for both groups of students—that the overall impact of AM on the effectiveness of her instruction does not depend on ELL status. Teacher B reported that AM promotes a great deal of cooperative learning for all of her students, regardless of ELL status, and also observed similar growth in skills over time. Like Teachers A and B, Teacher C believed that AM is equally useful for each of her students and she explained that her instructional strategies and overall expectations do not vary between groups.

Question 5b: What are some ways in which the effects of using Accelerated Math are different for your students who are ELLs? When asked to describe ways in which the effects of AM are different for ELLs, Teachers A and C reported observing no differences whatsoever. Teacher B noted that ELLs have a harder time using contextual clues to determine the meanings of unfamiliar terms, and believes that this can increase the level of challenge or frustration they may experience. She also noted that, although she has not seen average performance on STAR Math improving for either group in the current school year, she suspects some of her students who are ELLs exhibit stronger performance because they have to "put in a little more"—that they are accustomed to using greater effort in completion of academic tasks in general.

Question 6: To what extent is the content (such as practice or test items) in Accelerated Math accessible for your students who are ELLs? Teacher A expressed the opinion that there will always be linguistic barriers to some extent, and that as a teacher, it is up to her to identify these barriers and help her students around them. She noted that she also has the option to print all reports in Spanish. Like Teacher A, Teacher B believed that language differences are inseparable from math instruction and assessment. Teacher B offered the example of the reversed roles of commas and decimal points between Spanish and English (in Spanish, commas separate decimals, and periods separate thousands). Teacher B also noted that similar concepts are expressed differently across both languages (that there are different names for the same things). Teacher C indicated that she screens assignments based on the extent to which she feels each is dependent on English proficiency or specific cultural knowledge, and favors assignment of tasks that make use of pictures. In cases when she expects a student with limited English proficiency to complete work that uses much text, she provides read-aloud accommodation.

Question 7: What is your opinion on the effectiveness of Accelerated Math for the identification of instructional needs in math for English language learners? Questions 7 and 8 were both intended to encourage teachers to think about specific ways in which a student's proficiency in English might interact with different instructional variables, such as instructional match and provision of feedback. Teacher A described her approach to instruction (essentially the basic problem-solving model involving identification and analysis of observed problems) in which she identifies goals, analyzes skills, plans, delivers, and evaluates instruction; she finds that she needs to add an additional step for her students who are ELLs: is this learning problem based on language or cultural barriers? Teacher B answered this question in a general sense, without specific focus on ELLs. She believed AM has been helpful as an assessment of specific mathematics skills more than an assessment of progress over time. Teacher C said AM is "perfect" if one uses the diagnostic features as she does to form groups of students.

Question 8: In your experience, how has the use of Accelerated Math impacted the amount and frequency of corrective feedback (or reinforcement) you provide your students who are English Language Learners? In general, each of the three teachers interviewed responded to this question by describing how they feel AM increases their ability to provide feedback that is not only more frequent, but also of better quality. Teacher A explained her perspective on problem solving in the classroom, and how her students' performance on AM prompts her to think more critically about the ways in which her instruction and the language skills of her students may interact. She noted that this simply adds another step to her problem solving process.

Chapter 5: Discussion

English language learners are an important sub-population of students in schools in the United States, and the extent to which schools provide effective instruction for this group has been a topic of great concern. In addition to learning social and academic English, ELL students are faced with a variety of challenges beyond those faced by most students in the general population. Much of the research literature devoted to this group of students covers issues related to literacy instruction, or to methods of assessment for summative purposes. Authors who contribute to this body of literature often advocate for the use of formative assessment for ELLs (Abedi, 2009; Deno, 2003), yet to date, empirical examinations of the use of such assessment practices for ELLs appear to be rare.

Examination of the use of technology-enhanced formative assessment in mathematics for this group of students should illustrate the ways in which the outcomes of instruction based on formative assessment practices could vary as a function of language proficiency. This study was intended to provide information that might illustrate ways in which formative assessment in mathematics might function similarly for ELLs and their NES peers, to identify issues to be considered by educators, and to provide some guidance for future research. Five research questions were posed:

1. To what extent is technology enhanced formative assessment implemented differently for ELLs than for NESs?
2. Do ELLs and NESs begin a school year with equivalent levels of mathematics skills?

3. Do ELLs and NESs exhibit equivalent rates of growth in mathematics skill over time?
4. Do ELLs and NESs respond differently to implementation of formative assessment in mathematics?
5. What were teachers' of ELLs perceptions of barriers to implementation, as well as relative strengths and weaknesses of the technology-enhanced formative assessment system?

Review of Results

Question 1: Implementation. The way in which features of Accelerated Math were implemented for ELLs and NESs was evaluated through descriptive and inferential statistical analyses. Descriptive comparisons of the use of each feature of AM between both groups of students suggested that, on average, ELLs in the dataset appeared to receive a slightly greater level of implementation of AM than NESs. As shown in Table 7, the average number of practice, test, diagnostic, and exercise items was greater for ELLs than for NESs. Follow-up analyses using single-level generalized linear models provided results that appeared to confirm this observation statistically, in which ELLs appeared to have attempted items or mastered objectives at a rate that was significantly greater than that of their NES peers after controlling for time spent using AM and grade level. This trend was not observed when classroom-level effects were taken into account.

This result suggested that ELLs in this dataset tended to be found in the classrooms of teachers who implemented AM to a somewhat greater extent than their NES peers in the dataset. Analysis of the ways in which implementation occurred between both groups of students was included in this study to add perspective to the estimates of differences between initial status and growth in math skills over time. That there were no meaningful systematic differences observed in the ways in which AM was implemented for both groups of students within classrooms suggests that differences in growth in STAR Math scores over time between the two groups of students are not likely to be the result of patterns of implementation of this formative assessment system. This finding was similar to the results of the study by Ysseldyke and Bolt (2007), in which implementation, as indexed by mastery of objectives, appeared to be equivalent across demographic groups.

Question 2: Differences in initial skill levels. After controlling for grade level and level of implementation the average STAR Math scaled score for ELLs was about 14.67 points below their NES peers, with a 95% confidence interval spanning 20.23 to 9.10 points below the NES mean. The mean difference of -14.67 points was just slightly outside the 95% confidence interval of the intercept for NESs, indicating a significant, but small difference in initial status between the two groups.

One finding, which was not anticipated, was that students who attended classrooms with a higher level of implementation by the end of the school year also appeared to start that year with higher math skills. For each additional level of implementation, the intercepts for STAR Math scaled scores increased by 24.39 and 26.35 points; this boost was essentially the same for ELLs and NESs alike. Although these analyses cannot describe what caused this trend, two possible explanations are offered: observed skill differences between levels of implementation at the beginning of the year were actually produced by limitations in the design of this retrospective study, or were associated with instructional trends within the schools that produced these data.

The implementation data used to create the dummy-coded variable “Objectives” represented a level of implementation achieved by the end of a school year. These data did not add information regarding the rate or consistency of implementation of AM. Furthermore, the fact that STAR Math administrations happened on many days across the school year (over 30 of the weeks included), and that most students had about three measurements, lessens interpretability of estimates of initial skill levels. The initial skill level estimates obtained in the analyses reported in this study are based on the assumption that each student experienced continuous growth in math skills over time.

It is possible, perhaps even likely, that groups of students in this sample received AM as a form of periodic instruction or intervention, and that the true amount of growth associated with this use of this system was actually localized within the span of one or two months rather than an entire school year. If implementation of technology enhanced formative assessment actually promotes growth in skills, then differences in the schedule of implementation can impact parameter estimates of continuous growth models.

For example, consider the differences between the linear intercepts and slopes of students who experience the same amounts of growth before, during, and after use of AM when the window of time in which AM was implemented is at the beginning or at the end of the year. The intercept (at Week 0) for students who received AM earlier in the year will be greater than those who received AM later in the year; the reverse would be true for slope parameters. The extent to which such differences in timing of AM use affected parameter estimates in the final growth model of this study is unknown. The potential impact of this issue is limited, in this dataset, by the fact that the majority of the students in this sample started using AM near the beginning of the year, and tended not to stop using AM until about the end of the school year.

