EFFECTS OF FORMATIVE ASSESSMENT ON MIDDLE SCHOOL STUDENT ACHIEVEMENT IN MATHEMATICS AND READING

Abram J. Kline

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the School of Education (Educational Psychology, Measurement & Evaluation).

Chapel Hill
2013

Approved by:

Gregory J. Cizek

Jeff A. Greene

Tammy L. Howard

ABSTRACT

ABRAM J. KLINE: Effects of Formative Assessment on Middle School Student Achievement in Mathematics and Reading
(Under the direction of Dr. Gregory Cizek)

Working with a dataset from middle school students' mathematics and reading assessments, this study was conducted to gather evidence regarding effects of formative assessment on student achievement. The study used student usage statistics from an online formative assessment program to examine the effect of formative assessment on student growth scores from end-of-grade summative assessments. The major findings of this study suggest that formative assessments are positively related to student achievement in reading and mathematics. Results suggest that short-cycle reading formative assessments result in positive gains for students in reading. Both student and school-level short-cycle reading formative assessment frequency were observed to have a positive effect on student achievement in reading.

The results from this study also suggest that long-cycle mathematics formative assessments may result in positive gains for students. The interaction between student and school-level long-cycle mathematics assessment frequency suggested that students who attend schools that administer a greater number of long-cycle mathematics formative assessments experience positive gains in mathematics achievement. In addition, short-cycle mathematics formative assessments seem to have a particularly stronger positive effect on the achievement of students who are economically disadvantaged.

ACKNOWLEDGEMENTS

I would like to thank ClassScapes and the North Carolina Department of Public Instruction for allowing me to use their data, Dr. Cizek for his guidance and dedication, Dr. Greene for his recommendations and multilevel modeling expertise, and Dr. Howard for her always enthusiastic support. In addition, I’d like to thank Amy Powell-Moman for providing additional support with my SAS programming and multilevel modeling. Last but furthest from least, I’d like to thank Jocelyn for enduring the late nights, for being a constant source of encouragement, and also for delivering delicious meals to the office throughout this entire process.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

LIST OF EQUATIONS

Equation

CHAPTER ONE: INTRODUCTION

Introduction

The passage of *No Child Left Behind* (NCLB, 2001) legislation in the United States made summative assessments mandatory for public school students in an attempt to shrink national achievement gaps and increase student academic success. However, a report issued by the OECD in 2009 indicated that the United States has fallen to 21st of the top 26 OECD countries in terms of graduation rates (OECD, 2009). In addition, American students have shown little growth over the last decade in primary subjects such as mathematics and reading (U.S. Department of Education, 2011). These outcomes have left educators, administrators, and policymakers searching for more effective methods of improving student achievement. Race to the Top (RTTT), authorized under the American Recovery and Reinvestment Act (ARRA) of 2009, came with the same goals of closing national achievement gaps and increasing graduation rates (U.S. Department of Education, 2009). RTTT encouraged states to innovate their measures of student learning and achievement (Bill & Melinda Gates Foundation, 2010). Since the arrival of RTTT, the topic of formative assessment has garnered a great deal of interest among the national and international education community. Although past research has shown formative assessment to have a positive effect on student achievement (Burns et al, 2010; Bergan et al., 1991; Black & Wiliam, 1998; Fuchs & Fuchs, 1986; Martinez & Martinez, 1992; Sadler, 1989; White & Frederiksen, 1998), it can be a time-intensive endeavor. However, the recent interest in formative assessment has spurred the development of many online formative assessment programs (OFAP) designed to take some of the burden off the shoulders of instructors, allowing students to benefit from formative assessment without adding another time-intensive task to the instructor's already busy schedule.

Whereas it is exciting to see these types of educational innovations, it is also important to monitor how these tools function in terms of student outcomes. Research at the Gates' Foundation has suggested that an ideal scenario is one in which, "formative assessments are embedded in the curriculum and actually guide the design of the summative assessments; the two forms of assessment should be intertwined" (2010, p. 6). A logical next step then would be to ask the question: Does the practice of formative assessment in the classroom affect student performance on summative assessments?

This thesis attempts to answer that question. Working with a dataset of student information from an OFAP and student achievement data from state-mandated mathematics and reading summative assessments for 6th, 7th, and 8th grade students, this study investigated whether there are differences in student growth scores that may be attributable to formative assessment based on student use statistics from an OFAP.

Definitions

The following set of definitions will be used throughout this paper.

Academic Change Score (AC-Score): student achievement measure based on state summative assessment performance. This was the primary dependent variable in this study. A detailed calculation of this score is provided in a subsequent section.

Online Formative Assessment Program (OFAP): a formative assessment tool which is available online, includes test items which have been aligned to state standards, and provides detailed, student-level feedback to the instructor.

OFAP Assessment Count: the total number of OFAP assessments taken in the given school year. This study included student-level counts and school-level means for both Mathematics and Reading OFAP assessments.

OFAP Assessment Type: the category(ies) of OFAP assessments taken based on differing assessment characteristics. This study differentiates between short-cycle and long-cycle assessments as the two possible OFAP assessment types.

Background

Formative assessment can be defined as "frequent, interactive assessments of student progress and understanding to identify learning needs and adjust teaching appropriately" (OECD, 2005, p. 21). Properly administered formative assessment can provide useful data for educators so that they can understand in which areas their students are obtaining solid understanding and in which areas their students may need remediation.

Whereas formative assessment can occur organically within a classroom setting in the form of discussion, observed group work, or simple student-teacher interaction, these forms of formative assessment are, for obvious reasons, more difficult to quantify. In addition, information gathered from these types of formative assessments would need to be recalled by the instructor at a later time, which can easily lead to some students' remediation needs being forgotten. Furthermore, students who are struggling but do not vocalize their needs may not be apparent to the teacher until summative tests are administered. In situations such as these, a formative assessment system, such as the OFAP used for this study, could prove to be a very useful tool for the instructor, assuring that data is collected and stored for all students in the class regardless of how much they speak up. Although the purpose of this study is not to contrast computerized and non-computerized formative assessments, the fact that an OFAP is used as the formative assessment tool facilitates the collection of data that can be used to investigate the potential effects of formative assessment. Not only does the OFAP provide useful data for the teacher, but it also provides quantitative data on formative assessment which is otherwise more difficult to obtain.

As previously mentioned, many companies are beginning to offer technology-based formative assessment programs. However, as with any new technology, there is a learning curve, and any potentially positive or negative effects may take time to become evident. With this in mind, it is important to understand how systems such as the OFAP are being used in the classroom and what types of student outcomes result from their use. Again, whereas this study does not focus directly on the technological aspect of formative assessment, it may provide some insight as to how formative assessment functions within that context. In addition to investigating the effects of formative assessment, it is also important to address any potential differences between different subgroups of students such as economically disadvantaged students (EDS), students with disabilities (SWD), and limited English proficiency (LEP) students. These issues and the related literature are discussed further to provide support and justification for the study.

Assessment Types

When discussing student assessment it is important to differentiate between three different forms of assessment: formative assessment, summative assessment, and interim assessment. Each form of assessment carries its own strengths and weaknesses and is designed to serve different purposes. In addition, although each form of assessment tests students' knowledge, results from each type of assessment carry specific implications and therefore can only be effective to the extent that they are used as intended.

Formative assessment. Formative assessment is the most frequently occurring of the three forms of assessment presented and is defined by a few major characteristics. According to Cizek (2010), formative assessment is administered midstream, in the course of some unit of instruction with, the primary purpose of one or more of the following: 1) to identify the student's strengths and weaknesses; 2) to assist educators in the planning of subsequent instruction; 3) to aid students in guiding their own learning, revising their work, and gaining self-evaluation skills; and 4) to foster increased autonomy and responsibility for learning on the part of the student.

This is consistent with the largely agreed upon definition of formative assessment within the educational community. The keys to the definition of formative assessment that set it apart from interim and summative assessment are its timing and purposes. As formative assessment is intended to evaluate student understanding in order to adjust the instructional-learning model, it is only appropriate that it occurs throughout the course and not at the end. It is important to note, however, that assessments administered throughout the course of study are only truly formative if the results are used for the purpose of adjusting learning and instruction. This is an area which is often misconstrued and, as a result, often leads to the ineffective implementation of what is mistakenly thought of as formative assessment.

Summative assessment. A second type of assessment is summative assessment. Summative assessment is set apart from formative and interim assessment in that it typically occurs upon completion of coursework and is used primarily for measuring and evaluating student achievement levels (Cizek, 2010). Although results from summative assessments can be used to inform instruction, due to the timing of the administration, the results can only be used in making educational decisions that will impact future students. These types of assessments are often administered on a large-scale, usually at the state or national level. Summative assessments can carry important consequences for students such as grade retention, grade promotion, or graduation. Based on the current legislation, teachers and administrators also have a great deal at stake in regards to summative assessments as schools can lose funding or even be shut down if repeated failure to meet federally-mandated standards occurs (NCLB, 2002).

Interim Assessment. Interim assessment is somewhat of a mix between the two forms of assessment that have already been discussed. According to Perie et al. (2009), the two key components of interim assessments are that they "evaluate students' knowledge and skills relative to a specific set of academic goals, typically within a limited time frame, and are designed to inform decisions both at the classroom and beyond the classroom level, such as the school or district level" (pp. 6-7). Based on this definition, interim is similar to formative assessment in that it has an informative component and is intended to have instructional implications for the current students. The main difference between the two is that interim assessment occurs less frequently than formative assessment (typically marking the middle or end of a semester) and is also meant to inform administrative and policy decisions, by reporting outcomes of assessment so that comparisons can be made across the school or district. Perie et al. (2009) claim that many of the assessment tools currently being marketed as formative assessment systems are truly interim assessment systems because many of them focus on such things as predicting student scores and benchmarking.

The Future of Testing

In the past few decades, statewide summative assessments have been used as the barometer for student achievement and, as a result, have often become the focus of classroom instruction. This has led to the popular phrase "teaching to the test", which generally implies "that teachers are doing something special to help students do well on a test, often without helping them to better understand the underlying subject matter.(Firestone & Shorrs, 2004, p. 2). As *No Child Left Behind* (2002) requires that every child be tested and that schools be held accountable for student achievement based on testing outcomes, administrators and educators have experienced pressure to increase students' scores on state-run, large-scale summative assessments (Monfils et al, 2004). However, because summative assessments are much easier to administer on a large-scale, such as at the state-level, these currently remain the most realistic mechanisms to evidence student gains for accountability purposes.

Whereas it would not be prudent (or, currently, legal) to eschew large-scale summative assessment, it seems logical that utilizing formative assessments throughout the school year that have been aligned with the state curriculum would benefit students by ensuring that the necessary objectives and goals are mastered prior to taking the end-of-year summative assessment. Recent studies support this notion, calling for further exploration of the relationship between formative and summative assessment, research which could hold implications for a comprehensive assessment system that would incorporate both form, and would better serve the informational needs and educational interests of students, instructors, administrators, and policy-makers (Perie et al, 2009).

The context of the current study is the state of North Carolina. North Carolina, along with 43 other U.S. states and the District of Columbia, has recently adopted the Common Core State Standards (Common core states, 2011). However, as data for this study are from the 2010-11 school year, the formative assessment data in this study are aligned with the Standard Course of Study (SCS), which was the previously held state standard. Formative assessment used throughout the course of the year as a means of evaluating and ensuring student understanding and comprehension of the grade level material could prove to be effective in improving student achievement. If so, end-of-year test results would show evidence of this.

CHAPTER 2: THEORETICAL FRAMEWORK & LITERATURE REVIEW

Theoretical Framework

According to Dylan Wiliam, “increased use of formative assessment is one of the most educationally effective and most cost effective ways of increasing student achievement” (2010, p. 36). His chapter in the *Handbook of Formative Assessment* (2010) provides implications for a new theory of formative assessment, building on some of the already existing literature and theory. This chapter covers several key aspects of formative assessment, explaining how it can optimize learning and instruction in the classroom.

Moments of Contingency

Formative assessment focuses on moments of contingency in instruction in which agents of formative assessment (teachers, peers, and/or students) gather evidence from which to base decisions on how to best regulate the learning process (Furtak, 2005; Stiggins & DuFour, 2009; Wiliam, 2010). Whereas these moments are only a narrow slice of the educational process, they are a vital part which links directly to other important aspects of learning.

According to Wiliam (2010), moments of contingency can be synchronous (e.g. occur during discussion) or asynchronous (i.e. evidence is collected and used to provide feedback or adjust instruction at a later time). It is important to note that although synchronous moments of contingency directly affect the students from which the evidence was collected, asynchronous moments may not necessarily affect the current students. An example of an asynchronous moment of contingency failing to form instruction for the current students would be a teacher using test results from one class to adjust the instructional practices used in a different classroom. In a case such as this, even though the students from which the evidence was collected did not benefit, the assessment could technically be considered to be formative under the broader definition of formative assessment in that the evidence elicited ultimately resulted in the regulation of instruction. Although this is not to say that assessments which create asynchronous moments of contingency cannot be formative, it is fair to say that assessments which create synchronous moments of contingency fit more consistently with the more comprehensive definition of formative assessment in which evidence is elicited and used to make decisions about the next steps to take in the learning process for the current students.

Types of Formative Assessment

Based on the largely agreed upon definition in the literature (Black & Wiliam, 1998; Perie et al, 2007, Cizek, 2010) an assessment can be formative if it informs future instruction and learning. However, formative assessments can function in different ways depending on what type of evidence is elicited. Three types of formative assessment are proposed by Wiliam (2010). The first type of formative assessment proposed by Wiliam is the *monitoring assessment*. The monitoring assessment elicits the least amount of information of the three types, serving only to signal whether or not there has been a lack of understanding between the instructor and student throughout the lesson. An example would be the student’s overall score on a quiz. While the score information indicates the student’s achievement, indicating whether or not instruction was successful, it does not provide insight to the specific problem area(s).

The second type of formative assessment proposed by Wiliam (2010) is *diagnostic assessment*. Diagnostic assessments serve not only to indicate that a problem has occurred in the instructional process, but also to locate the specific area in which the student or students are experiencing trouble. An example of this would be an assessment which the instructor has access to information regarding the test items, thereby allowing for identification of the particular constructs (e.g. goals, objectives, etc) which were not fully comprehended by the student(s).

Whereas diagnostic assessments provide more detailed evidence of learning as compared to monitoring assessments, there still is room to expand. The shortcoming of the diagnostic assessment is its lack of ability to provide insight on how to go about overcoming the student's lack of understanding. This is where the third type of formative assessment comes in: *assessments providing instructionally tractable insights*. Wiliam explains that these type of assessments "situate the problem within a theory of action that can suggest measures that could be taken to improve learning" (2010, p. 27). An assessment that provides instructionally tractable insight elicits evidence of 1) which students are struggling, 2) in which areas they are struggling, 3) as well as the specific miscomprehensions that are causing these struggles. The third component provides traction for the instructor by indicating the next steps to take in attempts to overcome the problem.

It is important to note that the differentiation of the different types of assessment is not necessarily to suggest that all assessments need to be those that provide instructionally tractable insights. While it is true that, in comparison with monitoring or diagnostic assessments, instructionally tractable assessments provide the most detailed feedback, they also often require more time and effort. If a teacher is confident that the students have a good grasp of the material included in the lesson, a monitoring assessment may suffice, serving to provide evidence affirming that the instruction was successful and therefore no additional assessment is required. This illustrates one of Wiliam's (2010) key points in regard to formative assessment; that the evidence elicited through formative assessment does not necessarily need to lead to change in instructional practice. Whereas one of the primary purposes of formative assessment is to provide evidence to inform future instruction, affirmation of success in the current practice, although perhaps not leading to a change in instruction, still qualifies as formative in that the decision to continue forward with the current instructional practice is based on evidence gathered through assessment of the students' understanding.

Informative Questioning Cycle

Furtak (2005) proposed a three-step informative questioning cycle which involves a continuous elicitation of student understanding while moving towards specific learning goals. The first step in this cycle is 1) *eliciting* responses from students in order to understand where they stand in their learning. In a continuous context, this could occur in classroom discussion or question and answer sessions. Furtak suggests 12 types of questions for teachers to use in the process of eliciting evidence of understanding from students. A few examples of Furtak's types of questions for eliciting evidence of understanding are questions that formulate explanations, interpret data or patterns, compare/contrast others' ideas, elaborate, take votes on ideas, share predictions, and define concepts. Furtak suggests that teachers "use the list as a source of suggestions, tailoring the questions to fit their own activities" (2005, p. 23).

Once evidence is elicited, teachers should *recognize* the students' responses, acknowledging what the student has said and how the response fits in with the current classroom discussion. If a student's response provides evidence that a misunderstanding has occurred, the teacher should then return to the first step and elicit more information in order to understand where the misconception may be grounded in order to help move the student towards the learning goal.

The third part of the informative questioning cycle is to take action based on the evidence collected in the first two steps of the cycle. In this step, the teacher encourages students to discuss their conceptions in an attempt to reach a common understanding based on the facts and evidence that was involved in the discussion. Furtak provides four guiding question types for teachers to help initiate the action part of the cycle. Types of acting questions are those which promote argumentation, help relate evidence to explanations, provide descriptive or helpful feedback, or promote making sense (2005, p. 24).

Decisions for Future Action

Closely in line with the informative questioning cycle proposed by Furtak, Wiliam (2010), Ramaprasad (1983) indicates three keys to the instructional process based on the systems approach to formative assessment:

1) establishing where the learners are going;

2) establishing where the learners are in their learning; and

3) establishing what needs to be done to get them there.

Whereas the first step in the instructional process is obviously establishing and teaching the curriculum, the next logical step is to establish whether or not learning has occurred. This is where formative assessment fits into the instructional process. In order for an assessment to improve learning it must elicit evidence of knowledge from the student which can then be used to inform the instructor's decisions for future action. The results of a well-constructed formative assessment theoretically provide the instructor with the evidence necessary to make an informed decision as to what the next course of action should be in the learning process. Based on this, formative assessment can account for a great deal of the instructional process by providing evidence on where the students are in their learning and, if constructed in a manner which measures achievement on specific learning goals or objectives, what needs to be done to get them to the goal established at the beginning of the lesson. Formative assessment, therefore, can serve a major role in the instructional process, helping to increase the efficiency and efficacy of instruction.

Levels of Assessment

Stiggins and DuFour (2009) provide implications for assessment to be used formatively at three different levels – classroom, school/program, and institutional/accountability level. The type of evidence elicited and future action varies by level, but they all work together to drive success of students, instructors, and schools.

Classroom-level assessments. At the classroom level, the students and teachers are the primary benefactors of assessment. It is necessary for these agents in the educational process to have knowledge about where the students are in their learning and what the next steps are in the learning process. Teachers should be clear with the students about the standards and learning progression that will take place over time. Stiggins and DuFour suggest that “a balanced classroom assessment environment uses some assessments in a formative manner to support learning and some in a summative way to verify it, as at grading time” (2009, p. 641). The authors emphasize that learning is a continuous process which takes place over time, not instantaneously and as such, formative assessments should occur continuously in order to keep track of where each student is in his or her learning trajectory.