Another explanation for this observation may exist in prior-year instructional practices between schools. Students who attended classrooms that had higher levels of implementation by the end of the year might have received similar instruction in the preceding school year. Although the specific growth estimates obtained in this study may not be exact, the significance and the scope of the effect of implementation of AM appear to be reliable. That students in higher implementation classrooms also started the school year with high math skills could be a result of school-level instructional trends, or a result of the possibility that these classrooms were in schools with more experience in implementation of technology enhanced formative assessment. Whether differences in initial status were due to limitations in study design and analysis or instructional trends within schools is an issue that would require additional research to resolve.

Question 3: Do ELLs and NESs exhibit similar growth in mathematics skill over time? Without taking the effects of implementation of AM into account, ELLs exhibited a rate of growth that was not significantly different from that of their NES peers. A follow-up growth model, model H, that excluded classroom-level parameters, suggested that ELLs grew by 0.15 to 0.44 fewer STAR Math points per week than their NES peers. This difference disappeared once between-classroom variation is accounted for in the final model. Like the results of models used to evaluate potential differences in implementation of AM between groups, this result suggests that overall growth in STAR Math scaled scores is related more to classroom membership (and potentially to specific instructional methods) than student characteristics.

Question 4: Do ELLs and NESs respond differently to formative assessment in mathematics? Implementation of AM appeared to accelerate slopes of STAR Math scaled scores over time, but this effect was moderated by ELL status. A significant three-way interaction between ELL status, time, and level of implementation of AM suggested that on average, ELLs STAR Math performance increased by about 0.06 fewer points over time, per level of implementation. Inclusion of this three-way interaction in the full model produced a better fit to the original data than the reduced models, but the size of this difference in response to implementation of AM is quite small. It is possible that differences in response to use of AM are associated with language factors that impact communication between teachers and learners. It does not appear to be, in a systematic way, the result of instructional mismatch or other trends in implementation.

Question 5: Teacher perceptions of technology-enhance formative assessment for ELLs. Past studies on formative assessment practices (including studies specific to AM), have indicated that effects of instruction fed by formative assessment can vary by school, teacher, and student factors (Wiliam, Lee, Harrison, & Black, 2004; Bolt, Ysseldyke, & Patterson, 2010). For this reason, and in anticipation of significant differences in implementation of formative assessment between ELLs and NESs, several general education teachers were recruited for interviews for the qualitative portion of this study. The intention was to gather input from teachers who work in regions with schools that are either traditionally attended by higher numbers of ELLs, or who work in districts with a relatively high subpopulation of ELLs. Their input was sought to add perspective to the findings from the quantitative portion of this study. For instance, responses to questions about patterns of implementation or observations of growth in STAR Math scores could provide illustration of information obtained in statistical models, or could suggest issues for future consideration in research on formative assessment for ELLs.

When asked to describe their use of AM, and whether their use of AM is different for ELLs than for NESs, the teachers interviewed generally described uses of AM that were close to those for which it was intended. Barriers to implementation that were reported by teachers included issues of infrastructure, such as school-day scheduling, or adoption of a new mathematics curriculum. Teacher B, a teacher of 8th graders in southern Texas, also observed that the motivation of her students impacted the extent to which formative assessment was implemented in her classroom.

All three teachers reported that their use of AM did not vary substantially by language status, or that they did not feel that they used features of AM differently, or to a different extent. However, each teacher remarked about the way in which language proficiency or cultural background knowledge appeared to impact their ELLs understanding of items in AM. Teacher C also reported that her third grade ELLs, in a southern Minnesota school district, sometimes start the year with math skills below their own grade level, and that she appreciates the flexibility of AM, as it allows her to select objectives for these students that are at a more appropriate skill level, and still related to the focus of her instruction.

Teachers A and B noted that, although their use of AM does not differ between groups, they find that allowing their ELL students opportunities to work together is helpful in addressing barriers to comprehension (due to language or background knowledge in mathematics). These descriptions of cooperative learning paralleled one of the recommendations given in the qualitative study and literature review by Gersten and

Baker (2000). Teacher B also reported that allowing group work appears to promote the engagement of all of her 8th graders.

Conclusions

Descriptive and inferential analyses of the ways in which AM was implemented for a large sample of students from classrooms and schools across the U.S.A. suggested that ELLs, as a group, may receive slightly greater implementation of formative assessment than their NES peers in the general population. Yet the average likelihood that an ELL student would receive instruction based on formative assessment practices was no greater than that of NES students. These two observations suggested that the ELLs who were included in this sample were found in classrooms of teachers who made greater use of a formative assessment system.

The use of technology-enhanced formative assessment appears to be effective in promoting development of mathematics skills for students in the general population. Although the strength of this effect appears to be smaller for ELLs, the size of the difference does not reach practically meaningful levels. Still, what we know about the heterogeneity of academic English skills for ELLs would suggest that this effect could vary substantially for some students, which could have implications for instructional planning and decision making.

Limitations

The results reported and discussed in this paper are informative and unique, but should be interpreted with several points in mind. Limitations on the extent to which these results may be generalized to other samples of students in other settings were imposed by study design decisions, as well as the availability of data.

Limitations imposed by study design. This study required a fairly large sample, including individual measurements, students, and classrooms, to be able to model trends in implementation of a technology-enhanced formative assessment system and growth in mathematics across one school year for ELLs and NESs alike. The use of an extant database allowed for such analysis, but at the expense of generalizability. Participants were not selected at random, and unobserved participant characteristics could have influenced results.

Because of the retrospective nature of this study there was substantial imbalance in the measurement occasions over time. Although the computational methods used in these analyses were not affected. In other words, it was possible to obtain parameter estimates for each model fit, and to compare relative fits between each model. Yet the issue of missing data within each time point does have implications for the interpretation of the results. Growth was measured in weeks, the smallest unit of time for which STAR Math was designed to be sensitive. Many schools administered STAR Math three times per year; others administered the test more than three times per year. The distribution of STAR Math assessments across the school year appeared to be about even (see Figures 2 and 3 in the appendix). Still, because each student was measured two or three times, on average, out of a possible 41 weeks included in the dataset, individual students contributed very little information to the longitudinal models fit.

This problem, understood to be a problem of missing data, was discussed by Little and Rubin (1987), who introduced terminology to describe patterns in missing data, and made recommendations on ways in which the problem of missing data could be addressed. An attempt to support the assumption that the data in this dataset are "missing at random" (MAR) was made through the inclusion of classroom as the mechanism of missingness in the final model. Because the occurrence of missing data (or STAR Math administration) depended mostly upon classroom membership, and was not associated with other model parameters, the results of the final model are considered to be interpretable with caution.

Classification of students. In this study it was not possible to verify the correct classification of students as ELL or NES. One of the basic premises for the conduct of this study was that learning math in a second language adds difficulty to the process of learning for students above and beyond those challenges normally encountered by students in the general population. The results of this study are based on the assumption that ELLs in this dataset received instruction in mathematics that was delivered primarily in English, and by teachers who are not typically fluent in the average ELLs first language.

Potential model specification errors. A plot of the residuals for the final longitudinal model against expected values based on the normal curve suggested that the model committed more over-predictions and under-predictions than would be expected (Figure 6 in the appendix). The same pattern did not appear in the fully unconditional three-level model. It appeared once Week was added, and did not change when the functional form of growth over time was changed from linear to quadratic. This may be an indication that important predictors were not included in the model, and that there is left-over systematic variation in the residuals, or that Week was problematic as an indicator of growth over time.

Recommendations for Future Research and Practice

Accounting for variation in language proficiency. There are different requirements in research and in practice for issues that involve ELLs. Research requires better indicators of English proficiency as covariates: there is less information provided by a dichotomous label, and more to be gained from research involving specific, well defined language skills. ELL, as a label, is only meaningful as an indicator of different educational needs. Including this label in models of growth in mathematics skill over time produced results that were statistically significant, but were of limited value for most applications. This label is more meaningful from the perspective of practice and policy, as it pertains to accountability and the prevention and reduction of systemic inequity. Yet educational research for this subpopulation would be more informative with a greater focus on interactions between specific skills or characteristics (such as time spent in the USA) and instructional variables than on indicators of group membership.

Much of what is written on the topic of instruction in mathematics for ELLs asserts that mathematics and language are inextricably linked. Studies on large-scale assessment provided evidence that linguistic factors can affect the reliability and validity of scores for ELLs. The results reported here suggest that the ways in which linguistic issues affect the mathematics learning of ELLs in response to the use of formative assessment data are subtle, yet potentially important. Future research should at least divide ELL groups into smaller groups based on level of English language proficiency, or should include scores from a reliable and valid test of English proficiency as a covariate. That English proficiency was included in this study as a binary variable with questionable validity and it interacted significantly with levels of AM implementation over time suggests that there is potential for significant and meaningful variation in growth trends related to language proficiency.

This is also important because language proficiency typically changes over time, and any effect of an interaction between language and instructional or assessment practices should change over time as well. This suggests that the small interaction observed in this study could, in some conditions, be greater and have more meaningful implications for evaluation of growth in math skills over time. Because the average change in language proficiency over time may span years, the changes in the relationship between instruction, assessment, and language may vary across age level instead of school year.

Overcoming language barriers. Information obtained from interviews with teachers who use AM with ELLs and NESs suggested that future research on instruction or classroom assessment for ELLs should include examination of peer-assisted learning, and whether grouping based on similarities in language and cultural background plus differences in academic English proficiency influences the effectiveness of instruction based on formative assessment data. Such research could be accomplished on a scale smaller than that of the current study, and could make use of a fully controlled, experimental design.

The teachers interviewed for this study each noted that AM items heavily dependent on text comprehension were problematic for their ELLs. One teacher reported avoiding such content, whereas the other two discussed ways in which they attempt to help their students acquire unfamiliar vocabulary or concepts. A considerable body of research has already been devoted to avoiding such effects in assessments used for summative purposes. Future research on formative assessment in mathematics for ELLs should extend findings from such research—such as the work of Abedi and colleagues (2003, 2005)—to research on formative assessment systems such as AM, or other popular alternatives. This research might seek to identify specific linguistic and cultural barriers (such as national currency, or symbols used to express quantities), and should evaluate the effects associated with removal of such barriers in test materials.

Dynamic growth models. This study provided evidence that students who are ELLs learning math in a second language benefit to a somewhat lesser extent, on average, from technology enhanced formative assessment than students who are native English speakers. However, an examination of the ways in which this type of formative assessment was implemented for all students suggested that the specific growth estimates produced by these analyses may not represent the actual rate of growth educators should expect when implementing this particular method of formative assessment.

Earlier studies on the effectiveness of technology enhanced formative assessment evaluated effects in terms of magnitude of implementation, rather than timing of implementation. Their designs were controlled, and it is possible that they may not have included students that were not using AM within a consistent timeline. This was not true in the current study: there was considerable variation in the dates in which AM implementation started and stopped across students (see Figure 7 in the appendix for an illustration of this variation). Useful information may be obtained from additional research designed to detect discontinuous growth in skills in response to instruction. Such information should increase the efficiency and the accuracy with which decisions can be made about a student’s response to instruction. Current educational policies about eligibility for special education are sometimes based upon measurement of students' response to a process of evidence-based intervention. The amount of time and data required to make eligibility judgments can be a source of frustration for students, parents, and educators.

What is known about the effects of a specific instructional practice over time can be limited by generalizations made in research. For example, consider Figure 8 in the appendix. The STAR Math scores and dates of AM usage are depicted in these two figures. Student A was a third grader attending a school that might have been using AM as an intervention for students' who scored below some criterion upon fall screening. The first record of his use of AM for the year occurred several days after his first STAR Math score, and the final record of his use of AM for the year occurred several days before his second STAR Math score. Substantial growth is evident between these two scores, but a level growth trend is evident afterword. Student B started to use AM at about the same time, but reportedly continued to use AM throughout the rest of the school year. Examination of the four administrations of STAR Math suggests fairly even growth across the school year. By estimating a single slope based on time for a sample of students who received AM for different amounts of time, information about the relationship between the use of technology enhanced formative assessment and achievement is diminished. Singer and Willett (2003) describe methods that may be used in such situations, when growth cannot be assumed to be continuous over time. In one such method several slopes for time may be included in one model; for studies evaluating the effects of a treatment over time, this could include growth estimates for skill growth before, during, and after implementation.

Problem solving or instructional decision making. Because the data produced by formative assessment are used to evaluate the results of instruction or intervention, it is important for educators to have direction not only on the way in which such assessment or instruction should be provided, but also on the timelines in which instructional decisions should be made. More information about reasonable differences in response to intervention for determining special educational needs is necessary. The results of this study suggested that there are some students who are ELLs that may benefit less from use of a technology-enhanced formative assessment system than students who are native English speakers, and that there are other students who are ELLs that benefit to about the same extent. It is the responsibility of teams of teachers and other educators to determine whether a student's response to instruction or intervention was within the range expected for typically developed students, or whether a lack of response to intervention indicates a need for additional or special educational service.

References

Abedi, J. (2003). Impact of Student Language Background on Content-Based Performance: Analyses of Extant Data (CSE Tech. Rep. No. 603). Los Angeles: University of California, National Center on Evaluation, Standards, and Student Testing.

Abedi, J. (2004). The No Child Left Behind Act and English language learners: assessment and accountability issues. *Educational Researcher, 33*(1), 4-14.

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

Abedi, J. (2009). *Research and recommendations for formative assessment with English language learners.* In Andrade, H., & Cizek, G. (Eds.) *Handbook of Formative Assessment.* New York: Routledge.

Abedi, J. & Herman, J. (2010). Assessing English language learners' opportunity to learn mathematics: issues and limitations. *Teachers College Record, 112*(3).

Abedi, J., Lord, C., & Plummer, J. (1997). *Language background as a variable in NAEP mathematics performance* (CSE Tech. Rep. No.429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abella, R. Urrutia, J. & Shneyderman, A. (2005). An examination of English-language achievement test scores in an English language learner population. *Bilingual Research Journal, 29*(1), 127-144.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267-281). Akademiai Kiado, Budapest, Hungary.

Albers, C., Hoffman, A., & Lundahl, A. (2009). Journal coverage of issues related to English language learners across student service professions. *School Psychology Review*, *38*(1), 121-134.

Albus, D., Thurlow, M., & Liu, K. (2002). *1999-2000 participation and performance of English language learners reported in public state documents and web sites* (LEP Projects Report 3). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Amerein, A. L., & Berliner, D. C. (2002a). *An Analysis of Some Unintended and Negative Consequences of High-Stakes Testing*. The Great Lakes Center for Education Research & Practice. MI: East Lansing.

Amerein, A. L., & Berliner, D. C. (2002b). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10* (18).Retrieved 11/3/2008 from http://epaa.asu.edu/epaa/v10n18/.

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.

Bailey, A., & Drummond, K. (2006). Who is at risk and why? Teachers' reasons for concern and their understanding and assessment of early literacy. *Educational Assessment, 11* (3), 149-178.

Baker, S., Plasencia-Peinado, J., Lezcano-Lytle, V. (1998). The use of curriculum-based measurement with language minority students. In M. Shinn (Ed.) *Advanced applications of curriculum-based measurement (pp. 175-213).* New York: Guilford Press.

Batalova, J., Fix M., and Murray, J. (2007). *Measures of Change: The Demography and Literacy of Adolescent English Learners—A Report to Carnegie Corporation of New York.* Washington, DC: Migration Policy Institute.