Stiggins and DuFour also suggest that the development of common assessments to be used across classrooms may be helpful in identifying struggling students among different classrooms within the same school. This notion leads into the next level of assessments.

School-level assessments. While Stiggins and DuFour (2009) suggest that cross-classroom common assessments may be helpful for individual students, they are an absolutely imperative part of evaluating strengths and weaknesses in the current curriculum and instruction being practiced at the school. Results from school-level common assessments help to "identify components of an instructional program that are working effectively and those that are not" (p. 641). The authors indicate that teachers within the school should collaborate to form common assessments which address three formative purposes: 1) to identify curricular areas in which many students are struggling, 2) clarify each instructor's individual strengths and weaknesses, and 3) identify students who are in need of systematic interventions.

Institutional-level assessments. The last level described by Stiggins and DuFour draws attention to the need for accountability tests in order to provide evidence of institutional impact for superintendents, school boards, and legislators. The scope of these assessments is much larger than assessments at the other two levels, but this is only dictated by the scope that those using the evidence are making decisions from. In other words, the large scope of institutional-level assessments is necessary due to the audience it is intended to inform. As mentioned previously, although assessments at this level are typically considered summative, as long as the results from summative assessments are used to inform future instructional practice, they can still be considered formative. At best, they would be considered asynchronous formative assessments under Wiliam's (2010) definition.

Cycle Length

Building on the previous section, while interim or summative assessments can function formatively if the information gathered from them is used to inform future instruction, it is important to recognize that differences in cycle length exist for each of these types of assessments (Wiliam, 2010). The three formative assessment cycle lengths indicated by Wiliam and Thompson (2007) are 1) Long-Cycle, 2) Medium-Cycle, and 3) Short-Cycle. Descriptions of each of these can be found in Table 2.1. These categorical designations are intended to describe the length of the feedback loop – the time from assessment until the results are actionable. Furthermore, the current literature supports the popular assumption that shorter assessment cycle lengths are more likely to increase learning in students, while assessments with longer cycles are not likely to have much of an impact on learning (Cowie & Bell, 1999; Looney, 2005; Shephard 2007; Wiliam, 2010).

Table 2.1 *Focus and Cycle Lengths for Types of Formative Assessment*

Type	Focus	Length
Long-Cycle	Across marking periods, semesters, quarters, years	4 weeks to 1 year
Medium-Cycle	Within and between instructional units	1 to 4 weeks
Short-Cycle	Within and between lessons	Day by day; 24 to 48 hours Minute by minute: 5 seconds to 2 hours

Sensitivity to Instruction

One major aspect to consider when comparing different types of assessment is the assessment's sensitivity to instruction. Sensitivity to instruction refers to how closely an assessment measures the effects of instruction (Wiliam, 2010). Wiliam points out that "the learning outcome measures used in different studies are likely to differ significantly in their sensitivity to instruction" (p. 21). Furthermore, he identifies the assessment's distance from the curriculum it is intended to assess as the primary determinant of its sensitivity to instruction. Ruiz-Primo, Shavelson, Hamilton, and Klein (2002) proposed five categories indicating distance from curriculum as a standard for measuring sensitivity to instruction. The five categories are presented in Table 2.2.

Table 2.2 *Distance from Instruction Classifications*

Category	Example
1. Immediate	Science journals, notebooks, and classroom tests
2. Close	Formal embedded assessments
3. Proximal	Different assessment of the same concept, requiring some transfer
4. Distal	A large-scale assessment from a state assessment framework, in which the assessment task was sampled from a different domain, such as physical science, and where the problem, procedures, materials and measurement methods differed from those used in the original activities
5. Remote	Standardized national achievement tests

The results from Ruiz-Primo, et al (2002) illustrate the inverse relationship between distance from curriculum and sensitivity to instruction, suggesting that the closer the assessment is to the enactment of curriculum, the greater the sensitivity to instruction. In comparing the average effect size of a proximal intervention (.26) with the average effect size of a close intervention (1.26) their study illustrated the impact that distance from instruction can have on student outcomes.

Summary

As shown in the theoretical framework section, the current theories and literature are intertwined, all involving different aspects of the assessment process. From the setting to the intended purpose, comparisons between Wiliam's types of assessment, Furtak's assessment cycle, and Stiggins and DuFour's three levels of assessment illustrate a general understanding among researchers in the field of formative assessment.

Literature Review

Many studies since the late 1980s have shown the positive effects that formative assessment can have on student achievement (Burns et al, 2010; Bergan et al., 1991; Black & Wiliam, 1998; Foster & Poppers, 2009; Fuchs & Fuchs, 1986; Martinez & Martinez, 1992; Miesels et al, 2003; Peterson & Vali Siadat, 2009; Sadler, 1989; White & Frederiksen, 1998). However, federally-mandated summative assessments have remained the primary mechanism for measuring student knowledge. Studies have been conducted at all levels, from kindergarten through college, and have shown that formative assessment has the potential to close achievement gaps while raising student achievement overall (Fuchs & Fuchs, 1986). Furthermore, it has been shown that formative assessment is capable of compensating for differences in instructor ability (Martinez & Martinez, 1992). This is an area in which formative assessment could potentially help by providing relief to schools which have historically had trouble attracting the nation's best educators.

Closing Achievement Gaps

Past research has shown that formative assessment, implemented in various forms, has the ability to increase student achievement (Burns et al, 2010; Bergan et al., 1991; Black & Wiliam, 1998; Foster & Poppers, 2009; Fuchs & Fuchs, 1986; Martinez & Martinez, 1992; Miesels et al, 2003; Peterson & Vali Siadat, 2009; Sadler, 1989; White & Frederiksen, 1998).

Of particular concern in education are the achievement gaps that currently exist among different subgroups in the United States including gender, racial and ethnic minorities, English language learners, students with disabilities, and students from low-income families (National Education Association, 2012). Many studies have addressed gaps in achievement between subgroups by exploring the effects of formative assessment when applied to these at-risk populations.

Meisels et al (2003) investigated the effects of a curriculum-embedded performance system – Work Sample System (WSS) – on the change in student summative scores from third to fourth grade. The focus of their study was on a sample of students from a low-income, urban school district in Pittsburgh, Pennsylvania. The sample matched 96 students in WSS classrooms with 116 students in non-WSS classrooms by demographic characteristics. Classrooms were matched as closely as possible on race, income, mobility, school size, and number of parents in the home. The two comparison groups were also compared to the 2,922 third and fourth grade students in the Pittsburgh Public school district from 1996-98.

The study compared student's change in score on the Iowa Test of Basic Skills (ITBS) from third to fourth grade. The researchers utilized longitudinal data based on the presumption that the students' raw scores would be comparatively low due to the negative effects typically seen in low-income school districts. Using longitudinal data, therefore, would allow student gains to be evidenced regardless of score.

The WSS was a "curriculum-embedded performance assessment" designed for children from Pre-school to grade 5. The system involved a very rich and in-depth systematic approach to data collection involving information on instruction in the classroom, teachers' perceptions of students, as well as involving students and parents in the learning and assessment process.

The results from a three-step hierarchical regression, controlling for previous score, indicated that students in WSS classrooms displayed greater gains (27 and 20 points for reading and math, respectively) than their demographically matched comparison group (0 and 6 points for reading and math), as well as all other public school students at the same grade level (15 and 17 for reading and math). Furthermore, gains were shown for students who started with high skills as well as students who started with lower skills (Meisels et al, 2003). This evidence supports the notion that formative assessment may provide benefits for students coming from economically disadvantaged backgrounds.

Also addressing an at-risk population, Fuchs and Fuchs (1986) conducted a meta-analysis of educational research designed to measure the effects of formative assessment in the classroom on children with disabilities. The age levels ranged from preschool to grade twelve. Each of the studies included in the meta-analysis, which included comparisons between experimental and control groups, produced relatively large effect sizes of .70 or higher. These results suggest the potential effectiveness of classroom formative assessment on student achievement. In addition, the research design indicates that using student gains may be helpful when measuring effectiveness of an intervention on student populations that may historically perform lower than their peers in order to place emphasis on growth as opposed to overall achievement.

A few interesting findings stand out from Fuchs and Fuchs' (1986) results, each having implications for the area of formative assessment. First, although significant relationships were found in both experimental groups—that is, the group of teachers who collaborated with colleagues following collection of formative assessment information, and the group who collected formative assessment information but without the collaboration element--a stark difference was noted between those groups. Classrooms in which collaboration occurred displayed a much larger effect size compared to classrooms in which collaboration did not occur (effect sizes =.92 and .42, respectively). This difference displays the importance of collaboration between teachers in evaluating student understanding and planning towards future instruction. The authors also found that although all students benefited from the implementation of formative assessment, students with mild mental disabilities displayed greater gains.

Addressing another at-risk population, Bergan et al (1991) studied a sample of 838 kindergarten children who came from economically disadvantaged domestic homes. The sample was taken from six states, seven different districts, and 21 different schools. This study looked at the effects of a measurement and planning system (MAPS) on the promotion of students into traditional or special education tracks. The study was implemented over the course of eight weeks and produced results in support of formative assessment. Following the eight week evaluation period, the experimental group showed significantly higher gains in reading, math, and science as compared to the control group. In the control group one in every five students was placed in special education as compared to one in every 71 students in the experimental group. This finding is startling given the ramifications of placing a child in special education at the beginning of his or her schooling (Byrnes & Yamamoto, 1984; Niklason, 1984; Peterson, DeGracie, & Ayabe, 1987).

Frequency of Testing

The current literature provides evidence regarding the effects of assessment frequency on student achievement (Martinez & Martinez, 1992; Peterson & Siadat, 2009).

In a study investigating the effects of formative assessment on student achievement in a college-level elementary algebra class, Peterson and Siadat (2009) used a pre-post test design in order to measure student achievement over time. Data for this study was collected over three years from multiple iterations of the same course taught by 25 different instructors, but with the same materials, academic objectives, textbook, content, and homework. In addition, all students took the common midterms, finals, and COMPASS test. The only other difference between classes was the presence of frequent formative assessment with immediate feedback.

The sample was representative, consisting of 1,574 students of mixed gender, race, ethnicity, ability, and economic background. Students self-assigned to instructional groups upon enrollment. This self-assignment resulted in a disproportionally smaller test group of 222, students placed in ten sections taught by two instructors, whereas the control group consisted of 1,352 students, placed in 50 sections, taught by 23 different instructors. The instructors of the sections in the test group received specific training on how to formatively assess the students throughout the course.

The pre-test results indicated that all students, regardless of group were essentially equal in terms of knowledge of the course material at the beginning of the semester. Whereas all students evidenced gains from the beginning to the end of the course, students in the test groups showed greater achievement from pre to post-test. In addition, the majority of the students in the test groups passed the COMPASS exam, qualifying them to proceed to the next level math course. On the other hand, the majority of the students in the control groups did not pass the COMPASS exam.

The findings of Peterson and Siadat (2009) support the common claim that formative assessment can increase student achievement. Particularly, "formative assessment in the form of frequent, cumulative, time-restricted, multiple-choice quizzes with the immediate constructive feedback reveals the levels of conceptual understanding in a timely manner and improves student academic performance on the summative assessment instruments" (p. 100). Although the results of this study suggest that frequent formative assessment increases student achievement, the authors found that an increased frequency of formative assessment failed to produce a significant improvement in students' learning outcomes. This discovery was unexpected and contrary to other literature regarding frequency of formative assessment (Martinez & Martinez, 1992). The results of this study indicate that additional research in the area of frequency of assessment would be beneficial to the discourse on formative assessment.

Research conducted by Martinez and Martinez (1992) addressed the issue of differential teacher effectiveness. This problem in education is especially pertinent in low-income areas including predominantly urban and rural schools, which historically have had difficulty attracting top quality teachers. This research consisted of a sample of 120 students taking an introductory Algebra course over a period of 18 weeks at a U.S. college. The experimental design included four classes, two of which were control groups, receiving only one assessment at the end of each chapter, and two experimental groups that were assessed three times throughout each of the seven chapters. In addition to the differential in testing frequency, two groups were taught by an average, relatively inexperienced instructor, while the other two classes were taught by an instructor with extensive teaching experience, and a history of outstanding ratings from past courses.

Results from this study showed that frequency of assessment has a positive effect on achievement, although the gains were much smaller for the experienced teacher in comparison to the average instructor. The authors speculated that the higher achievement but lower overall gain observed for students in courses taught by the experienced instructor was due to the instructor’s ability to formatively assess his students without administering a formal assessment. That is, he was able to gauge student understanding through other means such as discussion, and therefore, account for the difference in frequency of assessment. The authors (Martinez & Martinez, 1992) presented implications for the previously mentioned under-funded and under-staffed schools, suggesting that the implementation of formative assessment in the classroom could compensate for the lack of highly experienced, expert teachers.

Formative Assessment and Instructional Technology

Over the past few decades, technology has been taking over the world of education. The recent interest in formative assessment has spurred testing companies have been turning out new technology-enhanced assessment tools meant to improve student performance and streamline the data collection process. In response to the release of these products, many researchers have begun to address the effectiveness of technology-enhanced formative assessments on student achievement (Burns et al, 2010; Kingston & Nash, 2011).

Kingston and Nash (2011) reviewed more than 300 studies in a meta-analysis of existent research that addresses the effects of formative assessment in kindergarten through grade 12. Of the 300-plus studies reviewed, only 13 studies with a total of 42 effect sizes were deemed acceptable based on a five-point criteria: 1) intervention had to be described as *formative* or *assessment for learning*, 2) participants had to be from a K-12 setting, 3) a control or comparison group design must have been used, 4) the appropriate statistics required for effect size must have been provided, and 5) the study had to have been published in 1988 or later. Of the 42 effect sizes selected for the meta-analysis, 19 were based on math formative assessment, 12 on reading, language arts, or writing, 10 on science, and one on music.

Kingston and Nash (2011) categorized each of the 42 effect sizes into five different treatment type categories based on common themes found throughout the literature review. The five categories were: 1) professional development, 2) curriculum-embedded assessment systems, 3) use of a computer-based formative assessment system, 4) use of student feedback, or 5) other types of formative assessment. Of the 42 effect sizes, 23 used professional development as the treatment, seven used curriculum-embedded assessment, six used computer-based formative assessment system, and the student feedback and other types of formative assessment categories accounted for three effect sizes each.

The results from this study, utilizing a random effects meta-analytic approach, produced a weighted mean effect size of .20 and a median effect size of .25. Although Kingston and Nash point out that these effect size estimates are markedly lower than the effect-sizes claimed by the oft-cited study by Black and Wiliam (1998b), their results still suggest that formative assessment has a significantly positive effect on student learning (Kingston & Nash, 2011).

In addition to overall effect size, the results from this study indicated that mean effects of formative assessment were moderated by both content area and treatment type.

Reading produced the largest mean effect size (.32) with math and science producing mean effect sizes of .17 and .09, respectively. Studies involving professional development activities as the treatment produced the largest mean effect among treatment types (.30). However, the authors note that all of the effect sizes for this treatment type came from the same study, in which the authors point out that the findings were difficult to interpret due to methodological issues (Wiliam et al, 2004). Computer-based formative assessments were also shown to have a moderately large mean effect size (.28). This implication is quite relevant in that technology continues to pervade the educational process. These findings may indicate that technology and formative assessment can be a formidable team of tools for teachers to utilize in the classroom.

Burns et al. (2010) conducted a study that examined the effects of technology-enhanced formative evaluation (TEFE) on student achievement. The authors hypothesized that computer-based formative assessment would increase student achievement as it would allow instructors to implement effective formative assessment programs in their classrooms without requiring a great deal of additional time and effort. The study included 360 non-charter elementary schools from across four geographically distinct states in the U.S. (Florida, Minnesota, New York, and Texas) which were randomly selected from a list of schools who had previously ordered the TEFE program from the publisher. The mean enrollment across schools was 522.62. The study examined if a higher percentage of students at schools using a TEFE system scored at the proficient level or higher on state-wide summative assessments.

The TEFE system for this study was a program called AM which is designed to "monitor student progress towards instructional goals and manage student practice of relevant instructional tasks" (Burns et al., 2010, p. 586). The program provides students with tasks and provides feedback as they complete them. The program allows the student to work relatively independently while sending data and monitoring reports to the teacher.

The study design addressed the duration of program implementation at each school. The schools in the study were classified as having used a TEFE system for 1) one year to four years and eleven months, 2) five years or more, or 3) not at all. In addition, relevant school-level variables were collected including: percent proficient in reading and math, student enrollment, average student/teacher ratio, percent free or reduced lunch (FRL), as well as ethnic variables.

Results showed that, in schools with TEFE programs, a higher percentage of students scored at or above the proficient level on the end-of-year state summative assessments as compared to students in schools that did not have a TEFE program at all. In particular, the schools that had been using a TEFE program for five or more years produced a rather large effect size of .78. Schools that had the TEFE program in place for one year to 4 years and 11 months produced an effect size of .51. This evidence suggests that not only are students at schools with TEFE programs more likely to achieve proficient or higher score-levels on state summative assessments, but students at schools that have had a TEFE program in place for a longer time show greater achievement. In addition, Burns et al found no significant difference in student achievement between race groups in schools that had a TEFE program for five or more years. This, however, was not the case with schools that did not have a TEFE program at all. In these schools the disparity in achievement between White students and minorities that has come to be expected in educational research remained true.

This study provides traction for additional research on the effects of technology-enhanced formative assessment on student summative assessment performance. In addition, additional investigation towards the ability of formative assessment to close the achievement gap would be beneficial to the literature.

Expanding the Literature

The existing literature provides a sufficient base knowledge from which further exploration on the effects of formative assessment in the classroom can build upon. Evidence from past research has shown that formative assessment has the potential to result in positive gains for all students, and possibly even more so for disadvantaged individuals (Black & Wiliam, 1998; Burns et al, 2010; Fuchs & Fuchs, 1986; Martinez & Martinez, 1992; Meisel et al, 2003; Peterson & Siadat, 2009; Sadler, 1998). However, there is certainly room for expansion in the research. Perie et al. (2009) called for future researchers to examine the ability of formative assessments to improve achievement on summative assessments. The study described in this thesis addresses these issues. Student achievement on the North Carolina end-of-grade (EOG) reading and mathematics assessments, specifically the change in scale score from 2009-2010 to 2010-2011, were compared with frequency of OFAP assessments taken throughout the 2010-11 school year.

In addition, as this study is multi-level, it is designed to investigate potential variables at both the school and student level. Student and school identification numbers were used to link the OFAP data to the corresponding student end-of-grade reading and mathematics data. This research is designed to address these areas, therefore providing insight into previously unexplored aspects of formative assessment.

Research Questions

This study was designed to examine student formative and summative assessment data for potential differences in student achievement based on number of formative assessments taken. In addition, assessments cycle-length categories were used to determine if results are consistent based on assessment cycle-length. The cycle-length variable was determined based on the timing and intention of the assessments (classroom quiz – short-cycle vs. district-wide benchmark assessment – long-cycle). Controls were included for gender, race, economically disadvantaged student (EDS), students with limited English proficiency (LEP), and students with disability (SWD). This thesis attempted to address these issues through the following research questions:

1. What are the effects of formative assessment frequency on student performance on reading and mathematics summative assessments for middle school students (grades six through eight)?

2. Do the effects of the formative assessment frequency differ based on assessment cycle-length?
3. Do the effects of formative assessment frequency differ for student subgroups (gender, race, EDS. LEP, and SWD)?

CHAPTER THREE: METHODS & PROCEDURES

Method

This study was conducted to gather evidence regarding the effects of formative assessment on student achievement. Specifically, the relationship between frequency of formative assessment and student gains on state-mandated, end-of-grade assessment were investigated. The study used student usage data from an Online Formative Assessment System (OFAP) and existing end-of-grade (EOG) Math and Reading assessment data. The data and methods are described in this chapter.

Participants

Participants were middle school students (grades 6, 7, or 8) in 2010-11 who took the North Carolina End-of-Grade (EOG) mathematics and/or reading assessment in the 2010-11 school year, and who were enrolled in a North Carolina Public school that used an online formative assessment program (OFAP) in the 2010-2011 school year. This study only included students from schools that had received formal training on how to use the OFAP to assess students in a formative manner. The training requirement was included in order to increase the validity of any claims regarding the effectiveness of formative assessment on student achievement as evidenced by use of the OFAP. Although the dataset provided no delineation between the different types of training offered by the OFAP provider, Table 3.1 provides a detailed description of each training type.

Table 3.1 - Training Types Offered by OFAP Provider

Test Administrator Training	- Designed for school-based leaders such as principals, assistant principals, curriculum coaches, technology facilitators, media specialists, who will be responsible for monitoring OFAP usage; - Focuses on the “back-end” of the program including bulk uploading student data, setting up teacher accounts, creating common assessments, viewing reports, and monitoring system usage - Occurs early in the school year (August or September) - 1-3 participants, 3 hours minimum, computer lab
Basic User Training	- Designed for classroom teachers who will be using the OFAP - Focuses on user basics such as setting up classes, creating and scheduling assessments, and viewing reports - Occurs after the successful completion of the Test Administrator Training (August, September, or October) - Maximum 25 participants, 2 hours minimum, computer lab
Reports/Data Analysis	- Designed for teachers and school staff who have given OFAP assessments - Focuses on how teachers can analyze data from the OFAP reports to determine instructional effectiveness, identify student and classroom needs, and create instructional intervention plans - Occurs at least one month after the successful completion of the Basic User Training (October-February) - Staff must have access to individual classroom data - Maximum 25 participants, 2 hours minimum, computer lab
Refresher	- Designed for teachers and school staff who have had prior OFAP training and need to refresh their skills - Covers setting up classes, creating and scheduling assessments, and analyzing reports - Can occur at any time during the school year - Maximum 25 participants, 3 hours minimum, computer lab
Customized/A La Carte Training Option	- A La Carte trainings are on-site training sessions and professional development workshops that can be ordered separately as needed. The A La Carte choices include: o Test Adminstrator Training o Basic User Training o Reports/Data Analysis Training o District Benchmark Administrator Training o Custom Options: designed for individual school or district needs

Student-level achievement data were provided by the North Carolina Department of Public Instruction (NCDPI) through the Division of Accountability Services. Student-level formative assessment data were obtained from a North Carolina-based OFAP provider. These two datasets were merged using unique student identification numbers. The final dataset included one observation for every sixth-, seventh-, and eighth-grade student who participated in the EOG Reading and Mathematics assessments in 2010-2011 and who was administered at least one Reading or Mathematics assessment using the OFAP. The total sample included 83,799 students at 413 schools. (Descriptive statistics for the demographic variables are included in Table 4.2 of the next chapter.)

Student Achievement Data

Dependent Variables. As mentioned previously, the NCDPI provided the student achievement data for this study. The measure used to represent student achievement was an academic growth score. Students had one growth score for each subject (mathematics and reading). These scores were calculated by NCDPI and included in the dataset provided for this study. The growth scores, referred to from here on as "AC-Scores" (academic change), measured each student's relative growth in Mathematics and/or Reading in comparison to their performance on the EOG assessment for the given subject in the two prior academic years. AC-Scores are based on an academic change scale, or C-Scale, which is defined by the NCDPI as, "a standardized scale, similar to z-scores, to measure student performance relative to standard performance for that grade level in a standard setting year" (North Carolina Department of Public Instruction\Accountability Services, 2011) . The formulas used to calculate these scores are presented in Table 3.1. Basic descriptive statistics for the AC-Scores are shown in Table 4.1 of the following chapter.

Table 3.2 *Academic Change Score Calculation*

	Variable Notation	Definition
Formula for AC-Score:	AC-Score = C-Score$_{\text{c-scale}}$ – (0.92 x ATPA$_{\text{c-scale}}$)	
Where:	AC	Academic Change
	ATPA	Average of two previous assessment Change Scores
	C-Score	Change Score on C-Scale for current year
Formula for C-Score:	C-Score = [(DSS) – (mean, SS year)] / (standard deviation, SS year)	
Where:	DSS	Developmental Scale Score
	SS Year	Standard Setting year for given assessment

Note. 0.92 in the AC-Score formula accounts for regression to the mean.

At-risk student control variables. The NCDPI dataset also included several student-level indicators for students with disadvantaged backgrounds. These included the following:

1) students with an economic disadvantage (EDS);
2) students with limited English proficiency (LEP); and
3) students with a learning disability (SWD).

Additional information on these subgroups can be found in the *Guide to Career and Technical Education's Special Populations – Challenge Handbook* on the NCDPI website (North Carolina Department of Public Instruction, 2011). The demographics for these indicators are shown in Table 4.2 of the next chapter. Based on the at-risk subgroups identified by the NEA (2012), gender and race were also included as control variables for this study.

Dummy variables were created for each of these variables for analysis. EDS, SWD, and LEP students were coded 1 and students not at-risk were coded 0. For gender, males were the reference group (coded 0) and females were coded 1. Caucasians served as the reference group for race (coded 0) and dummy variables were created for each of the other racial groups (Asian, African American, Hispanic, and Other) where 1 indicates that the individual belongs to that category.

Formative Assessment Data

Independent variables. The data provided by the OFAP included one observation for every North Carolina student in sixth, seventh, or eighth grade, who took at least one math or reading assessment using the OFAP in the 2010-2011 school year. Each student record included the total number of formative assessment administrations by subject. Examples of the formative assessment items as well as an objective-based report sample are provided in Figures A1 through A3 of Appendix A. In addition, assessments administered using the OFAP were classified into two categories to indicate differing cycle length. The two categories – short-cycle and long-cycle – were designated based on the assessment cycle categories developed by Wiliam and Thompson (2007) and illustrated in Table 2.2 of the previous chapter. A detailed description of the difference between the two assessment classifications is provided in the next section of this chapter. The total number of mathematics and reading assessments given for each of the cycle-length categories was also included in order to investigate whether assessment cycle-length has a significant effect on the relationship between formative assessment frequency and student achievement. Basic descriptive statics for the OFAP assessment data are shown in Table 4.1.

Distinguishing between short and long-cycle assessments. The type of assessment available in the OFAP that was classified as a short-cycle assessment (SCA) comes in the form of 10-15 item quizzes, each aligned with a specific objective included in the North

Carolina Standard Course of Study which can be found on the NCDPI website. The short cycle-length designation was ascribed based on the nature of the assessment (i.e. assessment administered by the instructor in the classroom) and the feedback loop. SCAs are available as pre-packaged quizzes and also as customizable quizzes. Pre-packaged quizzes are 10-question quizzes constructed by the OFAP contractor and made available to all instructors at OFAP enrolled schools, for classroom use. Instructors also have the ability to construct their own objectives-based quizzes, using items from the OFAP item bank. Each question in the OFAP item bank is designated by objective and difficulty-level, which have been ascribed by item writers, contracted by the OFAP provider. Before items can be added to the OFAP item pool, they are vetted by multiple educational professionals (also contracted by the OFAP) in order to ensure that the appropriate difficulty level and objective has been designated. The provision of these designations allows teachers to design assessments specifically to suit the needs of individual students.

The other type of assessment offered by the OFAP is the benchmark assessment. Given the nature of the OFAP benchmark assessments (i.e. assessment administered at the school-level) and the longer feedback loop, this particular assessment type was classified as a long-cycle assessment (LCAs). The OFAP LCAs are administered at the school or district-level and typically consist of 30-50 items which cover a range of objectives covered throughout a unit of instruction. LCAs typically mark the end of a quarter or semester. All students of the same grade level in the school or district are given the same LCA and results from these assessments are made available to administrators for the purpose of tracking student progress based on district-wide benchmarks.

In summary, the primary differences between SCAs and LCAs are 1) the length of the assessment (10-15 items vs. 30-50 items), 2) the breadth of material included (single objective vs. multiple objectives, and 3) the ability to tailor assessments to particular students' needs. By distinguishing SCAs from LCAs, any differences in student mathematics and reading achievement based on the cycle-length that was used to assess the student should be evidenced.

Statistical Methods

Given the nested nature of educational data, a multi-level model approach was employed to address the following specific research questions:

1. What are the effects of formative assessment frequency on student AC-Score for each subject?
2. Do the effects of the formative assessment frequency differ based on assessment cycle-length?
3. Do the effects of the formative assessment frequency differ for at-risk student subgroups (gender, race, EDS, LEP, and SWD)?
4. Do the effects of the formative assessment frequency differ based on school-level at-risk characteristics (%EDS, %Minority)?
5. Do the effects of the formative assessment frequency differ based on school-level assessment cycle length characteristics (Mean SCAs, Mean LCAs)?

To address these specific research questions, a multi-level model was constructed for each content area. As the model was constructed, each research question was addressed by testing for statistical significance for each of the specific relationships. Math AC-Score was the dependent variable for all models estimating student math achievement. Reading AC-Score was the dependent variable for all models estimating student reading achievement.

Unconditional Means Model

Each model was built from the bottom up. The first step was fitting an unconditional model for each content area. The unconditional model estimates the dependent variable without consideration of level 1 or level 2 predictors. These estimates provide a reference point for comparison to more parameterized models. Table 3.2 presents the model in two common forms. The Multi-Level model presents equations for each level whereas the Mixed-Effects model presents one single level-1 equation in which γ_{00} represents the grand mean AC-Score for the given subject across all students, u_{0j} represents the variability in AC-Score between schools, and r_{ij} represents the variability in AC-Score between students (i.e. the random error associated with ith student in the jth school). The term Y_{ij} represents the estimated student AC-Score for the given content area in both models and the term β_{0j} represents the sum of an intercept for the student's school in the Multi-Level Model. Both forms of the model formula were provided in this initial presentation to provide a reference point for readers who may only be familiar with one form or the other. From here on the formulas will be presented in the mixed-effects model format only.

Table 3.3 *Unconditional Model*

Multi-Level Model	Mixed-Effects Model
Level 1: $Y_{ij} = \beta_{0j} + r_{ij}$ Level 2: $\beta_{0j} = \gamma_{00} + u_{0j}$	$Yij = \gamma_{00} + u_{0j} + r_{ij}$

Intraclass Correlation Coefficient. The unconditional model, in addition to providing a reference point from which to compare more complex models, also provides the information necessary to calculate the intraclass correlation coefficient (ICC) and design effect. According to Hox (2002, p. 15), "ICC is the proportion of variance that can be explained by the clustering or grouping structure." The equation for ICC is shown in Equation 1. Design effect is defined by McCoach (2010, p. 134) as, "the degree to which the parameter estimates' standard errors are underestimated when assuming independence." The equation for design effect utilizes the ICC and average cluster size, and is illustrated in Equation 2.

$$\rho = \frac{\sigma_{u0}^2}{(\sigma_{u0}^2 + \sigma_r^2)} \tag{1}$$

where: ρ = intraclass correlation coefficient
σ_{u0}^2 = between-school variance
σ_r^2 = variance between students within schools

$$design\ effect\ = \sqrt{1 + \rho(\bar{n}_j - 1)} \tag{2}$$

where: ρ = intraclass correlation coefficient
$\bar{n}_j$ = mean school size

Random Coefficients Model

The second step of the model building process was to estimate a random coefficients model in which only level-1 predictors were included. In order to address Research Question 1, each random coefficients model was fit with a variable representing the total number of formative assessments taken along with control variables for gender, race, and at-risk students (EDS, LEP, and SWD). The formative assessment frequency, EDS, LEP, and SWD variables were all initially estimated as randomly varying by school. Any variance components determined to be statistically non-significantly different from 0 were then fixed. The control variables for race and gender were estimated as fixed across schools. All statistically significant variables were retained in the model. To address Research Question 2, another random coefficients model was fit in which the total number of assessments variable was replaced by two assessment count variables – one for short-cycle assessments (SCAs) and one for long-cycle assessments (LCAs). The model used going forward (total assessment count model vs. assessment count by cycle-length model) was the model with the better fit to the data. The equation for the total number of assessments (regardless of cycle-length) is presented in Table 3.3. The equation for the cycle-length specific model is presented in Table 3.4. It is important to note that, whereas all of these variables were present in the initial iteration of the analysis, any variables that were determined to be non-statistically significant were subsequently eliminated from the model. Therefore, the formulas presented below are the starting point and are subject to change based on statistical evidence.

Table 3.4 *Random Coefficient – Mixed-Effects Model (Total Assessments)*

$$Yij = \quad [\gamma_{00} + \gamma_{10}(\text{TotalAssmts})_{ij} + \gamma_{20}(\text{Gender})_{ij} + \gamma_{30}(\text{Asian})_{ij} + \gamma_{40}(\text{AfrAm})_{ij} + \gamma_{50}(\text{Hisp})_{ij} + \gamma_{60}(\text{Other})_{ij} + \gamma_{70}(\text{EDS})_{ij} + \gamma_{80}(\text{LEP})_{ij} + \gamma_{90}(\text{SWD})_{ij}] + $$

(*fixed effects*)

$$[u_{0j} + u_{1j}(\text{TotalAssmts})_{ij} + u_{2j}(\text{EDS})_{ij} + u_{3j}(\text{LEP})_{ij} + u_{4j}(\text{SWD})_{ij} + r_{ij}]$$

(*random effects*)

Where:

γ_{00}	school mean AC-Score (intercept) when all other predictors are 0
$\gamma_{10}(\text{TotalAssmts})_{ij}$	slope for total number of formative assessments predictor
$\gamma_{20}(\text{Gender})_{ij}$	slope for gender
$\gamma_{30}(\text{Asian})_{ij}$	slope for Asian students
$\gamma_{40}(\text{AfrAm})_{ij}$	slope for African American students
$\gamma_{50}(\text{Hisp})_{ij}$	slope for Hispanic students
$\gamma_{60}(\text{Other})_{ij}$	slope for students of Other race/ethnicity
$\gamma_{70}(\text{EDS})_{ij}$	slope for EDS students
$\gamma_{80}(\text{LEP})_{ij}$	slope for LEP students
$\gamma_{90}(\text{SWD})_{ij}$	slope for SWD students
$u_{1j}(\text{TotalAssmts})_{ij}$	variability in slope for total number of formative assessments
$u_{2j}(\text{EDS})_{ij}$	variability in slope for EDS students
$u_{3j}(\text{LEP})_{ij}$	variability in slope for LEP students
$u_{4j}(\text{SWD})_{ij}$	variability in slope for SWD students

Note: The terms Yij, u_{0j}, and r_{ij} were defined previously in this chapter;

Table 3.5 *Random Coefficient – Mixed-Effects Model (Cycle-Length Specific)*

$$Yij = \begin{array}{c}[\gamma_{00} + \gamma_{10}(\text{SCAs})_{ij} + \gamma_{20}(\text{LCAs})_{ij} + \gamma_{30}(\text{Gender})_{ij} + \gamma_{40}(\text{Asian})_{ij} + \gamma_{50}(\text{AfrAm})_{ij} + \\ \gamma_{60}(\text{Hisp})_{ij} + \gamma_{70}(\text{Other})_{ij} + \gamma_{80}(\text{EDS})_{ij} + \gamma_{90}(\text{LEP})_{ij} + \gamma_{100}(\text{SWD})_{ij}] + \\ (\textit{fixed effects}) \\ \\ [u_{0j} + u_{1j}(\text{SCAs})_{ij} + u_{1j}(\text{LCAs})_{ij} + u_{3j}(\text{EDS})_{ij} + u_{4j}(\text{LEP})_{ij} + u_{5j}(\text{SWD})_{ij} + r_{ij}] \\ (\textit{random effects})\end{array}$$

Where:

$\gamma_{10}(\text{SCAs})_{ij}$	slope for total number of short-cycle assessments predictor
$\gamma_{20}(\text{LCAs})_{ij}$	slope for total number of long-cycle assessments predictor
$u_{1j}(\text{SCAs})_{ij}$	variability in slope for short-cycle assessments predictor
$u_{2j}(\text{LCAs})_{ij}$	variability in slope for long-cycle assessments predictor

Note: The terms Yij, u_{0j}, and r_{ij} were defined previously in this chapter; all other terms are defined in Table 3.3;

Once both models have been fit, deviance statistics were calculated for each model in order to determine which model (total assessment count model vs. assessment count by cycle-length model) provided a better fit for the data. Once this decision was made, and in order to address Research Question 3, the better fit model was tested for interaction effects between the at-risk student variables and the variable(s) chosen to represent formative assessment frequency (either total assessment count or SCAs and LCAs). This step served to determine if the effect of formative assessment frequency on student achievement varies for different at-risk student sub-groups (EDS, LEP, SWD). These interaction terms were initially allowed to randomly vary across schools. Any variance components determined to be statistically non-significantly different from 0 were then fixed. Any statistically significant interactions were retained in the model as long as the addition resulted in an improved model fit. The equation for the total number of assessments (regardless of cycle-length) with interactions is presented in Table 3.5. The equation for the cycle-length specific model with interactions is presented in Table 3.6.