Beal, C., Adams, N., & Cohen, P. (2010). Reading proficiency and mathematics problem solving by high school English language learners. *Urban Education*, *45*, 58-74. DOI: 10.1177/0042085909352143

Betts, E. A. (1946). *Foundations of reading instruction.* New York: American Book.

Betts, J., Bolt, S., Decker, D., Muyskens, P., & Marston, D. (2009). Examining the role of time and language type in reading development for English language learners. *Journal of School Psychology, 47,* 143-166.

Bolker, B., Brooks, M., Clark, C., Geange, S., Poulsen, J., Stevens, H., & White, J. S. (2008). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution, 24* (3), 127 – 135.

Bolt, D., Ysseldyke, J., & Patterson, J. (2010). Students, teachers, and schools as sources of variability, integrity, and sustainability in implementing progress monitoring. *School Psychology Review, 39,* 612-630.

Black, P. &Wiliam, D. (1998a). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80 (2): 139-148. (Available online: http://www.pdkintl.org/kappan/kbla9810.htm.).

Black, P., &Wiliam, D. (1998b). Assessment and classroom learning. *Assessment in Education*, 5 (1): 7-74.

Bloom, B., Hastings, J. T., & Madaus, G. (1971). *Handbook on Formative and Summative Evaluation of Student Learning.* New York: McGraw-Hill.

Blot, K., Della-Piana, G., Turner, W. (1998). The development and employment of formative evaluation instruments to enhance students' opportunity to learn. *Proceedings of the Frontiers in Education Conference*, 1355-1360.

Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research & Evaluation*, 8(9). Retrieved July 22, 2008 from http://PAREonline.net/getvn.asp?v=8&n=9.

Bradby, D. (1992). *Language characteristics and academic achievement: A look at Asian and Hispanic eighth graders in NELS:88*. Washington, DC: U.S. Government Printing Office.

Buchanan, T. (2000). The efficacy of a world-wide web mediated formative assessment. *Journal of Computer Assisted Learning*, *16*, 193-200.

Burns, M. K., Codding, R. S., Boice, C. H., & Lukito G. (2010). Meta-analysis of acquisition and fluency math interventions with instructional and frustration level skills: Evidence for a skill-by-treatment interaction. *School Psychology Review, 39*, 69-83.

Burns, M., & Dean, V. (2005). Effect of drill ratios on recall and on-task behavior for children with learning and attention difficulties. *Journal of Instructional Psychology*, *32*(2), 118-126.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record*, *63,* 723-733.

Callahan, R. (2005). Tracking and high school English Learners: limiting opportunity to learn. *American Educational Research Journal*, *42*(2), 305-328.

Cizek, G. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, *20* (4), 19-27.

Cizek, G., Fitzgerald, S., & Rachor, R. (1995). Teachers' assessment practices: preparation, isolation, and the kitchen sink. *Educational Assessment*, *3* (2), 159-179.

Clarkson, P. C. (1992). Langauge and mathematics: a comparison of bilingual and monolingual students of mathematics. *Educational Studies in Mathematics*, *23,* 417-429.

Cleary, L. M. (2008). The imperative of literacy motivation when native children are being left behind. *Journal of American Indian Education*, *47* (1), 96-116.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58*, 438-481.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism*, *19,* 121-129.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California State Department of Education (Ed.) *Schooling and Language Minority Students: A Theoretical Framework.* Evaluation, Dissemination, and Assessment Center, California State University, Los Angeles.

Deno, S. (2003). Developments in curriculum-based measurement.*The Journal of Special Education, 37,* 184-192.

Diamond, J. & Spillane, J. P. (2004). High-stakes accountability in urban elementary schools: challenging or reproducing inequality? *Teachers College Record*, *106,* 1145-1176.

Dunn, K., & Mulvenon, S. (2009). A critical review of research on formative assessment: the limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research, & Evaluation, 14*. Retrieved from http://pareonline.net/getvn.asp?v=14&n=7.

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L.,Means, B., Murphy, R., Penuel, W., Javitz, H., Emery, D., & Sussex, W. (2007). *Effectiveness of reading and mathematics software products: Findings from the first student cohort.* Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Edl, H., Jones, M., & Estell, D. (2008). Ethnicity and English proficiency: Teacher perceptions of academic and interpersonal competence in European American and Latino students. *School Psychology Review, 37*(1), 38-45.

Elston, D. A., Moss, R., Boulinier, T., Arrowsmith, C., & X. Lambin (2001). Analysis of aggregation, a worked example: numbers of ticks on red grouse chicks. *Parasitology, 122*, 563-569. DOI: 10.1017/S0031182001007740.

Fry, R. (2008). *The Role of Schools in the English Language Learner Achievement Gap.* Washington, DC: Pew Hispanic Center, June 2008.

Fuchs, L. (1998). Computer applications to address implementation difficulties associated with curriculum-based measurement. In Shinn, M. (Ed.) *Advanced Applications of Curriculum-Based Measurement.* New York: The Guilford Press.

Fuchs, L. (2004). The past, present, and future of curriculum based measurement research. *School Psychology Review, 33,* 188-192.

Fuchs, L., & Deno, S. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57,* 488-500.

Fuchs, L., & Fuchs, D. (1986). Effects of systematic formative evaluation: a meta-analysis. *Exceptional Children, 53*, 199-208.

Fuchs, L., Fuchs, D., Hamlett, C., & Stecker, P. (1991). Effects of curriculum based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal, 28,* 617-641.

Fuchs, L., Fuchs, D., Hamlett, C., Walz, L. & Germann, G. (1993). Formative evaluation of academic progress: how much growth can we expect? *School Psychology Review, 22*, 27-48.

Fuchs, L., Fuchs, D., Hamlett, C., & Whinnery, K. (1991). Effect of goal line feedback on level, slope, and stability of performance within curriculum-based measurement. *Learning Disabilities Research and Practice, 6*, 66-74.

Gándara, P., Rumberger, R., Maxwell-Jolly, J. and Callahan, R., (2003). English learners in California schools: Unequal resources, unequal outcomes. *Education Policy Analysis Archives, 11*, 1-52.

Garibaldi, A. M. (1992). Educating and motivating African American males to succeed. *Journal of Negro Education, 61*, 4–11.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* New York: Cambridge University Press.

Gersten, R., & Baker, S. (2000). What we know about effective instructional practices for English language learners. *Exceptional Children, 66,* 454-470.

Gersten, R. Keating, T. & Irvin, L. (1995). The burden of proof: Validity of improvement of instructional practice. *Exceptional Chilren, 61,* 510-519.

Goddard, R., Hoy, W. K., & Woolfolk-Hoy, A. (2000). Collective teacher efficacy: its meaning, measure, and impact on student achievement. *American Educational Research Journal, 37,* 479-507.

Good, T., & Nichols, S. (2001). Expectancy effects in the classroom: a special focus on improving the reading performance of minority students in first grade classrooms. *Educational Psychologist, 36*, 113-126.

Hakuta, K., Buttler, Y. & Witt, D. (2000). *How long does it take English learners to attain proficiency?* The University of California Linguistic Minority Research Institute.

Jordan, N., & Levine, S. (2009). Socioeconomic variation, number competence, and mathematics learning difficulties in young children. *Developmental Disabilities Research Reviews, 15*, 60-68.

Kluger, A., & DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254-284.

Krashen, S., & Brown, C. L. (2005). The ameliorating effects of high socioeconomic status: a secondary analysis. *Bilingual Research Journal, 29*, 185-196.

Krueger, A., & Whitmore, D. (2002). Would smaller classes help close the Black-White achievement gap? In J. Chubb & T. Loveless (Eds.), *Bridging the Achievement Gap*. Washington, DC: Brookings Institution.

Lee, S. K. (2002). The significance of language and cultural education on secondary achievement: a survey of Chinese-American and Korean American students. *Bilingual Research Journal, 26*, 213–224.