Table 3.6 *Random Coefficient – Total Assessments Model with Interactions*

$$Yij = [\gamma_{00} + \gamma_{10}(\text{TotalAssmts})_{ij} + \gamma_{20}(\text{Gender})_{ij} + \gamma_{30}(\text{Asian})_{ij} + \gamma_{40}(\text{AfrAm})_{ij} + \gamma_{50}(\text{Hisp})_{ij} +$$

$$\gamma_{60}(\text{Other})_{ij} + \gamma_{70}(\text{EDS})_{ij} + \gamma_{80}(\text{LEP})_{ij} + \gamma_{90}(\text{SWD})_{ij} + \gamma_{100}(\text{TotalAssmts})_{ij}{*}(\text{EDS})_{ij} +$$

$$\gamma_{110}(\text{TotalAssmts})_{ij}{*}(\text{LEP})_{ij} + \gamma_{120}(\text{TotalAssmts})_{ij}{*}(\text{SWD})_{ij}] +$$

(*fixed effects*)

$$[u_{0j} + u_{1j}(\text{TotalAssmts})_{ij} + u_{2j}(\text{EDS})_{ij} + u_{3j}(\text{LEP})_{ij} + u_{4j}(\text{SWD})_{ij} +$$

$$u_{5j}(\text{TotalAssmts})_{ij}{*}(\text{EDS})_{ij} + u_{6j}(\text{TotalAssmts})_{ij}{*}(\text{LEP})_{ij} +$$

$$u_{7j}(\text{TotalAssmts})_{ij}{*}(\text{SWD})_{ij} + r_{ij}]$$

(*random effects*)

Where:

$\gamma_{100}(\text{TotalAssmts})_{ij}{*}(\text{EDS})_{ij}$	slope for interaction between total number of formative assessments predictor and EDS
$\gamma_{110}(\text{TotalAssmts})_{ij}{*}(\text{LEP})_{ij}$	slope for interaction between total number of formative assessments predictor and LEP
$\gamma_{120}(\text{TotalAssmts})_{ij}{*}(\text{SWD})_{ij}$	slope for interaction between total number of formative assessments predictor and SWD
$u_{5j}(\text{TotalAssmts})_{ij}{*}(\text{EDS})_{ij}$	variability in slope for interaction between total number of formative assessments predictor and EDS
$u_{6j}(\text{TotalAssmts})_{ij}{*}(\text{LEP})_{ij}$	variability in slope for interaction between total number of formative assessments predictor and LEP
$u_{7j}(\text{TotalAssmts})_{ij}{*}(\text{SWD})_{ij}$	variability in slope for interaction between total number of formative assessments predictor and SWD

Note: The terms Yij, u_{0j}, and r_{ij} were defined previously in this chapter; all other terms not defined in this table are defined in Table 3.3;

Table 3.7 *Random Coefficient – Cycle-Length Specific Model with Interactions*

Yij =

$$[\gamma_{00} + \gamma_{10}(\text{SCAs})_{ij} + \gamma_{20}(\text{LCAs})_{ij} + \gamma_{30}(\text{Gender})_{ij} + \gamma_{40}(\text{Asian})_{ij} + \gamma_{50}(\text{AfrAm})_{ij} + \gamma_{60}(\text{Hisp})_{ij} + \gamma_{70}(\text{Other})_{ij} + \gamma_{80}(\text{EDS})_{ij} + \gamma_{90}(\text{LEP})_{ij} + \gamma_{100}(\text{SWD})_{ij} + \gamma_{110}(\text{SCAs})_{ij}{*}(\text{EDS})_{ij} + \gamma_{120}(\text{SCAs})_{ij}{*}(\text{LEP})_{ij} + \gamma_{130}(\text{SCAs})_{ij}{*}(\text{SWD})_{ij} + \gamma_{140}(\text{LCAs})_{ij}{*}(\text{EDS})_{ij} + \gamma_{150}(\text{LCAs})_{ij}{*}(\text{LEP})_{ij} + \gamma_{160}(\text{LCAs})_{ij}{*}(\text{SWD})_{ij}] +$$

(*fixed effects*)

$$[u_{0j} + u_{1j}(\text{SCAs})_{ij} + u_{1j}(\text{LCAs})_{ij} + u_{3j}(\text{EDS})_{ij} + u_{4j}(\text{LEP})_{ij} + u_{5j}(\text{SWD})_{ij} + u_{6j}(\text{SCAs})_{ij}{*}(\text{EDS})_{ij} + u_{7j}(\text{SCAs})_{ij}{*}(\text{LEP})_{ij} + u_{8j}(\text{SCAs})_{ij}{*}(\text{SWD})_{ij} + u_{9j}(\text{LCAs})_{ij}{*}(\text{LEP})_{ij} + u_{10j}(\text{LCAs})_{ij}{*}(\text{SWD})_{ij} + u_{11j}(\text{LCAs})_{ij}{*}(\text{LEP})_{ij} + r_{ij}]$$

(*random effects*)

Where:

$\gamma_{110}(\text{SCAs})_{ij}{*}(\text{EDS})_{ij}$ slope for interaction between total number of short-cycle assessments predictor and EDS

$\gamma_{120}(\text{SCAs})_{ij}{*}(\text{LEP})_{ij}$ slope for interaction between total number of short-cycle assessments predictor and LEP

$\gamma_{130}(\text{SCAs})_{ij}{*}(\text{SWD})_{ij}$ slope for interaction between total number of short-cycle assessments predictor and SWD

$\gamma_{140}(\text{LCAs})_{ij}{*}(\text{EDS})_{ij}$ slope for interaction between total number of long-cycle assessments predictor and EDS

$\gamma_{150}(\text{LCAs})_{ij}{*}(\text{LEP})_{ij}$ slope for interaction between total number of long-cycle assessments predictor and LEP

$\gamma_{160}(\text{LCAs})_{ij}{*}(\text{SWD})_{ij}$ slope for interaction between total number of long-cycle assessments predictor and SWD

$u_{6j}(\text{SCAs})_{ij}{*}(\text{EDS})_{ij}$ variability in slope for interaction between total number of short-cycle assessments predictor and EDS

$u_{7j}(\text{SCAs})_{ij}{*}(\text{LEP})_{ij}$ variability in slope for interaction between total number of short-cycle assessments predictor and LEP

$u_{8j}(\text{SCAs})_{ij}{*}(\text{SWD})_{ij}$ variability in slope for interaction between total number of short-cycle assessments predictor and SWD

$u_{9j}(\text{LCAs})_{ij}{*}(\text{EDS})_{ij}$ variability in slope for interaction between total number of long-cycle assessments predictor and EDS

$u_{10j}(\text{LCAs})_{ij}{*}(\text{LEP})_{ij}$ variability in slope for interaction between total number of long-cycle assessments predictor and LEP

$u_{11j}(\text{LCAs})_{ij}{*}(\text{SWD})_{ij}$ variability in slope for interaction between total number of long-cycle assessments predictor and SWD

Note: The terms Yij, u_{0j}, and r_{ij} were defined previously in this chapter; all other terms are defined in Table 3.3;

Full Contextual Model

The final step in building the multi-level model for this study was adding relevant school-level variables to the model in order to address Research Questions 4 and 5. In order to address Research Question 4, variables representing percentage of EDS students (%EDS) as well as percentage of minority students (%Minority) for the given school were added to the model. In order to address Research Question 5, variables representing assessment frequency average for the given school were added to the model. If the total assessments count model was determined to be the best fit model, the school total number of assessments mean was used. However, if the assessment count by cycle-length model was determined to have the best fit, two school-level means were added – one for mean number of SCAs and one for mean number of LCAs. In the full contextual model formulas illustrated in Tables 3.7 (total assessment count model) and 3.8 (assessment count by cycle-length model) the level-2 variables predicted variance in the intercept. As was done with the random coefficients model, any statistically non-significant variables were eliminated in an effort to retain the most parsimonious model possible.

Table 3.8 *Full Contextual – Mixed-Effects Model (Total Assessments)*

$$
\begin{aligned}
Yij = \; & [\gamma_{00} + \gamma_{01}(\text{%EDS})_j + \gamma_{02}(\text{%Minority})_j + \gamma_{03}(\text{MeanAssmts})_j + \gamma_{10}(\text{TotalAssmts})_{ij} + \\
& \gamma_{20}(\text{Gender})_{ij} + \gamma_{30}(\text{Asian})_{ij} + \gamma_{40}(\text{AfrAm})_{ij} + \gamma_{50}(\text{Hisp})_{ij} + \gamma_{60}(\text{Other})_{ij} + \gamma_{70}(\text{EDS})_{ij} + \\
& \gamma_{80}(\text{LEP})_{ij} + \gamma_{90}(\text{SWD})_{ij} + \gamma_{100}(\text{TotalAssmts})_{ij}*(\text{EDS})_{ij} + \gamma_{110}(\text{TotalAssmts})_{ij}*(\text{LEP})_{ij} + \\
& \gamma_{120}(\text{TotalAssmts})_{ij}*(\text{SWD})_{ij}] + \\
& (\textit{fixed effects}) \\
& [u_{0j} + u_{1j}(\text{TotalAssmts})_{ij} + u_{2j}(\text{EDS})_{ij} + u_{3j}(\text{LEP})_{ij} + u_{4j}(\text{SWD})_{ij} + \\
& u_{5j}(\text{TotalAssmts})_{ij}*(\text{EDS})_{ij} + u_{6j}(\text{TotalAssmts})_{ij}*(\text{LEP})_{ij} + \\
& u_{7j}(\text{TotalAssmts})_{ij}*(\text{SWD})_{ij} + r_{ij}] \\
& (\textit{random effects})
\end{aligned}
$$

Where:

$\gamma_{01}(\text{%EDS})_j$	slope for percentage of EDS students at school j
$\gamma_{02}(\text{%Minority})_j$	slope for percentage of minority students at school j
γ_{03} $(\text{MeanAssmts})_j$	slope for mean number of formative assessments at school j

Note: The terms Yij, u_{0j}, and r_{ij} were defined previously in this chapter; all other terms are defined in Table 3.3;

Table 3.9 *Full Contextual – Mixed-Effects Model (Cycle-Length Specific)*

$$
\begin{aligned}
Yij = \; & [\gamma_{00} + \gamma_{01}(\%\mathrm{EDS})_j + \gamma_{02}(\%\mathrm{Minority})_j + \gamma_{03}(\mathrm{MeanSCAs})_j + \gamma_{04}(\mathrm{MeanLCAs})_j + \\
& \gamma_{10}(\mathrm{SCA})_{ij} + \gamma_{20}(\mathrm{LCA})_{ij} + \gamma_{30}(\mathrm{Gender})_{ij} + \gamma_{40}(\mathrm{Asian})_{ij} + \gamma_{50}(\mathrm{AfrAm})_{ij} + \gamma_{60}(\mathrm{Hisp})_{ij} + \\
& \gamma_{70}(\mathrm{Other})_{ij} + \gamma_{80}(\mathrm{EDS})_{ij} + \gamma_{90}(\mathrm{LEP})_{ij} + \gamma_{100}(\mathrm{SWD})_{ij} + \gamma_{110}(\mathrm{SCAs})_{ij}{*}(\mathrm{EDS})_{ij} + \\
& \gamma_{120}(\mathrm{SCAs})_{ij}{*}(\mathrm{LEP})_{ij} + \gamma_{130}(\mathrm{SCAs})_{ij}{*}(\mathrm{SWD})_{ij} + \gamma_{140}(\mathrm{LCAs})_{ij}{*}(\mathrm{EDS})_{ij} + \\
& \gamma_{150}(\mathrm{LCAs})_{ij}{*}(\mathrm{LEP})_{ij} + \gamma_{160}(\mathrm{LCAs})_{ij}{*}(\mathrm{SWD})_{ij}] + \\
& (\textit{fixed effects}) \\
& [u_{0j} + u_{1j}(\mathrm{SCAs})_{ij} + u_{1j}(\mathrm{LCAs})_{ij} + u_{3j}(\mathrm{EDS})_{ij} + u_{4j}(\mathrm{LEP})_{ij} + u_{5j}(\mathrm{SWD})_{ij} + \\
& u_{6j}(\mathrm{SCAs})_{ij}{*}(\mathrm{EDS})_{ij} + u_{7j}(\mathrm{SCAs})_{ij}{*}(\mathrm{LEP})_{ij} + u_{8j}(\mathrm{SCAs})_{ij}{*}(\mathrm{SWD})_{ij} + \\
& u_{9j}(\mathrm{LCAs})_{ij}{*}(\mathrm{LEP})_{ij} + u_{10j}(\mathrm{LCAs})_{ij}{*}(\mathrm{SWD})_{ij} + u_{11j}(\mathrm{LCAs})_{ij}{*}(\mathrm{LEP})_{ij} + r_{ij}] \\
& (\textit{random effects})
\end{aligned}
$$

Where:

$\gamma_{01}(\%\mathrm{EDS})_j$	slope for percentage of EDS students at school j
$\gamma_{02}(\%\mathrm{Minority})_j$	slope for percentage of minority students at school j
$\gamma_{03}(\mathrm{MeanSCAs})_j$	slope for mean number of short-cycle assessments at school j
$\gamma_{04}(\mathrm{MeanLCAs})_j$	slope for mean number of long-cycle assessments at school j

Note: The terms Yij, u_{0j}, and r_{ij} were defined previously in this chapter; all other terms are defined in Table 3.3;

Once the model was determined to be satisfactory, in order to address Research Questions 4 and 5, interactions between relevant school-level variables and assessment and the variable(s) chosen to represent formative assessment frequency (either total assessment count or SCAs and LCAs) were tested. In the full contextual model formulas illustrated in Tables 3.9 (total assessment count model) and 3.10 (assessment count by cycle-length model) the level-2 variable interaction terms predicted variance in the level-1 slopes. Any statistically significant interactions were retained for the final model. The full contextual model equation for the total number of assessments (regardless of cycle-length) with interactions is presented in Table 3.9. The full contextual model equation for the cycle-length specific model with interactions is presented in Table 3.10.

Table 3.10 *Full Contextual – Total Assessments Model with Interactions*

Yij =

$$[\gamma_{00} + \gamma_{01}(\%EDS)_j + \gamma_{02}(\%Minority)_j + \gamma_{03}(MeanAssmts)_j + \gamma_{10}(TotalAssmts)_{ij} + \gamma_{20}(Gender)_{ij} + \gamma_{30}(Asian)_{ij} + \gamma_{40}(AfrAm)_{ij} + \gamma_{50}(Hisp)_{ij} + \gamma_{60}(Other)_{ij} + \gamma_{70}(EDS)_{ij} + \gamma_{80}(LEP)_{ij} + \gamma_{90}(SWD)_{ij} + \gamma_{100}(TotalAssmts)_{ij}{*}(EDS)_{ij} + \gamma_{110}(TotalAssmts)_{ij}{*}(LEP)_{ij} + \gamma_{120}(TotalAssmts)_{ij}{*}(SWD)_{ij} + \gamma_{11}(TotalAssmts)_{ij}{*}(\%EDS)_j + \gamma_{12}(TotalAssmts)_{ij}{*}(\%Minority)_j + \gamma_{13}(TotalAssmts)_{ij}{*}(MeanAssmts)_j] +$$

(*fixed effects*)

$$[u_{0j} + u_{1j}(TotalAssmts)_{ij} + u_{2j}(EDS)_{ij} + u_{3j}(LEP)_{ij} + u_{4j}(SWD)_{ij} + u_{5j}(TotalAssmts)_{ij}{*}(EDS)_{ij} + u_{6j}(TotalAssmts)_{ij}{*}(LEP)_{ij} + u_{7j}(TotalAssmts)_{ij}{*}(SWD)_{ij} + r_{ij}]$$

(*random effects*)

Where:

$\gamma_{11}(TotalAssmts)_{ij}{*}(\%EDS)_j$	slope for interaction between total number of formative assessments predictor and school percent EDS students
$\gamma_{12}(TotalAssmts)_{ij}{*}(\%Minority)_j$	slope for interaction between total number of formative assessments predictor and school percent minority students
$\gamma_{13}(TotalAssmts)_{ij}{*}(MeanAssmts)_j$	slope for interaction between total number of formative assessments predictor and school mean number of formative assessments

Note: The terms Yij, u_{0j}, and r_{ij} were defined previously in this chapter; all other terms are defined in Tables 3.3 and 3.5;

Table 3.11 *Full Contextual – Cycle-Length Specific Model with Interactions*

$$
\begin{aligned}
Yij = \; & [[\gamma_{00} + \gamma_{01}(\%\text{EDS})_j + \gamma_{02}(\%\text{Minority})_j + \gamma_{03}(\text{MeanSCAs})_j + \gamma_{04}(\text{MeanLCAs})_j + \\
& \gamma_{10}(\text{SCA})_{ij} + \gamma_{20}(\text{LCA})_{ij} + \gamma_{30}(\text{Gender})_{ij} + \gamma_{40}(\text{Asian})_{ij} + \gamma_{50}(\text{AfrAm})_{ij} + \gamma_{60}(\text{Hisp})_{ij} + \\
& \gamma_{70}(\text{Other})_{ij} + \gamma_{80}(\text{EDS})_{ij} + \gamma_{90}(\text{LEP})_{ij} + \gamma_{100}(\text{SWD})_{ij} + \gamma_{110}(\text{SCAs})_{ij}*(\text{EDS})_{ij} + \\
& \gamma_{120}(\text{SCAs})_{ij}*(\text{LEP})_{ij} + \gamma_{130}(\text{SCAs})_{ij}*(\text{SWD})_{ij} + \gamma_{140}(\text{LCAs})_{ij}*(\text{EDS})_{ij} + \\
& \gamma_{150}(\text{LCAs})_{ij}*(\text{LEP})_{ij} + \gamma_{160}(\text{LCAs})_{ij}*(\text{SWD})_{ij} + \gamma_{11}(\text{SCAs})_{ij}*(\%\text{EDS})_j + \\
& \gamma_{12}(\text{SCAs})_{ij}*(\%\text{Minority})_j + \gamma_{13}(\text{SCAs})_{ij}*(\text{MeanSCAs})_j + \gamma_{14}(\text{SCAs})_{ij}*(\text{MeanLCAs})_j + \\
& \gamma_{21}(\text{LCAs})_{ij}*(\%\text{EDS})_j + \gamma_{22}(\text{LCAs})_{ij}*(\%\text{Minority})_j + \gamma_{23}(\text{LCAs})_{ij}*(\text{MeanSCAs})_j + \\
& \gamma_{24}(\text{LCAs})_{ij}*(\text{MeanLCAs})_j] + \\
& (\textit{fixed effects}) \\
& [u_{0j} + u_{1j}(\text{SCAs})_{ij} + u_{1j}(\text{LCAs})_{ij} + u_{3j}(\text{EDS})_{ij} + u_{4j}(\text{LEP})_{ij} + u_{5j}(\text{SWD})_{ij} + \\
& u_{6j}(\text{SCAs})_{ij}*(\text{EDS})_{ij} + u_{7j}(\text{SCAs})_{ij}*(\text{LEP})_{ij} + u_{8j}(\text{SCAs})_{ij}*(\text{SWD})_{ij} + \\
& u_{9j}(\text{LCAs})_{ij}*(\text{LEP})_{ij} + u_{10j}(\text{LCAs})_{ij}*(\text{SWD})_{ij} + u_{11j}(\text{LCAs})_{ij}*(\text{LEP})_{ij} + r_{ij}] \\
& (\textit{random effects})
\end{aligned}
$$

Where:

$\gamma_{11}(\text{SCAs})_{ij}*(\%\text{EDS})_j$ slope for interaction between total number of short-cycle assessments predictor and school percent EDS students

$\gamma_{12}(\text{SCAs})_{ij}*(\%\text{Minority})_j$ slope for interaction between total number of short-cycle assessments predictor and school percent minority students

$\gamma_{13}(\text{SCAs})_{ij}*(\text{MeanSCAs})_j$ slope for interaction between total number of short-cycle assessments predictor and school mean number of short-cycle formative assessments

$\gamma_{14}(\text{SCAs})_{ij}*(\text{MeanLCAs})_j$ slope for interaction between total number of formative assessments predictor and school mean number of long-cycle formative assessments

$\gamma_{21}(\text{LCAs})_{ij}*(\%\text{EDS})_j$ slope for interaction between total number of long-cycle assessments predictor and school percent EDS students

$\gamma_{22}(\text{LCAs})_{ij}*(\%\text{Minority})_j$ slope for interaction between total number of long-cycle assessments predictor and school percent minority students

$\gamma_{23}(\text{LCAs})_{ij}*(\text{MeanSCAs})_j$ slope for interaction between total number of long-cycle assessments predictor and school mean number of short-cycle formative assessments

$\gamma_{24}(\text{LCAs})_{ij}*(\text{MeanLCAs})_j$ slope for interaction between total number of long-cycle assessments predictor and school mean number of long-cycle formative assessments

Note: The terms Yij, u_{0j}, and r_{ij} were defined previously in this chapter; all other terms are defined in Tables 3.3 and 3.6;

Model Comparison

All models were tested for goodness of fit using the deviance statistic (-2LL) as well as the Akaike Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC). Deviance is calculated based on the number of parameters being estimated and is therefore more sensitive when comparing models differing in number of parameters (Luke, 2005). On the other hand, the AIC and BIC penalize models with more parameters and, therefore, are less sensitive when comparing models with differing number of parameters (Luke, 2005). Although none of these statistics can be interpreted directly, they can be used to compare multiple models to one another.