López, E., Ehly, S., & Garcia-Vázquez, E. (2002). Acculturation, social support, and academic achievement of Mexican and Mexican-American high school students: an exploratory study. *Psychology in the Schools, 39*, 245-257.

Little, R., & Rubin, D. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

Marsh, C. (2007). A critical analysis of the use of formative assessment in schools. *Educational Research Policy and Practice, 6*, 25-29.

McNair, S., Bhargava, A., Adams, L., Edgerton, S., & Kypros, B. (2003). Teachers speak out on formative assessment practices. *Early Childhood Education Journal, 31*, 23-31.

Molnar, A., Smith, P., Zahorik, J., Palmer, A., Halbach, A., & Erle, K. (1999). Evaluating the SAGE program: a pilot program in targeted teacher-pupil reduction in Wisconsin. *Educational Evaluation and Policy Analysis, 21*, 165-177.

National Mathematics Advisory Panel. (2008). Foundations for success: The final report of the National Mathematics Advisory Panel. Washington, DC: U.S. Department of Education.

No Child Left Behind Act of 2001. (2001). Pub. Law No. 107.110.

Olson, J., & Goldstein, A. (1997). *The inclusion of students with disabilities and limited English proficiency in large-scale assessments: a summary of recent progress.* U.S. Department of Education, Office of Educational Research and Improvement, NCES 97-482.

Peugh, J. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*, 85-112.

Pressey, S. (1927). A machine for automatic teaching of drill material. *School and Society, 25*, 549-552.

Renaissance Learning, Inc. (1998a).*Accelerated Math.* Wisconsin Rapids, WI: Advantage Learning Systems (http://www.renaissancelearning.com).

Renaissance Learning, Inc. (1998b). *STAR Math*. Wisconsin Rapids, WI: Renaissance Learning.

Reyes, L. H., & Stanic, G. (1988). Race, sex, socioeconomic status and mathematics. *Journal for Research in Mathematics Education*, *19*, 26-43.

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education*, *17*, 1-24.

Rumberger, R. & Gándara, P. (2004). Seeking equity in the education of California’s English Learners. *Teachers College Record*, *106*, 2032-2056.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119-144.

Salvia, J., Ysseldyke, J., & Bolt, S. (2010). *Assessment: In special and inclusive education* (11th edition). Boston: Houghton-Mifflin.

Shafer Willner, L., Rivera, C., & Acosta, B. (2008). *Descriptive Study of State Assessment Policies for Accommodating English Language Learners.* Arlington, VA: The George Washington University, Center for Equity and Excellence in Education.

Shernoff, D. J., Csikszentmihalyi, M., Schneider, B., Shernoff, E. S. (2003). Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly, 18*, 158-176.

Shinn, M., & Bamonto, S. (1998). Advanced applications of curriculum-based measurement: “big ideas” and avoiding confusion. In Shinn, M. (Ed.) *Advanced Applications of Curriculum-Based Measurement.* New York: The Guilford Press.

Shyyan, V., Thurlow, M., & Liu, K. (2008). Instructional strategies for improving achievement in reading, mathematics, and science for English language learners with disabilities. *Assessment for Effective Intervention, 33*(3), 145-155.

Singer, J., & Willett, J. (2003). *Applied Longitudinal Data Analysis.* New York: Oxford University Press.

Spicuzza, R., & Ysseldyke, J. (1999). *Using Accelerated Math to enhance instruction in a mandated summer program*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Stevens, R., Butler, F., & Castellon-Wellington, M. (2000). *Academic Language and Content Assessment: Measuring the Progress of English Language Learners (ELLs).* Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Stiggins, R. (2005). From formative assessment to assessment for learning: a path to success in standards-based schools. *Phi Delta Kappan*, 87, 324-328.

Stiggins, R., & Chappuis, J. (2005). Using student-involved classroom assessment to close achievement gaps. *Theory into Practice, 44,* 1-18. doi: 10.1207/s15430421tip4401_3

Tate, W. (1997). Race-ethnicity, SES, gender, and language proficiency trends in mathematics achievement: an update. *Journal for Research in Mathematics Education, 28*, 652-679.

Teelucksingh, E., Ysseldyke, J., Spicuzza, R., & Ginsburg-Block, M. (2001). *Enhancing the learning of English language learners: consultation and a curriculum based monitoring system.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurber, R., Shinn, M., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity curriculum-based mathematics measures. *School Psychology Review, 31*, 498-513.

Tierney, R. (2006). Changing practices: influences on classroom assessment. *Assessment in Education, 13,* 239-264.

White, K. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*, 461-481.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles, Policy & Practice, 11*, 49-65.

Ysseldyke, J., Betts, J., Thill, T., & Hannigan, E. (2004). Use of an instructional management system to improve mathematics skills for students in Title 1 programs. *Preventing School Failure, 48*, 10-14.

Ysseldyke, J., & Bolt, D. (2007). Effect of technology-enhanced continuous progress monitoring on math achievement. *School Psychology Review, 36*, 453-467.

Ysseldyke, J., Spicuzza, R., Kosciolek, S., Teelucksingh, E., Boys, C., & Lemkuil, A. (2003). Using a curriculum based instructional management system to enhance math achievement in urban schools. *Journal for the Education of Students Placed at Risk*, *8*, 247-265.

Ysseldyke, J., & Tardrew, S. (2007). Use of a progress monitoring system to help teachers differentiate mathematics instruction. *Journal of Applied School Psychology, 24*, 1-28.

Ysseldyke, J., Tardrew, S., Betts, J., Thill, T., & Hannigan, E. (2004). Use of an instructional management system to enhance the mathematics instruction of gifted and talented students. *Journal for the Education of the Gifted, 27*, 293-310.

Zehler, A., Fleischman, H., Hopstock, P., Stephenson, T., Pendzick, M., & Sapru, S. (2003). *Descriptive study of services to LEP students and LEP studentswith disabilities* (Policy Report Contract No.ED-00-CO-0089). Arlington, VA: Development Associates.

Appendix: Supplemental Figures

Figure 1: Use of Features of Accelerated Math between Groups

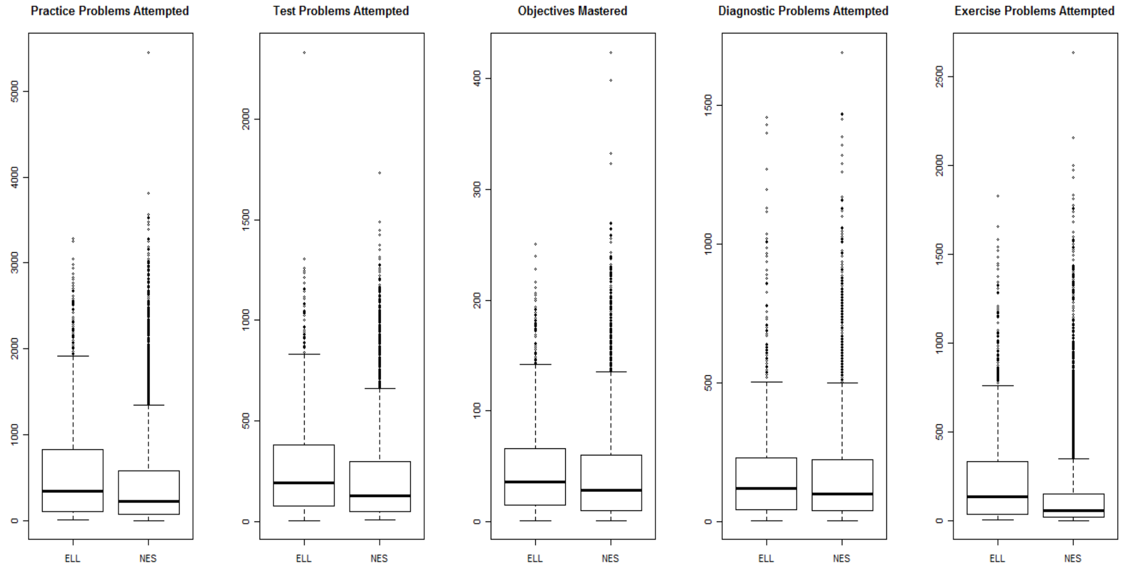


Figure 2: Number of STAR Math Administrations per Week across Grades

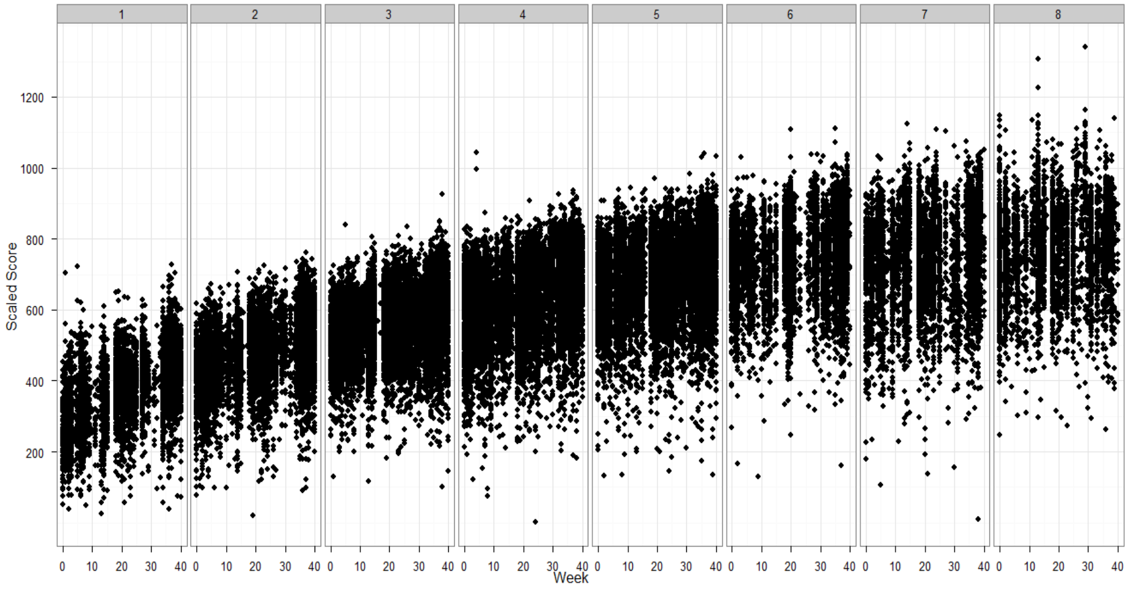


Figure 3: Number of STAR Math Administrations between Groups over Time.