In addition, the final models were tested for predictive ability by estimating the proportional reduction in prediction error at level-1 and level-2. The unconditional model was considered the baseline model and the best fit model between the random coefficients model and the full contextual model served as the fitted model for this comparison. This statistic was calculated at both levels for each subject. The equations for level-1 and level-2 proportional reduction in prediction error are given in Equations 3 and 4, respectively.

$$R_1^2 = 1 - \frac{(\hat{\sigma}^2 + \hat{\tau}_{00})_f}{(\hat{\sigma}^2 + \hat{\tau}_{00})_b} \qquad (3)$$

where: R_1^2 = proportional reduction in prediction error for level-1

$\hat{\sigma}^2$ = estimated level-1 variance

$\hat{\tau}_{00}$ = estimated level-2 variance

and where,

$(\hat{\sigma}^2 + \hat{\tau}_{00})_f$ = unexplained variance in the final model

$(\hat{\sigma}^2 + \hat{\tau}_{00})_f$ = unexplained variance in the baseline model

$$R_2^2 = 1 - \frac{(\frac{\hat{\sigma}^2}{n_j} + \hat{\tau}_{00})_f}{(\frac{\hat{\sigma}^2}{n_j} + \hat{\tau}_{00})_b} \qquad (4)$$

where: R_2^2 = proportional reduction in prediction error for level-2

n_j = number of students in school j

and where,

$(\frac{\hat{\sigma}^2}{n_j} + \hat{\tau}_{00})_f$ = prediction error for the final fitted model

$(\frac{\hat{\sigma}^2}{n_j} + \hat{\tau}_{00})_b$ = prediction error for the baseline model

CHAPTER 4: RESULTS

The main goal of this study was to investigate the effects of formative assessment on middle school student achievement on state-mandated standardized tests. Multiple level 1 predictors as well as three level 2 predictors were used in a multilevel model to investigate this relationship. Results of the analyses are presented in the following sections of this chapter.

Software and Parameter Estimation

Software. The analyses for this study were done using *SAS version 9.2*. The SAS Proc Mixed procedure was used for fitting the multi-level models (MLM) for this study. SAS was chosen for this study due to its ability to sufficiently handle two-level data sets with normally distributed response variables. Singer's (1998) article was helpful in specifying the appropriate SAS code needed to answer the specific research questions for this study.

Parameter estimation. The two parameter estimation methods most commonly used for MLMs with normal response variables are the *maximum likelihood* (ML) and the *restricted maximum likelihood* (REML) (McCoach, 2010). Although REML is the default method of estimation for the SAS Proc Mixed procedure, an option was included directing SAS to use the ML estimation technique. Given the large number of clusters included in this study (*N*=413) ML and REML would very likely produce similar estimates of variance components and fixed effects. However, ML is preferable over REML for testing model fit when comparing models with different fixed and/or random effects (McCoach, 2010). The framework of this study, in which models fit with different fixed and random effects were compared (Research Question 1 estimated the effects of frequency of assessment and student performance and Research Question 2 estimated the effect of frequency of assessment based on cycle length), dictated that ML was the most appropriate estimation technique to employ.

Error Covariance Structure

The models fit in this study were assumed to have the error covariance structure referred to as *compound symmetry* (Singer, 1998). This structure assumes that: 1) the total residual variance for each student in the model is the sum of the within school residual (σ^2) and the between-school residual (τ_{00}); 2) the covariance between any two students in the same school is τ_{00}; and 3) the residual covariance between students in different schools is equal to zero (McCoach, 2010).

Descriptive Statistics

Student-level (level-1) continuous variables. Basic descriptive statistics for the continuous level-1 variables in this study are provided in Table 4.1.

AC-Score. The AC-Scores for reading and math are comparable, with the mean reading AC-Score approximately 0.02 lower than the mean math AC-Score. The difference in standard deviation between the subjects was also negligible (SDx = 0.439 for math, 0.424 for reading). The minimum AC-Score point for math was higher in comparison to reading, but so was the corresponding maximum AC-Score point.

Cycle length. As one might expect, the maximum number of short-cycle assessments is greater than the maximum number of long-cycle assessments for both reading and math.

However, the maximum number of short-cycle assessments for math (99) is much greater than for reading (24). Upon investigating the data, it was found that there were multiple schools that utilized the short-cycle assessments heavily for mathematics resulting in a greater mean, standard deviation, and maximum frequency for mathematics short-cycle assessments. However, given the sample size in terms of schools (N = 413) and students (n = 83,799) in addition to the relatively small effect on the overall mean short-cycle assessment count for math (2.361 for math as compared to 1.317 for reading), it was determined to be reasonable to retain these schools and students for the analysis. The statistics for long-cycle assessments were comparable between subject areas.

In Table 4.1, Total Count represents the number of formative assessments per student by subject area regardless of cycle length; thus, the mean total assessment count for each subject is the sum of the mean numbers of long and short cycle assessment means.

Table 4.1 *Descriptive Statistics for Continuous Level-1 variables*

	Mathematics				Reading			
	Mean	Std Dev	Min	Max	Mean	Std Dev	Min	Max
AC-Score	0.109	0.439	-2.22	2.59	0.079	0.424	-2.695	2.283
Total Count	3.72	4.591	0	103	2.711	2.249	0	24
Long-Cycle Count	1.359	1.307	0	11	1.394	1.206	0	7
Short-Cycle Count	2.361	4.418	0	99	1.317	2.154	0	24

Student-level (level-1) dichotomous variables. Frequencies and proportions for the student-level demographic variables are provided in Table 4.2. These variables were used to control for at-risk status.

Table 4.2 *Demographics of Participants*

		Number	Percentage
Gender			
	Male	41,826	50.0%
	Female	41,973	50.0%
Race			
	Asian	970	1.2%
	African American	18,686	22.3%
	Hispanic	8,930	10.7 %
	Other	2,781	3.3%
	White	51,842	61.9%
	EDS	44,828	53.5%
	LEP	3,248	3.9%
	SWD	5,578	6.7%

Note: *Proportions may not lead to 100% due to rounding;*

Gender. The gender dichotomy represented in Table 4.2 is what one would expect for a large sample such as the one used for this study. There were slightly more females than males, but the difference is negligible resulting a nearly 50%-50% split.

Race. The racial distribution for the sample was also similar to would be expected for representative sample from North Carolina schools. For the 2010-11 school year, NCDPI reported proportions for ethnicities very similar to those illustrated in Table 4.2 (see NCDPI, 2010). The sample consisted largely of white students (61.9%) following by African

American students (22.3%), Hispanic students (10.7%), Other students (3.3%), and Asian students (1.2%).

Disadvantaged students. Similar to gender and race, the disadvantaged student categories also displayed distributions similar to what was expected. As is typical in North Carolina, slightly more than half (53.5%) of students fell into the economically disadvantaged category (EDS). Limited English proficiency (LEP) students and students with disabilities (SWD) each accounted for small proportions of the overall sample with 3.9% and 6.7% respectively.

School-level (level-2) assessment variables. Basic descriptive statistics for the school-level assessment variables are provided in Table 4.3.

Table 4.3 *Descriptive Statistics for Continuous Level-2 variables*

	Mathematics		Reading	
	Mean	Std Dev	Mean	Std Dev
Total Count	3.39	3.12	2.65	1.65
Long-Cycle Count	1.07	1.25	1.04	1.08
Short-Cycle Count	1.98	2.66	1.17	1.42

Cycle length. Table 4.3 illustrates that the school-level means for frequency of formative assessment are very similar to the level-1 means. Interestingly, the level-2 short-cycle assessment count is higher than the long-cycle assessment count (1.17 as compared to 1.04) for reading whereas the opposite was seen for the same means at level-1 for reading.

School-level (level-2) demographic variables. Basic descriptive statistics for the demographic school-level variables of interest are presented in Table 4.4.

Table 4.4 *Demographics of Schools*

	Mean	Standard Deviation
School N	202.9	260.0
Percent EDS	66.0%	28.0
Percent Minority	48.0%	35.0

Demographics. A statistic of particular importance presented in Table 4.4 is the average number of students per school in the sample (202.9). This statistic was necessary for computing the design effect. These statistics are presented in a subsequent section of this chapter. The mean percent EDS and percent minority students per school was similar what was expected based on the level-1 variables measuring these same at-risk student characteristics. The addition of these level-2 variables in the model building may wash out any effects of the level-1 at-risk predictors.

Correlation matrices. Four correlation matrices are presented in this section. Two tables are presented for level-1 variables for each subject area (Tables 4.5 and 4.6) as well as another two tables for each subject's level-2 variables (Tables 4.7 and 4.8).

Table 4.5 *Bivariate Correlations for Individual Level Variables – Mathematics*

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Math AC-Score	1.00	0.04***	-0.01***	0.05***	0.06***	0.03***	-0.01	-0.01*	0.00	0.00	-0.06***	0.00	-0.03***
2. Total Count		1.00	0.27***	0.96***	0.00	-0.01**	0.04***	0.03***	-0.02***	-0.04***	0.05***	0.02***	-0.02***
3. Short-Cycle Count			1.00	-0.01**	0.00	0.00	0.05***	0.01	0.01	-0.06***	0.06***	0.02***	-0.01**
4. Long-Cycle Count				1.00	0.00	-0.01**	0.02***	0.03***	-0.02***	-0.03***	0.03***	0.01**	-0.02***
5. Gender					1.00	0.01	0.01	0.00	0.00	-0.01**	0.01**	-0.02***	-0.09***
6. Asian						1.00	-0.06***	-0.04***	-0.02***	-0.14***	-0.01**	0.09***	-0.02***
7. Afr. Amer.							1.00	-0.19***	-0.1***	-0.68***	0.29***	-0.1***	0.04***
8. Hispanic								1.00	-0.06***	-0.44***	0.24***	0.51***	-0.02***
9. Other									1.00	-0.24***	0.04***	-0.02***	0.00
10. White										1.00	-0.42***	-0.24***	-0.02***
11. EDS											1.00	0.16***	0.06***
12. LEP												1.00	0.02***
13. SWD													1.00

*Note: *** p<.0001; ** p <.01; *p<.05*

Table 4.6 *Bivariate Correlations for Individual Level Variables – Reading*

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Reading AC-Score	1.00	0.00	0.01**	-0.01***	0.00	0.02***	-0.05***	0.01**	0.01*	0.03***	-0.04***	0.00	-0.02***
2. Total Count		1.00	0.85***	0.35***	0.00	-0.01**	0.04***	0.02***	-0.01**	-0.04***	0.05***	0.01**	-0.04***
3. Short-Cycle Count			1.00	-0.2***	0.00	-0.01*	0.02***	0.01**	-0.02***	-0.02***	0.02***	0.00	-0.03***
4. Long-Cycle Count				1.00	0.00	0.00	0.04***	0.01	0.01*	-0.05***	0.05***	0.02***	-0.02***
5. Gender					1.00	0.01	0.01	0.00	0.00	-0.01**	0.01**	-0.02***	-0.09***
6. Asian						1.00	-0.06***	-0.04***	-0.02***	-0.14***	-0.01**	0.09***	-0.02***
7. Afr. Amer.							1.00	-0.19***	-0.1***	-0.68***	0.29***	-0.1***	0.04***
8. Hispanic								1.00	-0.06***	-0.44***	0.24***	0.51***	-0.02***
9. Other									1.00	-0.24***	0.04***	-0.02***	0.00
10. White										1.00	-0.42***	-0.24***	-0.02***
11. EDS											1.00	0.16***	0.06***
12. LEP												1.00	0.02***
13. SWD													1.00

*Note: *** p<.0001; ** p <.01; *p<.05*

Table 4.7 *Bivariate Correlations for School Level (Level-2) Variables - Mathematics*

Variables	1	2	3	4	5
1. Mathematics Count Mean	1.00	0.93***	0.28***	0.16***	0.1***
2. Short-Cycle Mean		1.00	-0.06***	0.12***	0.08***
3. Long-Cycle Mean			1.00	0.2***	0.11***
4. Percent EDS				1.00	0.72***
5. Percent Minority					1.00

*Note: *** p<.0001; ** p <.01; *p<.05*

Table 4.8 *Bivariate Correlations for School Level (Level-2) Variables - Reading*

Variables	1	2	3	4	5
1. Reading Count Mean	1.00	0.28***	0.76***	0.23***	0.12***
2. Short-Cycle Mean		1.00	-0.32***	0.18***	0.11***
3. Long-Cycle Mean			1.00	0.13***	0.05***
4. Percent EDS				1.00	0.72***
5. Percent Minority					1.00

*Note: *** p<.0001; ** p <.01; *p<.05*

Unconditional Means Model

Unconditional model results for mathematics. The results from the unconditional model for mathematics are presented in Table 4.9. Covariance and fixed effects are discussed in the following sections.

Table 4.9 *Unconditional Models for Mathematics*

Fixed Effect	*Coefficient*	*Standard Error*	*DF*	*T-Value*	*P-value*
Math AC-Score Intercept (γ_{00})	0.0956	0.0078	412	12.32	<.0001
Random Effects	*Estimate*	*Standard Error*	*DF*	*Z-Value*	*P-value*
Intercept (u_{0j})	0.0154	0.0016	412	9.37	<.0001
Residual (r_{ij})	0.1829	0.0009	412	204.24	<.0001

Covariance parameter estimates. The estimated school-level variation (τ_{00}) in mathematics AC-Score was estimated to be 0.0154. The student-level variance (σ^2) was estimated to be 0.1829. The hypothesis tests indicate that these estimates were determined to be statistically significantly different from 0 (p <.0001). This suggests that schools differ in average mathematics AC-Score and that there is even more variation among students within schools ($\sigma^2$is nearly 12 times larger than τ_{00}).

Fixed effects parameter estimates. The estimated fixed effect for Math AC-Score was 0.0956. This represents the average mathematics AC-Score across schools. This estimate is slightly lower than the average mathematics AC-Score (0.109).

Unconditional model results for reading. The results from the unconditional model for Reading are presented in Table 4.10. Covariance and fixed effects estimates are discussed in the following sections.

Table 4.10 *Unconditional Models for Reading*

Fixed Effect	*Coefficient*	*Standard Error*	*DF*	*T-Value*	*P-value*
Reading AC-Score Intercept (γ_{00})	0.0739	0.005	412	16.47	<.0001
Random Effects	*Variance Component*	*Standard Error*	*DF*	*Z-Value*	*P-value*
Intercept (u_{0j})	0.0041	0.0006	412	6.99	<.0001
Residual (r_{ij})	0.1769	0.0009	412	204.18	<.0001

Covariance parameter estimates. The estimated school-level variation (τ_{00}) in reading AC-Score was 0.0041. The estimated student-level variation in reading AC-Score (σ^2) was estimated to be 0.1769. The hypothesis tests indicate that these estimates were determined to be statistically significantly different from 0. As was found in the mathematics analyses, the results from the reading analysis suggest that there is variation in reading AC-Score among schools and that there is even greater variation between students within schools ($\sigma^2$is more than 43 times τ_{00}).

Fixed effects parameter estimates. The estimated fixed effect for reading AC-Score was 0.0739. This represents the average reading AC-Score across schools. This estimate is slightly lower than the average student reading AC-Score (0.079) which was indicated in Table 4.1.

Intraclass Correlation Coefficient & Design Effect

The statistics necessary to calculate the ICC and design effect for each subject were obtained by fitting an unconditional model for each subject-specific AC-score. A detailed description and equation for the unconditional model was described in the previous chapter. The results from these analyses are illustrated in Tables 4.9 and 4.10.

ICC and design effect for mathematics. Based on the results from the unconditional model for mathematics, the ICC was calculated to be 0.078. The interpretation of this statistic is that schools account for approximately 7.8% of the variability in mathematics AC-Score between students. According to McCoach (2010, p. 134), in school effects research, “ICCs typically range from .10 to .20”. Although the ICC for mathematics was at the lower end of what may be considered the typical range, it still provided sufficient evidence suggesting that a multilevel model may be beneficial. Based on the calculated ICC of .078 and the average school size ($\overline{n}_j$) of 202.9 students in our sample, the design effect for mathematics was determined to be 4.1. This indicated that the standard errors would be inflated by a factor of 4.1 if a multilevel approach was not used for this study and independence of observations was assumed. For these reasons it as determined that a multi-level approach would be beneficial for the mathematics analysis in this study.

ICC and design effect for reading. Based on the results from the unconditional model for reading, the ICC was calculated to be 0.023. The interpretation of this statistic is that schools account for approximately 2.3% of the variability in reading AC-Score between students. Based on the previously mentioned school effects range of .10 to .20, the ICC for reading is relatively small. The design effect for reading was determined to be 2.38, thus indicating that the standard errors would be inflated by a factor of 2.38 if a multilevel approach was not utilized for the reading analysis. McCoach suggests that, "design effects below 2.0 are considered fairly small;" however, she goes on to state that, "the Type 1 error rate is already noticeably inflated, even with such a small design effect" (2010, p. 135). Based on the ICC and design effect statistics calculated, and the nested nature of the data, it was determined that a multilevel approach would be beneficial for this study.

Random Coefficients Models

As mentioned previously, the random coefficients model only includes level-1 predictors. The basic statistical model was presented in the previous chapter in mixed-effects model form. As illustrated in the previously presented models (Tables 3.3 and 3.4), the level-1 assessment frequency slopes (TotalAssmts, SCAs, and LCAs), as well as the slopes for predictors representing at-risk students (EDS, LEP, and SWD), were estimated as randomly varying across level-2 units. Any of these variables determined to have non-statistically significant variance across schools in the initial analysis were then fixed. The intercept (AC-Score), was also allowed to vary by school. Slopes for all other level-1 predictors (gender, race/ethnicity) were estimated as fixed across schools. This assumes that, whereas the intercept for AC-Score and the slopes for formative assessment frequency and at-risk student subgroups may vary by school, the effects of gender and race remain fixed across schools.