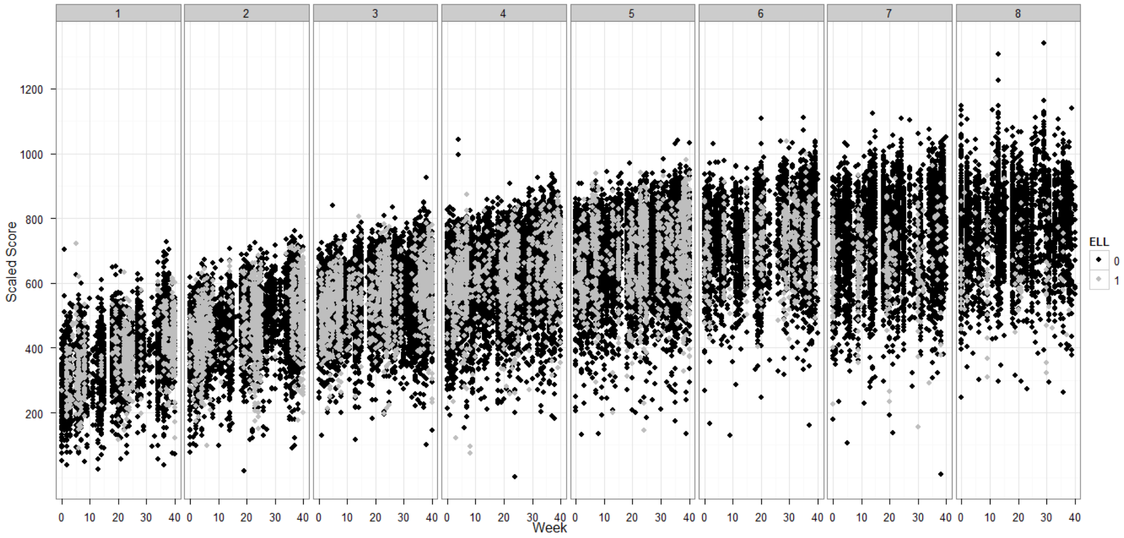


Figure 4: Normality of Student Intercepts

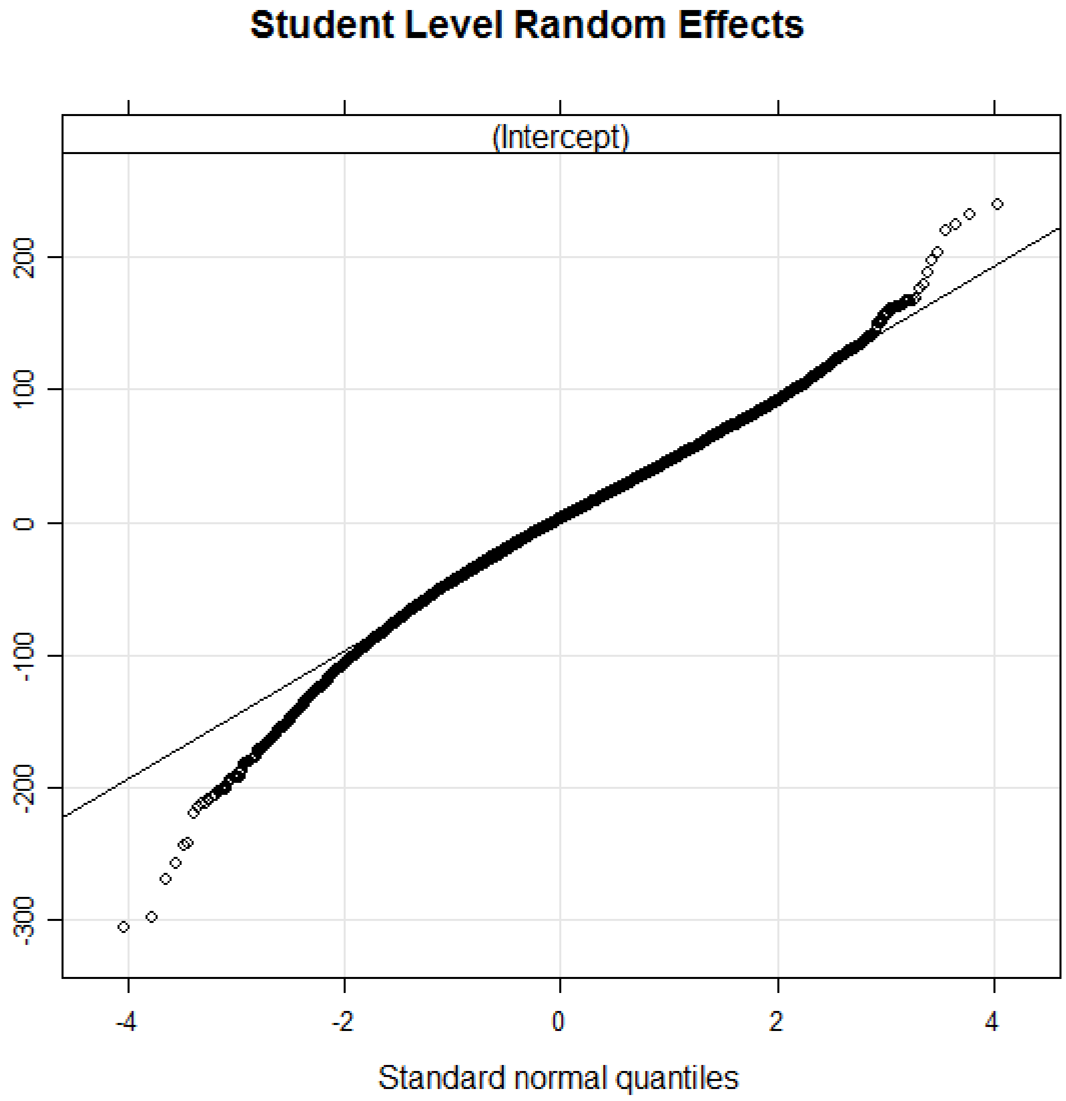


Figure 5: Normality of Classroom Intercepts and Random Slopes

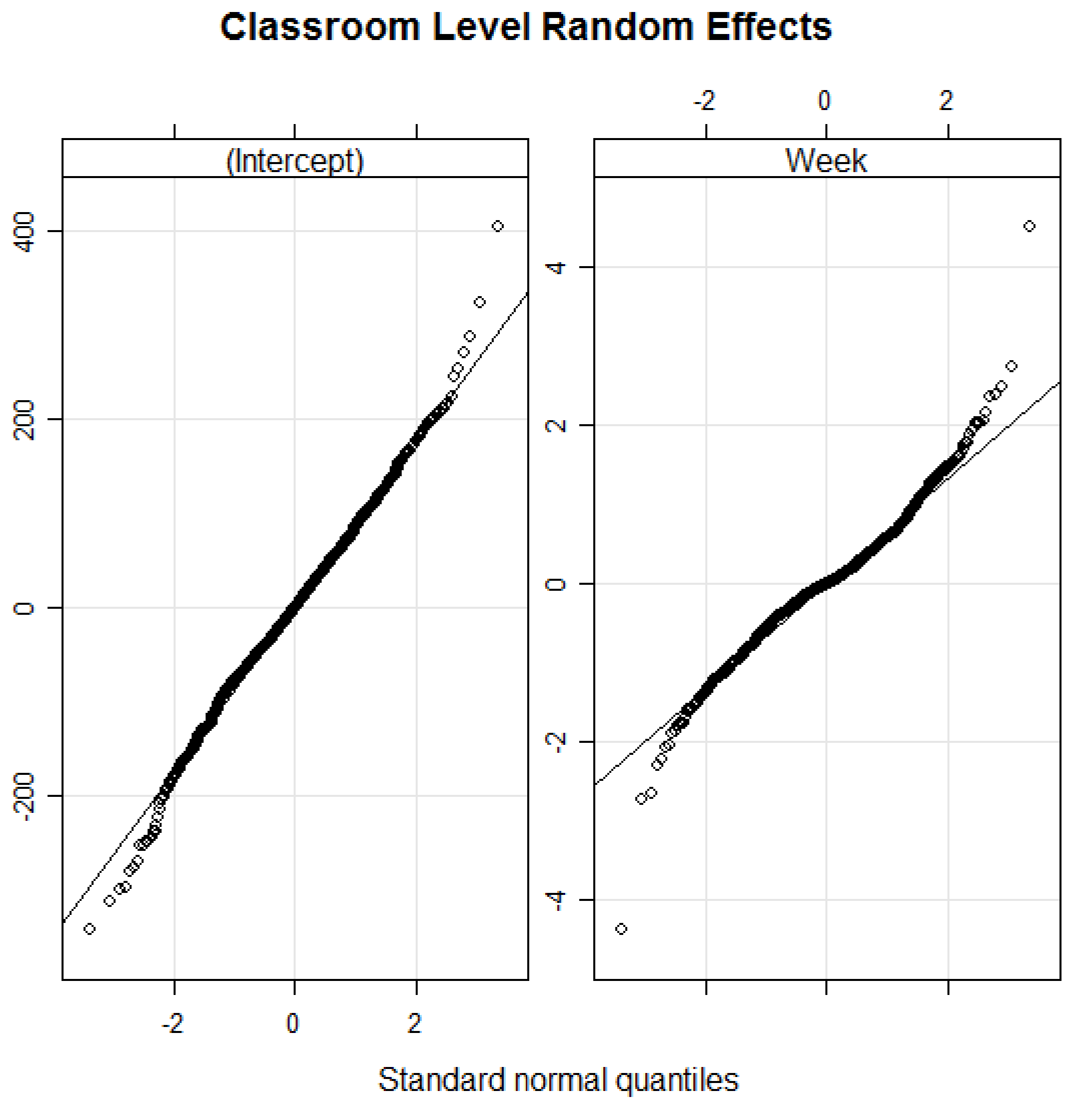


Figure 6: Normality of Final Model Residuals

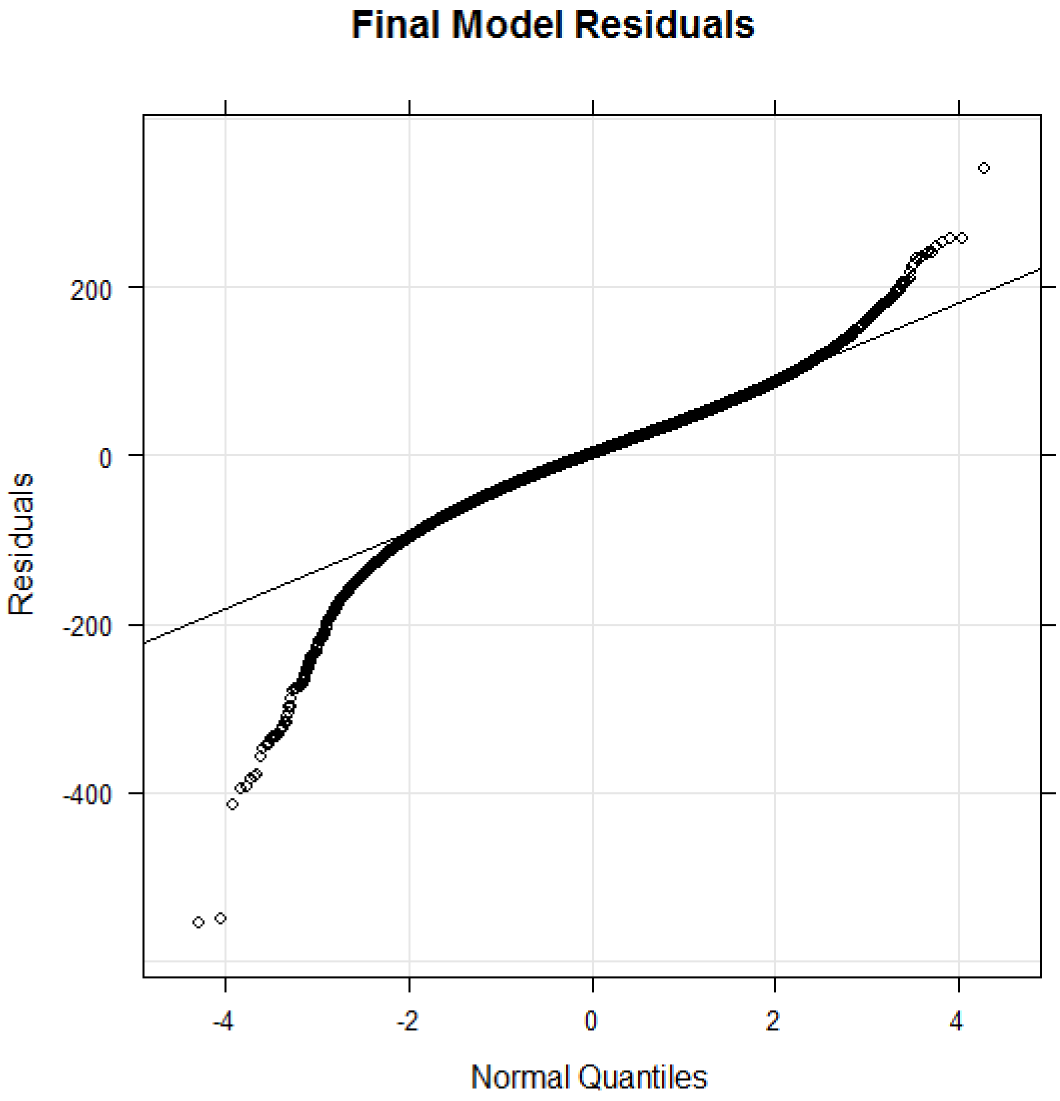