In order to address Research Questions 1 and 2, two models were fit. First, addressing Research Question 1, a model was tested using total number of formative assessments taken along with control variables for gender, race, and at-risk student indicators (EDS, LEP, and SWD) as the predictors of AC-Score. In order to address Research Question 2, the short-cycle assessment (SCA) and long-cycle assessment (LCA) count variables were used as the primary predictor variables in place of the total formative assessment count variable. Variables meeting one of the two following criteria were retained in the model: 1) variable effects estimate must have been determined to be statistically significantly different from 0, and/or 2) variable variance across schools was determined to be statistically significantly different from 0. Model deviance was then calculated for each of the two models in order to determine which model provided a better fit to the data. The model providing a better fit to the data was used going forward. Interactions were then tested between the assessment frequency variable(s) and at-risk indicators in order to address potentially differing effects of formative assessment for at-risk students (Research Question 3). The results from the random coefficients model analyses for each subject are discussed in the following sections. Tables are provided for the best fit random coefficients model for each subject.

Random Coefficients Mathematics Model Results

Formative assessment frequency model analysis. The initial analysis of the total assessment count random coefficients model for mathematics indicated that total mathematics assessment frequency ($\text{TotalAssmts}_{\text{math}}$) was not statistically significantly different from 0 (p = 0.5319), however, the random effect for $\text{TotalAssmts}_{\text{math}}$ (u_{1j}) was determined to be statistically different from 0, indicating that the relationship between $\text{TotalAssmts}_{\text{math}}$ and AC-Score varies across schools and, therefore, the addition of level-2 variables may reveal such a relationship. In addition, in order to address research question 3, interactions were to be tested between at-risk student variables and TotalAssmts$_{\text{math}}$. For these reasons, the TotalAssmts$_{\text{math}}$ variable was retained in the model and remained estimated as randomly varying across schools. All other variables initially included in this model were determined to be statistically significantly different from 0 (p <.05). Each at-risk variable (EDS, LEP, and SWD) was determined to have variance across schools statistically significantly different from 0 and, as such, these variables continued to be estimated as randomly varying in the model. The analysis of the assessment count by cycle-length random coefficients model indicated that both short-cycle assessment count and long-cycle assessment length were not statistically significant predictors of mathematics AC-Score (p = .9018 and .6958, respectively). However, as was seen with the total assessment count analysis, the random effects for SCAs$_{\text{math}}$ (u_{1j}) LCAs$_{\text{math}}$ (u_{2j}) were determined to be statistically different from 0. As with the TotalAssmts$_{\text{math}}$ variable in the previous model, these variables were retained in the model and continued to be estimated as randomly varying across schools. The model fit statistics for each of these models suggested that the cycle-length specific model provided a better fit to the data. The deviance (-2LL) statistic for the total assessment count random coefficients model was 94382.8 as compared to 94005.8 for the cycle-specific model. The comparison of the BIC statistics for each model also suggested that the cycle-specific model (BIC=94114.2) provided a better fit to the data as compared to the total assessment count random coefficients model (BIC=94479.2). Based on these results, the cycle-specific model including SCAs$_{\text{math}}$, LCAs$_{\text{math}}$, gender, race (*Asian, African American, Hispanic, Other*), and at-risk student subgroups (*EDS, LEP, SWD*) was used going forward with the mathematics analyses.

Interactions. Although both the $\text{SCAs}_{\text{math}}$ and $\text{LCAs}_{\text{math}}$ frequency variables were not found to be a statistically significant predictors of mathematics AC-Score, interactions between at-risk students and $\text{SCAs}_{\text{math}}$ and $\text{LCAs}_{\text{math}}$ frequency were still tested to see if a statistically significant effect of existed among the different subgroups (Research Question 3).

The initial analyses testing the interactions between mathematics SCAs and LCAs and the three at-risk subgroups estimated the interactions as randomly varying across schools. The only interaction found to have statistically significant variance across schools was between $\text{LCAs}_{\text{math}}$ and SWD. In addition, the introduction of this interaction to the model as a randomly varying slope caused the random effect for the SWD slope to be statistically non-significant. Based on these results, the analysis was run again with SWD and all interactions fixed with the exception of the $\text{LCAs}_{\text{math}}$ and SWD interaction. Once each variable and interaction term was fixed as necessary and the analysis was re-run, the results suggested that the only interaction term estimate statistically significant from 0 ($p = 0.0101$) was the interaction between $\text{SCAs}_{\text{math}}$ and EDS. This coefficient estimate (.00195) suggested that $\text{SCAs}_{\text{math}}$ may have a positive effect for EDS students. All non-significant interactions were removed from the model moving forward. Although the $\text{SCAs}_{\text{math}}$ and $\text{LCAs}_{\text{math}}$ predictors were still non-significant predictors of mathematics AC-Score alone, $\text{SCAs}_{\text{math}}$ was retained due to the statistically significant interaction between this variable and EDS. In addition, $\text{LCAs}_{\text{math}}$ was retained as the randomly varying slope for this variable was still statistically significantly different from 0.

Table 4.11 *Random Coefficients Model for Mathematics*

Fixed Effect	*Coefficient*	*Standard Error*	*DF*	*T-Value*	*P-value*
Math AC-Score Intercept (γ_{00})	0.0644	0.010	412	6.31	<.0001
SCAs (γ_{10})	-0.0007	0.004	83,798	-0.20	0.8400
LCAs (γ_{20})	0.0023	0.006	83,798	0.41	0.6823
Gender (γ_{30})	0.0477	0.003	83,798	16.28	<.0001
Asian (γ_{40})	0.1069	0.014	83,798	7.63	<.0001
Afr. American (γ_{50})	0.0299	0.006	83,798	6.59	<.0001
Hispanic (γ_{60})	0.0245	0.006	83,798	4.07	<.0001
Other (γ_{70})	0.0207	0.008	83,798	2.70	0.0070
EDS (γ_{80})	-0.0429	0.004	83,798	-9.75	<.0001
LEP (γ_{90})	0.0282	0.010	83,798	2.76	0.0057
SWD (γ_{100})	-0.0224	0.006	83,798	-3.77	0.0002
EDS*SCAs (γ_{110})	0.002	0.001	83,798	2.59	0.0097
Random Effects	*Estimate*	*Standard Error*	*DF*	*Z-Value*	*P-value*
Intercept (u_{0j})	0.0187	0.0022	412	8.54	<.0001
SCAs (u_{1j})	0.0023	0.0003	412	7.13	<.0001
LCAs (u_{2j})	0.0028	0.0005	412	5.11	<.0001
EDS (u_{3j})	0.0008	0.0003	412	2.67	0.0038
LEP (u_{4j})	0.0025	0.0014	412	1.75	0.0403
Level-1 Variance (r_{ij})	0.1763	0.0009	412	203.19	<.0001

Final random coefficient mathematics model fixed effects results. As formulated above, the final random coefficients mathematics model included SCAs$_{\text{math}}$, LCAs$_{\text{math}}$, gender, race (Asian, African American, Hispanic, and Other – White as reference group), EDS, LEP, SWD status, and the SCAs$_{\text{math}}$ and EDS interaction term. The results from this analysis are presented in Table 4.11. In this model the intercept (γ_{00}) is no longer interpreted as the grand mean mathematics AC-Score. It is now interpreted as the expected mathematics AC-Score when the predictor variables are all 0 (i.e. a white, male student who took no $\text{SCAs}_{\text{math}}$ or $\text{LCAs}_{\text{math}}$, and is not EDS, LEP, or SWD). According to the data presented in Table 4.11, the reference student described above would be expected to have a mathematics AC-Score of 0.0644.

Frequency of assessment effects for mathematics. While mathematics $\text{SCAs}_{\text{math}}$ and $\text{LCAs}_{\text{math}}$ alone were not determined to be a statistically significant predictors of mathematics AC-Score, the interaction between $\text{SCAs}_{\text{math}}$ and EDS status estimate ($\gamma_{110} = 0.002$) was estimated to be statistically significantly different than 0. The results from this analysis suggested that EDS students who took more $\text{SCAs}_{\text{math}}$ achieved significantly (p <.0001) higher mathematics AC-Scores as compared to EDS students who took fewer. Mathematics AC-score for an EDS student would be expected to increase 0.002 for each additional $\text{SCAs}_{\text{math}}$ taken.

Gender and ethnicity effects for mathematics. The estimate for the gender predictor was determined to be statistically significant from 0 (p <.0001) for the final random coefficients mathematics model. The coefficient estimate for gender (γ_{30}) suggested that female students are expected to achieve a mathematics AC-Score 0.0477 points higher than male students. The results also suggested that Asian, African American, Hispanic, Other students were expected to achieve higher mathematics AC-Scores as compared to white students (Asian (γ_{40}): 0.1069, African American (γ_{50}): 0.0299, Hispanic (γ_{60}): 0.0245, and Other (γ_{70}): 0.0207).

Disadvantaged student effects for mathematics. All disadvantaged student predictor estimates were determined to be statistically significantly different from 0 (p <.05). The EDS (γ_{80}) estimate was -0.0429, which suggests that EDS students achieve mathematics AC-Scores 0.0429 lower as compared to non-EDS students. The LEP (γ_{90}) student predictor estimate was 0.0282, indicating that LEP students would be expected to produce a mathematics AC-Score 0.0282 higher than non-LEP students. This was an interesting finding and will be discussed in further depth in the following chapter. The coefficient estimate for SWD (γ_{100}) was -0.0224, suggesting that SWD students produce mathematics AC-Scores .0224 lower than non-SWD students.

Final random coefficients mathematics model random effects results. It is important to note that the random effects portion of the model results presented in Table 4.11 should not be thought of as effects but instead, as evidence of the un-modeled variability in the model. The variance components representing random effects for SCAs$_{\text{math}}$, LCAs$_{\text{math}}$, EDS, and LEP were all statistically significantly different from 0, suggesting that these slopes vary across schools. The random effects estimate for the mathematics AC-Score intercept (u_{oj} =.0187) was also significantly different from 0, suggesting that there is additional variation in school mean mathematics AC-Score that is not explained by the predictors and interaction terms included in this model and that additional school-level predictors would likely be beneficial to the model. Additional variables were added in the third and final model presented later in this chapter.

Random Coefficients Reading Model Results

Assessment count model analysis. The initial analysis of the total assessment count random coefficients reading model suggested that total reading assessment count (TotalAssmts$_{\text{read}}$) was statistically significantly different from 0 (p < .0001). The covariance parameter estimates for $\text{LCAs}_{\text{read}}$ and LEP were not statistically significantly different from 0, indicating that there was not significant variance in these variables across schools. Therefore, these variables were fixed for all reading analyses going forward. Of the remaining variables, the coefficient estimates for gender (p = .63) and the ethnicity category *Other* (p = .5736) were not statistically significant from 0, suggesting that these variables are not significant predictors of reading AC-Score. The variable *Other* was a dummy-coded variable, it was retained in the model despite being a statistically non-significant predictor. On the other hand, the gender variable was omitted from the model going forward. All other variables were determined to be statistically significantly different from 0 (p<.05) and, therefore, were retained in the model.

The analysis of the assessment count by cycle-length random coefficients reading model suggested that reading SCA frequency ($\text{SCAs}_{\text{read}}$) was a statistically significant predictor (p = .0059) of reading AC-Score whereas, reading LCAs ($\text{LCAs}_{\text{read}}$) were not ($p$ = .1039). However, the $\text{LCAs}_{\text{read}}$ variable was retained in order to allow for testing of interactions between the assessment frequency variables and the at-risk student variables (EDS, LEP, & SWD), in order to address research question 3.

The model fit estimates automatically calculated using the SAS Proc Mixed procedure are the -2LL, AIC, and BIC. Whereas the two models being compared had differing number of parameters being estimated, BIC was the most appropriate measure for model comparison as it penalizes models with more parameters in order to compensate for the fact that models with more parameters tend to have a lower deviance (-2LL). The BIC for the total formative assessment frequency reading model was 92921.5 and the BIC for the cycle-length specific model was 92909.2 indicating that the latter provides a better fit to the data. As such, the formative assessment frequency by cycle-length random coefficients reading model including $\text{SCAs}_{\text{read}}$, $\text{LCAs}_{\text{read}}$, race (*Asian, African American, Hispanic, Other*), and at-risk student subgroups (*EDS, LEP, SWD*) was used was used for the remaining reading analyses.

Interactions. Once the random coefficients model was fit, interactions between both $\text{SCAs}_{\text{read}}$ and $\text{LCAs}_{\text{read}}$ and each of the three at-risk student subgroups were tested to see if $\text{SCAs}_{\text{read}}$ or $\text{LCAs}_{\text{read}}$ had effects on reading AC-Score for different subgroups (Research Question 3). After testing, the only interaction determined to be statistically significantly different from 0 was between $\text{SCAs}_{\text{read}}$ and SWD students ($p = .0290$). As a result, this interaction was retained in the model and all others were removed. In addition, whereas the $\text{LCAs}_{\text{read}}$ variable was determined to not have statistically significant variance across schools, and also did not appear to have a statistically significant interaction with any of the at-risk student subgroups, the $\text{LCAs}_{\text{read}}$ was removed from the model as well.

Table 4.12 *Random Coefficients Model for Reading*

Fixed Effect	*Coefficient*	*Standard Error*	*DF*	*T-Value*	*P-value*
Reading AC-Score Intercept (γ_{00})	0.0861	0.005	412	17.64	<.0001
SCAs$_{read}$ (γ_{10})	0.0050	0.002	83,798	2.91	0.0036
Asian (γ_{20})	0.0565	0.014	83,798	4.07	<.0001
Afr. American (γ_{30})	-0.0363	0.004	83,798	-8.22	<.0001
Hispanic (γ_{40})	0.0258	0.006	83,798	4.32	<.0001
Other (γ_{50})	-0.0043	0.004	83,798	0.57	0.5715
EDS (γ_{60})	-0.0171	0.004	83,798	-4.62	<.0001
LEP (γ_{70})	-0.0237	0.009	83,798	-2.67	0.0077
SWD (γ_{80})	-0.0191	0.009	83,798	-2.16	0.0306
SWD*SCA$_{read}$ (γ_{90})	-0.0071	0.003	83,798	-2.18	0.0290
Random Effects	*Estimate*	*Standard Error*	*DF*	*Z-Value*	*P-value*
Intercept (u_{0j})	0.0033	0.0005	412	6.21	<.0001
SCA (u_{1j})	0.0002	0.0001	412	3.31	0.0036
EDS (u_{2j})	0.0004	0.0002	412	1.82	0.0347
SWD (u_{3j})	0.0049	0.0013	412	3.71	0.0001
Residual (r_{ij})	0.1757	0.0009	412	203.41	<.0001

Final random coefficient reading model fixed effects results. As formulated above, the final random coefficients reading model included SCAs$_{read}$, race (*Asian, African American, Hispanic, Other*), and at-risk student subgroups (*EDS, LEP, SWD*) and the interaction between SCAs$_{read}$ and SWD as the statistically significant level-1 predictors of reading AC-Score. The results from the analysis of this model are presented in Table 4.12.

As mentioned in the mathematics results section, the intercept (γ_{00}) in this model is no longer interpreted as the mean reading AC-Score. It is now interpreted as the expected reading AC-Score when the predictor variables are all 0 (i.e. a student who is white, male, took zero reading SCAs, and is not EDS, LEP, or SWD). According to the data presented in Table 4.12, a white male student who did not take any reading SCAs, and is not EDS, LEP, or SWD would be expected to achieve a reading AC-Score of 0.0861.

Cycle-length effects for reading. The coefficient estimate for $\text{SCAs}_{\text{read}}$ (γ_{10}) was 0.005 and statistically significantly different from 0 (p = 0.0036), suggesting that students who take more reading SCAs achieve higher reading AC-scores than students who take fewer. Reading AC-score would be expected to increase 0.005 for each additional reading SCA taken.

Ethnicity effects for reading. Again, all ethnic groups included in the final random coefficients reading model, with the exception of *Other*, were statistically significant, indicating that Asian, African American, and Hispanic students differ significantly in reading AC-score in comparison to Caucasian students. Whereas Asian and Hispanic students were determined to have higher expected AC-Scores as compared to white students (Asian (γ_{20}): 0.0565; Hispanic (γ_{40}): 0.0258), African American students were expected to produce lower reading AC-Scores than Caucasian students (African American (γ_{30}): -0.0363).

Disadvantaged student effects for reading. All disadvantaged student effects estimates (EDS, LEP, and SWD) were determined to be statistically significantly different from 0 (p <.05) and negative, suggesting that disadvantaged students are expected to exhibit lesser gains than non-disadvantaged students. EDS (γ_{60}) effects were estimated to be -0.0171, which suggests that the average EDS student is expected to produce a reading AC-Score 0.0171 lower as compared to non-EDS students. LEP (γ_{70}) student effects were estimated at -

0.0237, suggesting that LEP students would be expected to produce a reading AC-Score -0.0237 lower than non-LEP. The effect of being a SWD (γ_{80}) on reading AC-score was -0.0191 as compared to non-SWD students.

Interaction effects. The only level-1 interaction effect included in the final random coefficient reading model was the interaction term between $\text{SCAs}_{\text{read}}$ and SWD. This effect estimate of -0.007 (γ_{90}) suggested that SWD students who take more reading SCAs achieve lower reading AC-scores.

Final random coefficients model random effects results for reading. As mentioned in the mathematics final random coefficients model section, it is important to note that the random effects portion of the model results presented in Table 4.12 should not be thought of as "effects" but instead, as evidence of the un-modeled variability in the model. The variance components for $\text{SCAs}_{\text{read}}$ (u_{0j}) and the at-risk slopes EDS (u_{1j}) and SWD (u_{2j}) were significantly different from 0, suggesting that these slopes vary across schools. In addition, the random effect estimate for the reading AC-Score intercept was significantly different from 0, suggesting that additional school-level predictors would likely be beneficial to the model. Additional variables were added in the third and final model presented later in this chapter.

Full Contextual Models

The final models to be fit for this study were the full contextual models which were fit to the same data as the random coefficient models in the previous section. These models included both level-1 and level-2 predictors. The full-contextual statistical models for mathematics and reading were presented in the previous chapter in mixed-effects model form (Tables 3.7 – 3.10). Although the previously presented models indicated that the formative assessment frequency and at-risk student variables, as well as interactions between these variables would be allowed to vary randomly across schools, this only remained the case for slopes which were found to vary across schools with statistical significance (p<.05). All other level-1 variables (gender, race/ethnicity) were estimated as fixed across all level-2 units. As in the random coefficients models, the intercepts were also allowed to vary by school. This assumes that student performance (AC-Score) varies across schools. All level-1 variables from the final random coefficients models were initially included in the full model. Any estimated level-1 effects that were no longer statistically significant after the addition of level-2 variables were then eliminated in order to arrive at the best fitting, most parsimonious model for the data. The results from the full contextual models for each subject are presented in the following tables. A comprehensive table comparing the three models for each subject will follow the results for the full contextual model.