Figure 7: Distribution of AM Start and Stop Dates

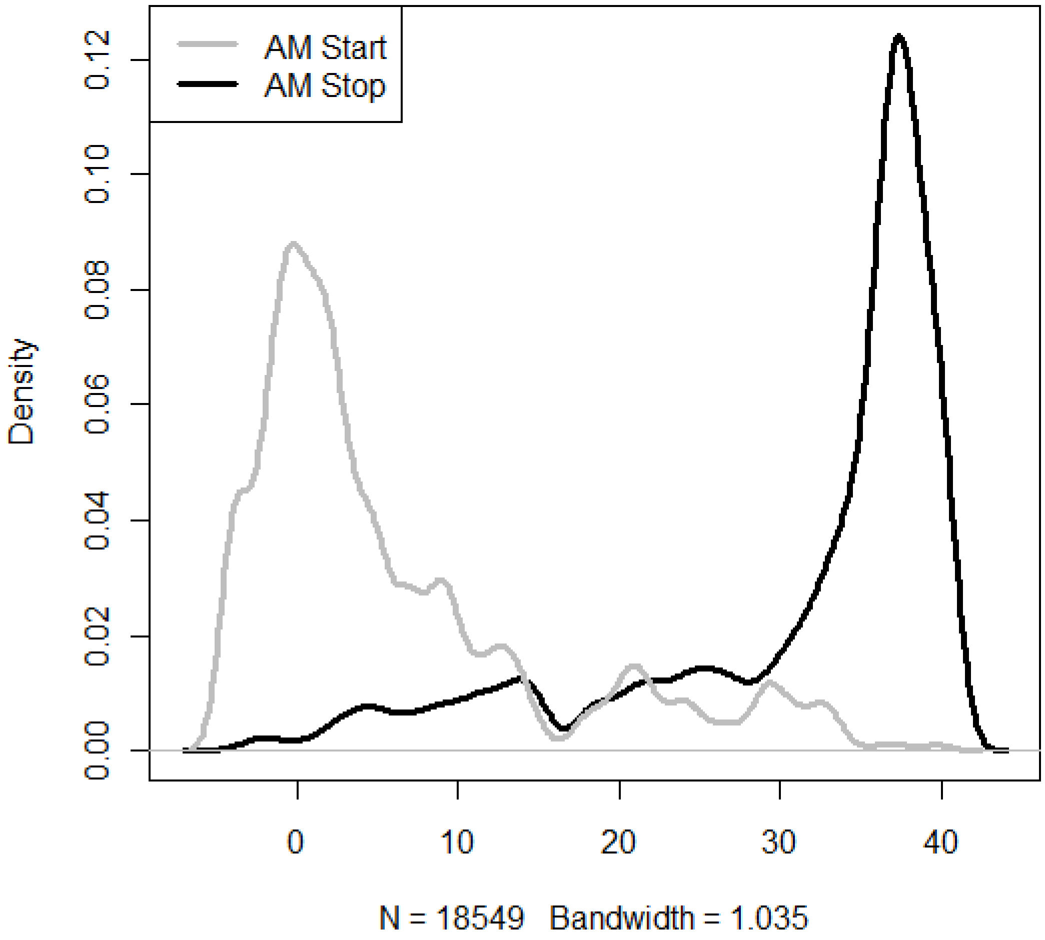
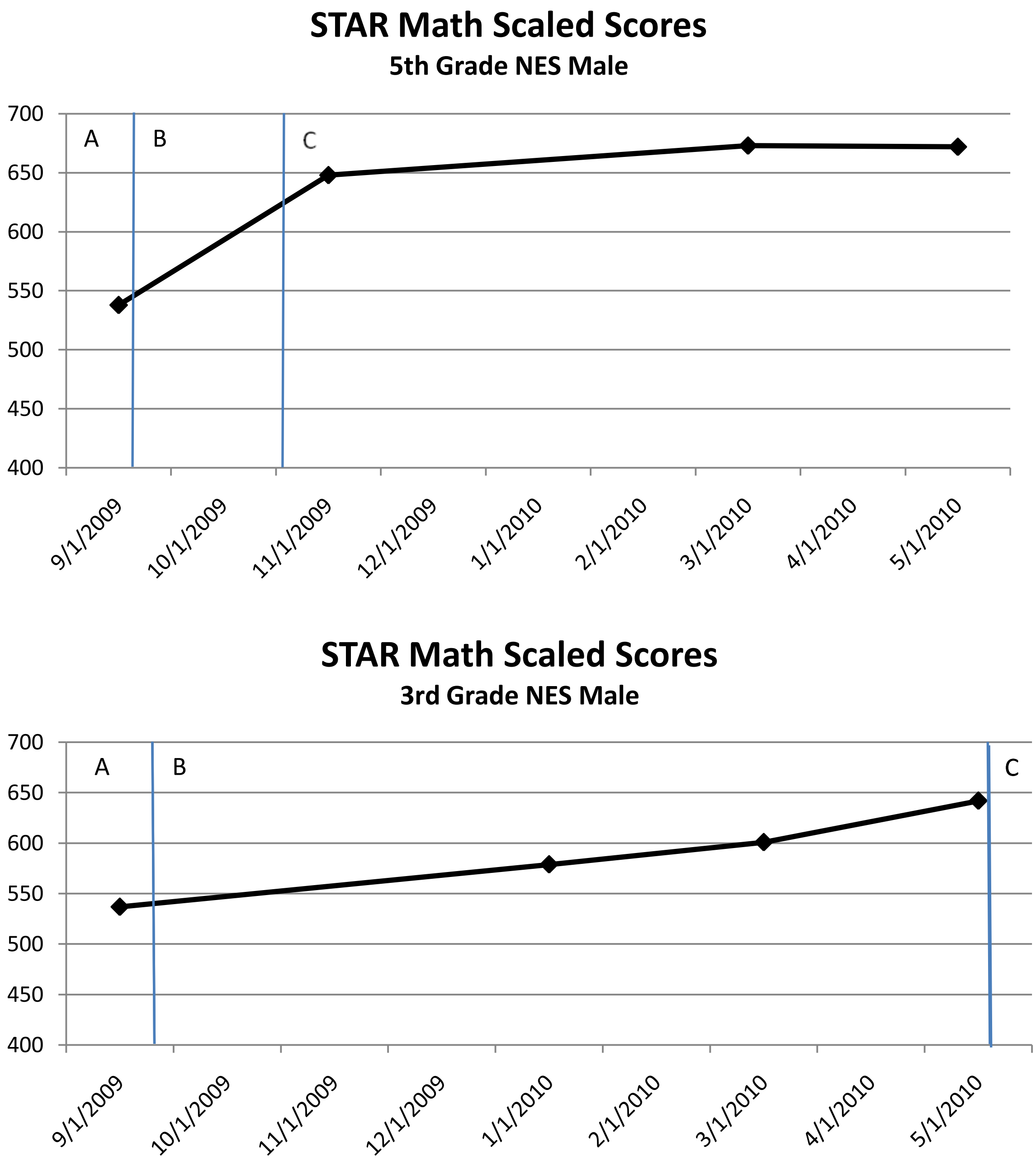


Figure 8: Example Differences in Schedule of Implementation of AM



STAR Math Scaled Scores

5th Grade NES Male

STAR Math Scaled Scores

3rd Grade NES Male