Full Contextual Mathematics Model Results

Model analysis. The first iteration of the full contextual model for mathematics included all of the level-1 variables and interaction terms presented in the random coefficients mathematics model (Table 4.11). In addition, in order to address research question 4, two level-2 variables (%EDS and %minority) and four interaction terms (%EDS*SCAs$_{math}$, %EDS*LCAs$_{math}$, %minority*SCAs$_{math}$, and %minority*LCAs$_{math}$) were added to the model and tested. Any statistically significant variables or interactions were retained in the model before adding the final two level-2 variables (school-level SCAs$_{math}$ mean and school-level LCAs$_{math}$ mean) and four interactions (school-level SCAs$_{math}$ mean*SCAs$_{math}$, school-level SCAs$_{math}$ mean*LCAs$_{math}$, school-level LCAs$_{math}$ mean*SCAs$_{math}$, and school-level LCAs$_{math}$ mean*LCAs$_{math}$). Again, any statistically significant predictors or interactions were retained in what would represent the final full contextual mathematics model.

Of the variables and interactions tested to specifically address research question 4, only %EDS was shown to have an effect which was statistically significantly different from 0 (p = .0004). As a result, %EDS was the only level-2 variable retained in the model before adding the variables and interactions addressing research question 5. Of the variables and interactions addressing research question 5, only school-level $\text{LCAs}_{\text{math}}$ mean and the school-level $\text{LCAs}_{\text{math}}$ mean*$\text{LCAs}_{\text{math}}$ interaction were found to be statistically different from 0. Based on the results from these analyses, the final full contextual model for mathematics included $\text{SCAs}_{\text{math}}$, $\text{LCAs}_{\text{math}}$, Gender, Asian, African American, Hispanic, Other, EDS, LEP, SWD, $\text{SCAs}_{\text{math}}$*EDS, %EDS, school-level $\text{LCAs}_{\text{math}}$ mean, and school-level $\text{LCAs}_{\text{math}}$ mean*$\text{LCAs}_{\text{math}}$. The results from this model are presented in Table 4.13 and are discussed in detail in the next section.

Table 4.13 *Full Contextual Model for Mathematics*

Fixed Effect	*Coefficient*	*Standard Error*	*DF*	*T-Value*	*P-value*
Math AC-Score Intercept (γ_{00})	0.1712	0.027	410	6.38	<.0001
% EDS (γ_{01})	-0.1386	0.041	410	-3.41	0.0007
LCAs Mean (γ_{02})	-0.0257	0.009	410	-3.01	0.0028
SCAs (γ_{10})	-0.0010	0.004	83,798	-0.27	0.7879
LCAs (γ_{20})	-0.0331	0.013	83,798	-2.64	0.0083
LCAs*LCAs$_{mean}$ (γ_{21})	0.0186	0.005	83,798	3.66	0.0002
Gender (γ_{30})	0.0476	0.003	83,798	16.26	<.0001
Asian (γ_{40})	0.1066	0.014	83,798	7.61	<.0001
Afr. American (γ_{50})	0.0307	0.005	83,798	6.75	<.0001
Hispanic (γ_{60})	0.0245	0.006	83,798	4.07	<.0001
Other (γ_{70})	0.0210	0.008	83,798	2.74	0.0062
EDS (γ_{80})	-0.0416	0.004	83,798	-9.40	<.0001
LEP (γ_{90})	0.0285	0.010	83,798	2.74	0.0054
SWD (γ_{100})	-0.0257	0.006	83,798	-3.75	0.0002
EDS*SCAs (γ_{110})	0.0019	0.001	83,798	2.48	0.0130

Random Effects	*Estimate*	*Standard Error*	*DF*	*Z-Value*	*P-value*
Intercept (u_{0j})	0.0168	0.0002	410	8.43	<.0001
SCAs (u_{1j})	0.0023	0.0003	410	7.14	<.0001
LCAs (u_{2j})	0.0025	0.0005	410	5.12	<.0001
EDS (u_{3j})	0.0008	0.0003	410	2.68	0.0037
LEP (u_{4j})	0.0026	0.0015	410	1.76	0.0393
Level-1 Variance (r_{ij})	0.1763	0.0009	410	203.21	<.0001

Note: Be wary of interpreting negative main effects in the presence of statistically significant interactions (LCAs);

Final full contextual mathematics model fixed effects results. In this model, as in the random coefficients mathematics model, the intercept (γ_{00}) was interpreted as the expected mathematics AC-Score when the predictor variables are all 0 (i.e. white, male student who took no $\text{SCAs}_{\text{math}}$ or $\text{LCAs}_{\text{math}}$, and is not EDS, LEP, or SWD, and attends a school with 0% EDS). According to the results presented in Table 4.13, the expected mathematics AC-Score (γ_{00}) when all predictors are 0 is 0.1712.

Level-2 effects for mathematics. The %EDS coefficient (γ_{01}) of -0.1386 suggests that students at schools with a higher percentage of EDS students exhibit lesser gains (remember that AC-Score is a measure of change in student score from year to year) than students at schools with a lower percentage of EDS students.

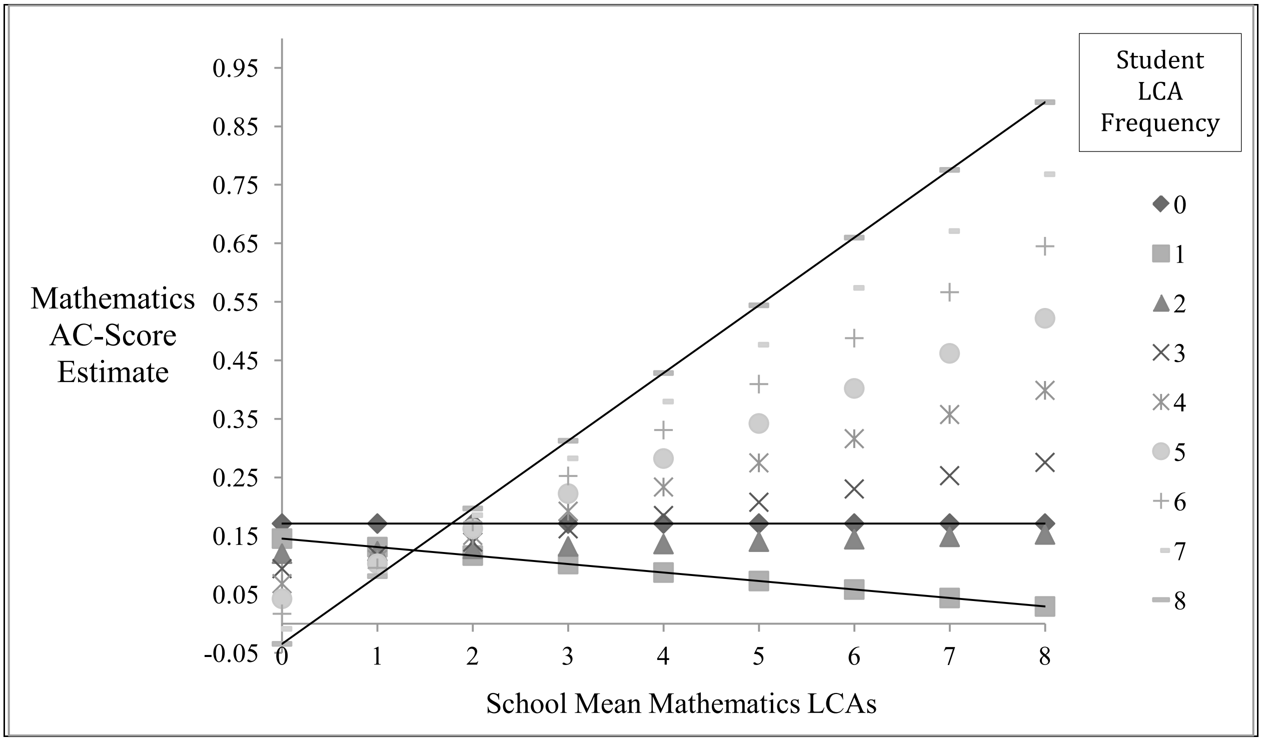
Formative assessment frequency effects for mathematics. The coefficient estimate for $\text{SCAs}_{\text{math}}$ (γ_{10}) again failed to be proven statistically significantly different from 0 (p = .7879). This finding was not entirely unexpected as this variable had proven to be non-significant throughout the entire model building process. However, it again was retained as the estimate for the interaction between $\text{SCAs}_{\text{math}}$ and EDS, discussed in a subsequent section, was again statistically significant from 0. On the other hand, with the addition of the school-level $\text{LCAs}_{\text{math}}$ mean and the interaction between school-level $\text{LCAs}_{\text{math}}$ mean and $\text{LCAs}_{\text{math}}$, the coefficient estimate for the $\text{LCAs}_{\text{math}}$ variable (γ_{20}), which had not previously been statistically significantly different from 0, was now statistically significantly different from 0 (p = .0083). The coefficient estimate of -0.0331 for $\text{LCAs}_{\text{math}}$ (γ_{20}) suggests that each $\text{LCAs}_{\text{math}}$ taken by a student results in a mathematics AC-Score 0.0331 lower as compared to students who took none. The $\text{SCAs}_{\text{math}}$*EDS and school-level $\text{LCAs}_{\text{math}}$ mean*$\text{LCAs}_{\text{math}}$ interaction effects are discussed in a later section.

Gender and ethnicity effects for mathematics. The coefficient for gender (γ_{30}) suggested that females are expected to produce a mathematics AC-Score 0.0476 points higher as compared to males. All other ethnic groups were determined to have higher expected mathematics AC-scores as compared to white students (Asian (γ_{40}): 0.1066, African American (γ_{50}): 0.0307, Hispanic (γ_{60}): 0.0245, and Other (γ_{70}): 0.021).

Disadvantaged student effects for mathematics. All effects estimates for disadvantaged student status were determined to be statistically significantly different from 0. EDS (γ_{80}) effects were estimated to be -0.0416, which suggests that EDS students achieve mathematics AC-Scores 0.0416 lower than non-EDS students. The LEP student coefficient estimate (γ_{90}) was 0.029, suggesting that LEP students would be expected to produce mathematics AC-Scores 0.029 higher than non-LEP students. The SWD (γ_{100}) coefficient estimate was -0.022, suggesting that SWDs achieve mathematics AC-Scores 0.022 lower than non-SWD students. These values are similar to those from the random coefficients model analysis (Table 4.11).

Interaction effects. The interaction effect between mathematics SCA frequency and EDS (γ_{110}) was estimated to be 0.0019, suggesting that mathematics SCAs may have a positive effect on mathematics performance for EDS students. This is similar to what was evidenced in the mathematics random coefficient model. The coefficient estimate for the interaction between school-level $\text{LCAs}_{\text{math}}$ mean and $\text{LCAs}_{\text{math}}$ ($\gamma_{21} = 0.0186$) was statistically significantly different from 0 ($p = .0002$) and suggested that $\text{LCAs}_{\text{math}}$ frequency may have a positive effect on student mathematics achievement in schools that have higher $\text{LCAs}_{\text{math}}$ means. For example, a student who took four $\text{LCAs}_{\text{math}}$ in a school with an $\text{LCAs}_{\text{math}}$ mean of one would be expected to score 0.0746 higher than a student who took no $\text{LCAs}_{\text{math}}$. Even more, a student who took four LCAs$_{math}$ at a school with an LCAs$_{math}$ mean of four would be expected to score 0.2982 higher than a student who took no LCAs$_{math}$. This was an interesting finding given that these two variables alone produced statistically significant estimates indicating a negative effect on student mathematics AC-Score. Figure 4.1 illustrates the effects of this interaction within the context of the negative effects of LCAs$_{math}$ mean (γ_{02}) and LCAs$_{math}$ (γ_{20}) alone.

Figure 4.1 *Mathematics AC-Score Estimates for Baseline Student – Based on LCAs x School LCAs Mean Interaction*



**Baseline student = White, Male, non-EDS, non-SWD, non-LEP, 0% EDS school;*

Full contextual model random effects results for mathematics. The random effects estimates for the SCAs$_{math}$ (u_{1j}), LCAs$_{math}$ (u_{2j}), EDS (u_{3j}), and SWD (u_{4j}) slopes remained statistically significantly different from 0. The estimates for level-1 (r_{ij} = .1807) and level-2 (u_{0j} = .01366) variation in mathematics AC-Score were also significantly different from 0 which suggests that there is still some unexplained between-school and between-individual variance in mathematics AC-score.

Full Contextual Reading Model Results

Model analysis. The first iteration of the full contextual model for reading included all of the level-1 fixed effects presented in the random coefficients reading model (Table 4.12). In addition, in order to address research question 4, two level-2 variables (%EDS and %minority) and two interaction terms (%EDS*SCAs$_{read}$ and %minority*SCAs$_{read}$) were added to the model and tested. Any statistically significant variables or interactions were retained in the model before adding the final level-2 variable (school-level SCAs$_{read}$ mean) and interaction term (school-level SCAs$_{read}$ mean*SCAs$_{read}$). Again, any statistically significant predictors or interactions were retained in what would represent the final full contextual reading model.

Of the variables and interactions tested to specifically address research question 4, only %EDS and %minority were statistically significantly different from 0 (p<.05). As such, these level-2 predictors were retained in full contextual reading model. The interactions between SCAs$_{read}$ and these variables were not statistically significant and, as a result, were removed from the model. The school-level SCAs$_{read}$ mean variable and school-level SCAs$_{read}$ mean*SCAs$_{read}$ interaction were then tested to address research question 5. Results indicated that school-level SCAs$_{read}$ mean is a statistically significant predictor of reading AC-Score but that the interaction between school-level SCAs$_{read}$ mean and SCAs$_{read}$ is not. Based on the results from these analyses, the final full contextual model for mathematics included SCAs$_{read}$, Asian, African American, Hispanic, Other, EDS, LEP, SWD, SCAs$_{read}$*SWD, %EDS, %Minority, and school-level SCAs$_{read}$ mean. The results from this model are presented in Table 4.14 and are discussed in detail in the next section.

Table 4.14 *Full Contextual Model for Reading*

Fixed Effect	*Coefficient*	*Standard Error*	*DF*	*T-Value*	*P-value*
Reading AC-Score Intercept (γ_{00})	0.1494	0.013	409	11.36	<.0001
%EDS (γ_{01})	-0.1582	0.030	409	-5.29	<.0001
%Minority (γ_{02})	0.0459	0.021	409	2.18	0.0327
SCAs$_{read}$ Mean (γ_{03})	0.0063	0.003	409	2.09	0.0365
SCAs$_{read}$ (γ_{10})	0.0048	0.002	83,798	2.75	0.0060
Asian (γ_{20})	0.0564	0.014	83,798	4.06	<.0001
Afr. American (γ_{30})	-0.0349	0.005	83,798	-7.70	<.0001
Hispanic (γ_{40})	0.0260	0.006	83,798	4.35	<.0001
Other (γ_{50})	0.0050	0.008	83,798	0.65	0.5139
EDS (γ_{60})	-0.0146	0.004	83,798	-3.89	<.0001
LEP (γ_{70})	-0.0229	0.009	83,798	-2.58	0.0099
SWD (γ_{80})	-0.0118	0.009	83,798	-2.14	0.0327
SWD*SCA (γ_{90})	-0.0070	0.003	83,798	-2.15	0.0315
Random Effects	*Estimate*	*Standard Error*	*DF*	*Z-Value*	*P-value*
Intercept (u_{0j})	0.0027	0.0005	409	5.85	<.0001
SCAs (u_{1j})	0.0002	0.0001	409	3.27	0.0005
EDS (u_{2j})	0.0004	0.0002	409	1.91	0.0277
SWD (u_{3j})	0.0048	0.0013	409	3.69	0.0001
Residual (r_{ij})	0.1758	0.0009	409	203.43	<.0001

Final full contextual model fixed effects results. The first iteration of the full contextual model for reading included all of the level-1 variables from the final random coefficients reading model. In addition, the three level-2 variables were included in the final full contextual reading model.

In this model, as in the random coefficients reading model, the intercept (γ_{00}) was interpreted as the expected reading AC-Score when the predictor variables are all 0 (i.e. a student who is white, male, took zero $\text{SCAs}_{\text{read}}$, is not EDS, LEP, or SWD, attends a school with 0% EDS, 0% minority students, and an $\text{SCAs}_{\text{read}}$ mean of 0). According to the data presented in Table 4.14, the expected reading AC-Score (γ_{00}) when all predictors are 0 is 0.1494.

Level-2 effects for reading. All three level-2 variables were determined to be statistically significantly different from 0 for the full contextual reading model. As was illustrated in the mathematics model, %EDS coefficient ($\gamma_{01} = -0.1582$) suggests that students at schools with a higher percentage of EDS students are estimated to achieve lower reading AC-scores than students at schools with a lower percentage of EDS students. On the other hand, students at schools with a higher %minority students were estimated to achieve higher reading AC-scores (γ_{02}=0.0459). Students at schools with a higher $\text{SCAs}_{\text{read}}$ mean were also estimated to have higher reading AC-Scores (γ_{03}=0.0063).

Cycle-length effects for reading. The coefficient estimate for reading SCAs (γ_{10} = 0.0048) suggests that students who take more $\text{SCAs}_{\text{read}}$ produce higher AC-scores for reading. Reading AC-score would be expected to increase 0.0048 for each additional reading SCA taken. Figure 4.2 illustrates the positive relationship between $\text{SCAs}_{\text{read}}$ and $\text{SCAs}_{\text{read}}$ Mean on student reading AC-Score.

Gender and ethnicity effects for reading. As in the random coefficient reading model, Asian (γ_{20}= 0.056) and Hispanic (γ_{40}= 0.026) students were determined to have higher expected reading AC-scores as compared to white students whereas African American ($\gamma_{30=}$ -0.035) students were estimated produce lower reading AC-Scores in comparison to white students. Students indicated as *Other* (γ_{50}) were not determined to be statistically significantly different from white students (p = 0.5139).

Disadvantaged student effects for reading. All disadvantaged student effects estimates were determined to be statistically significantly different from 0 (p <.05). EDS (γ_{60}) effects were estimated to be -0.0146, which suggests that EDS students produce reading AC-Scores 0.0146 lower than non-EDS students. LEP (γ_{70} = -0.0229) students were estimated to produce lower reading AC-Scores than non-LEP students. The fixed effect for SWD students (γ_{80} =-0.0188) suggested that SWD students produce lower reading AC-scores as compared to non-SWD students. In addition, the fixed effect estimate for the interaction between SWD and $\text{SCAs}_{\text{read}}$ was negative (γ_{90} =-0.007) and statistically significantly different from 0 (p=.0327), suggesting that SWDs who took more $\text{SCAs}_{\text{read}}$ achieved lower reading AC-Scores.

Full contextual model random effects results for reading. The random effects estimates for $\text{SCAs}_{\text{read}}$, EDS, and SWD slopes were all statistically significantly different from 0, suggesting that these slopes vary across schools. The random effects estimate for the intercept and residual remained virtually unchanged from the random coefficients reading model. This suggests that the addition of the level-2 predictors failed to explain a great deal of the between school variance in reading AC-score.

Model Comparisons

For ease of interpretation and to best illustrate how the model fit changed between the unconditional, random coefficient, and full contextual models, Tables 4.15 and 4.16 present the fixed effects parameter estimates for each model side by side. Tables 4.17 and 4.18 present the variance components as well as deviance estimates for each model. A discussion of these results is provided in the following section.

Figure 4.2 *Reading AC-Score Estimates for Baseline Student – Based on SCAs x School SCAs Mean*

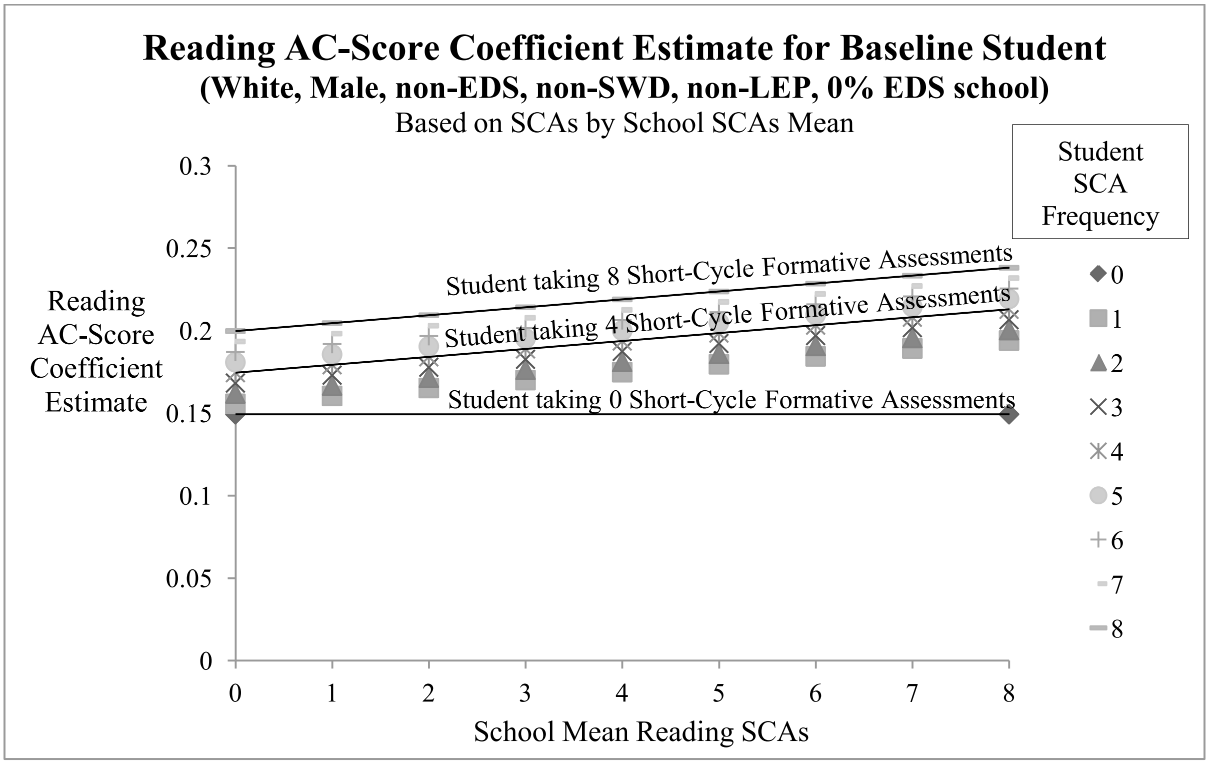


Table 4.15 *Model Comparison - Mathematics*

	Model 1 (Unconditional)				*Model 2 (Random Coefficients)*				*Model 3 (Full Contextual)*			
Fixed Effects	*Coef.*	*SE*	*T-ratio*	*p*	*Coef.*	*SE*	*T-ratio*	*p*	*Coef.*	*SE*	*T-ratio*	*p*
For Intercept ($\boldsymbol{\beta_{0j}}$)												
Intercept (γ_{00})	0.096	0.008	12.32	<.0001	0.064	.010	6.31	<.0001	0.1712	.027	6.38	<.0001
% EDS (γ_{01})									-.1386	.041	-3.41	0.0007
School $\overline{\mathbf{LCAs_{math}}}$($\gamma_{02}$)									-.0257	.009	-3.01	0.0002
For $\mathbf{SCAs_{math}}$ slope ($\boldsymbol{\beta_{1j}}$)												
$\mathbf{SCAs_{math}}$ (γ_{10})					-.001	.004	-0.20	0.8400	-.0010	.004	-0.27	0.7879
For $\mathbf{LCAs_{math}}$ slope ($\boldsymbol{\beta_{2j}}$)												
$\mathbf{LCAs_{math}}$ (γ_{20})					0.002	.006	0.41	0.6823	-.0331	.013	-2.64	0.0083
$\mathbf{LCAs_{math}}$*School $\overline{\mathbf{LCAs_{math}}}$ (γ_{21})									0.0186	.005	3.66	0.0002
For Gender slope ($\boldsymbol{\beta_{3j}}$)												
Gender (γ_{30})					0.048	.003	16.28	<.0001	0.048	.003	16.26	<.0001
For Asian slope ($\boldsymbol{\beta_{4j}}$)												
Asian slope (γ_{40})					0.107	.014	7.63	<.0001	0.107	.014	7.61	<.0001
For Afr. Amer. slope ($\boldsymbol{\beta_{5j}}$)												
Afr. Amer. slope (γ_{50})					0.030	.005	6.59	<.0001	0.031	.005	6.75	<.0001
For Hispanic slope ($\boldsymbol{\beta_{6j}}$)												
Hispanic slope (γ_{60})					0.025	.006	4.07	<.0001	0.024	.006	4.07	0.0002
For Other slope ($\boldsymbol{\beta_{7j}}$)												
Other slope (γ_{70})					0.021	008	2.70	0.0070	0.021	.008	2.74	0.0062
For EDS slope ($\boldsymbol{\beta_{8j}}$)												
EDS slope (γ_{80})					-0.043	.004	-9.75	<.0001	-0.042	.004	-9.40	<.0001
For LEP slope ($\boldsymbol{\beta_{9j}}$)												
LEP slope (γ_{90})					0.028	.010	2.76	0.0057	0.029	.010	2.78	0.0054
For SWD slope ($\boldsymbol{\beta_{10j}}$)												
SWD slope (γ_{100})					-0.022	.006	-3.77	0.0002	-0.022	.006	-3.75	0.0013
For EDS*$\mathbf{SCAs_{math}}$ ($\boldsymbol{\beta_{11j}}$)												
EDS*$\mathbf{SCAs_{math}}$ (γ_{110})					0.002	.001	2.59	0.0097	0.0019	.001	2.48	0.0002

Table 4.16 *Model Comparison - Reading*

	Model 1 (Unconditional)				*Model 2 (Random Coefficients)*				*Model 3 (Full Contextual)*			
Fixed Effects	*Coef.*	*SE*	*T-ratio*	*p*	*Coef.*	*SE*	*T-ratio*	*p*	*Coef.*	*SE*	*T-ratio*	*p*
For Intercept ($\boldsymbol{\beta_{0j}}$)												
Intercept (γ_{00})	0.074	0.005	16.47	<.0001	0.086	.005	17.64	<.0001	0.149	0.013	11.36	<.0001
% EDS (γ_{01})									-0.158	0.030	-5.28	<.0001
% Minority (γ_{02})									0.046	0.021	2.14	0.0327
School $\overline{\mathbf{SCAs}_{\mathbf{read}}}$ (γ_{03})									0.006	0.003	2.10	0.0365
For $\mathbf{SCAs}_{\mathbf{read}}$slope ($\boldsymbol{\beta_{1j}}$)												
$\mathbf{SCAs}_{\mathbf{read}}$ (γ_{10})					0.005	0.002	2.91	0.0036	0.005	0.002	2.75	0.0060
For Asian slope ($\boldsymbol{\beta_{2j}}$)												
Asian slope (γ_{20})					0.057	0.014	4.07	<.0001	0.056	0.014	4.06	<.0001
For Afr. Amer. slope ($\boldsymbol{\beta_{3j}}$)												
Afr. American slope (γ_{30})					-0.036	0.004	-8.22	<.0001	-0.035	0.005	-7.70	<.0001
For Hispanic slope ($\boldsymbol{\beta_{4j}}$)												
Hispanic slope (γ_{40})					0.026	0.006	4.32	<.0001	0.026	0.006	4.35	<.0001
For Other slope ($\boldsymbol{\beta_{5j}}$)												
Other slope (γ_{50})					0.004	0.008	0.57	0.5715	0.005	0.008	0.65	0.5139
For EDS slope ($\boldsymbol{\beta_{6j}}$)												
EDS slope (γ_{60})					-0.017	0.004	-4.62	<.0001	-0.015	0.004	-3.89	<.0001
For LEP slope ($\boldsymbol{\beta_{7j}}$)												
LEP slope (γ_{70})					-0.024	0.009	-2.67	0.0077	-0.023	0.009	-2.58	0.0099
For SWD slope ($\boldsymbol{\beta_{80j}}$)												
SWD slope (γ_{80})					-0.019	0.009	-2.16	0.0306	-0.019	0.009	-2.14	0.0327
For SCAs*SWD ($\boldsymbol{\beta_{90j}}$)												
$\mathbf{SCAs}_{\mathbf{read}}$*SWD ($\gamma_{90}$)					-0.007	0.003	-2.18	0.0290	-0.007	0.003	-2.15	0.0315

Table 4.17 *Model Comparison - Mathematics*

	Model 1 (Unconditional)				*Model 2 (Random Coefficients)*				*Model 3 (Full Contextual)*			
Random Effects	*SE*	*Var. Comp.*	*z*	*p*	*SE*	*Var. Comp.*	*z*	*p*	*SE*	*Var. Comp.*	*z*	*p*
Intercept ($u_{\mathbf{0}j}$)	0.0016	0.0154	9.37	<.0001	0.0022	0.0187	8.54	<.0001	0.0002	0.0168	8.43	<.0001
$\mathbf{SCAs_{math}}$ (u_{1j})					0.0003	0.0023	7.13	<.0001	0.0003	0.0024	7.14	<.0001
$\mathbf{LCAs_{math}}$ (u_{2j})					0.0005	0.0028	5.11	<.0001	0.0005	0.0025	5.12	<.0001
EDS (u_{3j})					0.0003	0.0008	2.67	0.0038	0.0003	0.0008	2.68	0.0005
LEP (u_{4j})					0.0014	0.0025	1.75	0.0403	0.0015	0.0026	1.76	0.0393
Residual ($r_{\mathbf{ij}}$)	0.0009	0.1829	204.24	<.0001	0.0009	0.1763	203.19	<.0001	0.0009	0.1763	203.21	<.0001
Model Fit	*Deviance*	*Param.*	*AIC*	*BIC*	*Deviance*	*Param.*	*AIC*	*BIC*	*Deviance*	*Param.*	*AIC*	*BIC*
	96223.2	0	96229.2	96241.2	94019.1	11	94055.1	94127.5	93989.2	14	94031.2	94115.7

Table 4.18 *Model Comparison - Reading*

	Model 1 (Unconditional)				*Model 2 (Random Coefficients)*				*Model 3 (Full Contextual)*			
Random Effects	*SE*	*Var. Comp.*	*z*	*p*	*SE*	*Var. Comp.*	*z*	*p*	*SE*	*Var. Comp.*	*z*	*p*
Intercept ($u_{\mathbf{0}j}$)	0.0006	0. 0041	6.99	<.0001	0.0005	0.0033	6.21	<.0001	0.0005	0.0027	5.85	<.0001
$\mathbf{SCAs_{read}}$ (u_{1j})					0.0001	0.0002	3.31	0.0005	0.0001	0.0002	3.27	0.0005
EDS (u_{2j})					0.0002	0.0004	1.82	0.0347	0.0002	0.0004	1.92	0.0277
SWD (u_{3j})					0.0013	0.0049	3.71	0.0001	0.0013	0.0048	3.69	0.0001
Residual ($r_{\mathbf{ij}}$)	0.0009	0.1769	204.18	<.0001	0.0009	0.1757	203.41	<.0001	0.0009	0.1758	203.43	<.0001
Model Fit	*Deviance*	*Param.*	*AIC*	*BIC*	*Deviance*	*Param.*	*AIC*	*BIC*	*Deviance*	*Param.*	*AIC*	*BIC*
	93128.7	0	93134.7	93146.8	92816.7	9	92846.7	92907.0	92782.8	12	92818.8	92891.2

Deviance and Model Fit

By default, SAS Proc Mixed provides the deviance statistic (-2LL) with the output for each model. Also provided are the Akaike Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC). As mentioned in the previous chapter, none of these statistics can be interpreted directly but, they can be used to compare multiple models to one another. Tables 4.17 and 4.18 present the variance components for each of the fit models as well as the model fit statistics mentioned above. Discussion of these comparisons is provided in the following two sections.

Deviance and model fit for mathematics. As illustrated in Table 4.17, the deviance statistic decreased from the unconditional model to the random coefficients model indicating that the addition of the level-1 variables provided a better fit to the data as compared to the unconditional model, which was expected. The deviance statistic decreased yet again from the random coefficients reading model to the full contextual mathematics model, suggesting that the addition of the level-2 variables provided a better fit to the data. However, as was mentioned previously, the deviance statistic is affected by the number of parameters being estimated, whereas a model with more parameters is more likely to produce a lower deviance statistic and indicate a better fit as compared to a model with fewer parameters.

However, the AIC and BIC penalize models with more parameters and, therefore, are less sensitive when comparing models with differing number of parameters. As with the deviance statistic, both the AIC and BIC decreased from the random coefficients model to the full contextual model thus indicating that the full contextual mathematics model provided the best fit for the data.

Deviance and model fit for reading. The deviance statistics for the reading models presented in Table 4.18, illustrate the decrease in deviance as each reading model increases in complexity, indicating that the full contextual reading model provides a better fit as compared to the random coefficients reading model. Although a more parsimonious model is preferable, in this case, the addition of the three level-2 variables resulted in a model fit that was significantly better fit for these data. In addition to the statistically significant difference in model fit based on the deviance statistics, the AIC and BIC statistics both suggested that the full contextual model was a better fit as well.

Predictive Ability

Given that the models fit for this study were multi-level in nature, included random intercepts and random slopes, it was necessary to calculate the proportional reduction in prediction error in a few steps. First, the proportional reduction in prediction error for predicting the level-1 outcome (R_1^2) was calculated. Then, the same was done for the level-2 mean (R_2^2). In calculating each of these statistics, the unconditional model for each subject served as the baseline model and the best fit, final model was used as the comparison model. As discussed in the previous chapter, in order to account for the additional variance estimates introduced due to the randomly varying slopes, the final comparison model was re-fit with the randomly varying slopes omitted so that only the level-1 and level-2 variance components needed for the calculations were produced. The equations for proportional reduction in prediction error are illustrated in the previous chapter. The results from these calculations are provided in the following sections.

Predictive ability for mathematics. Whereas the full contextual model provided the best fit for the data as compared to the random coefficients model for mathematics, the level-

1 and level-2 variance components from the former were used for calculating the proportional reduction in prediction error for mathematics. The level-1 proportional reduction in prediction error for the mathematics AC-Score intercept was 0.0181. The level-2 proportional reduction in prediction error for mathematics was estimated to be 0.116. These results indicated that the full contextual mathematics model, by including both level-1 and level-2 variables was able to improve the predictive ability of the model compared to the unconditional model by approximately 1.8% to 11.6%.

Predictive ability for reading. The full contextual reading model was estimated to be the best fit model for the data and, as such, it was used as the comparison model for this calculation. The level-1 proportional reduction in prediction error for the reading AC-Score intercept was 0.011. The level-2 proportional reduction in prediction error for reading was estimated to be 0.262. These results indicated that the full contextual mathematics model, by including seven level-1 predictors and one interaction, was only able to improve the predictive ability of the model compared to the unconditional model by approximately 1.0% to 26.2%.

Summary of Results

Assessment Frequency and Assessment Cycle-Length

Results for both mathematics and reading in regards to formative assessment frequency and assessment cycle-length were mixed. The mathematics analysis suggested that mathematics short-cycle formative assessments produced no effect on mathematics AC-Score that was statistically significantly different from 0. On the other hand, mathematics long-cycle formative assessments were found to be statistically significantly different from 0, but the coefficient estimate was negative (γ_{20} = -.0331), suggesting that long-cycle formative mathematics assessments have a negative effect on student mathematics AC-Score. In addition, school-level mean mathematics long-cycle formative assessment frequency was found to have a statistically significant, negative effect on student mathematics AC-Score (γ_{02} = -0.0257). However, the interesting finding came from the interaction between student mathematics long-cycle formative assessment frequency and school-level mean mathematics long-cycle formative assessment frequency. This interaction was shown to have a statistically significant, positive effect on student mathematics AC-Score. The illustration provided in Figure 4.1 shows that, despite the negative effects of each of these variables alone, their interaction can result in gains above what would be expected for a comparable student who took no mathematics long-cycle formative assessments. For example, a student who takes four long-cycle mathematics formative assessments in a school that averages three long-cycle mathematics formative assessments per student would meet or exceed what would be expected for a student taking no long-cycle mathematics formative assessments. A possible explanation for this relationship is that the level of school involvement or commitment to implementing the long-cycle formative assessments within the school may be a major factor in long-cycle formative assessments being successful on the individual level.

The reading analysis suggested that reading short-cycle formative assessments have a statistically significant, positive effect on student reading AC-Scores (γ_{10} = .005). Furthermore, schools with a higher mean number of reading short-cycle formative assessments were shown to have a statistically significant, positive effect on student reading AC-Score (γ_{03} = .006). Figure 4.2 illustrates the effect that the student and school-level short-cycle formative assessment frequency combined has on student reading AC-Score. Reading LCA frequency was determined to be a statistically non-significant predictor AC-Score for reading. Results from the interaction between reading SCAs and SWDs suggest that SWDs who take a higher frequency of reading SCAs achieve lower reading AC-Scores as compared to SWDs who take fewer reading SCAs. However, the SWD category is rather broad and delineation among different levels of SWDs could potentially produce different results.

Gender, Ethnicity, and Disadvantaged Students

There was no statistically significant difference between genders in predicting reading AC-Score, but there was for mathematics. Females were estimated to achieve mathematics AC-Scores approximately .048 higher in comparison to their male counterparts.

All ethnic groups in the study were statistically significantly different from white students for mathematics AC-score. Results suggest that Asian, African American, Hispanic, and students of *other* ethnicity achieve higher mathematics AC-Scores as compared to White students. However, for reading, only Asian and Hispanic students were estimated to achieve higher reading AC-Scores as compared to White students. Results suggest that African American students achieve lower reading AC-Scores as compared to White students, and that students of *Other* ethnicities were not statistically significantly different from White students.

Results suggested that all at-risk students (EDS, LEP, and SWD) have lower estimated reading AC-Scores as compared to non-disadvantaged students whereas only EDS and SWD students were predicted to achieve lower mathematics AC-Scores. LEP students were expected to have higher mathematics AC-scores as compared to non-disadvantaged students. This was an interesting finding; however, it would seem that LEP students would be less vulnerable to struggling in mathematics as opposed to reading since numbers are universal and not language specific.

School-level EDS was a statistically significant predictor of student AC-score for both mathematics and reading. For both subjects, students attending schools with a higher percentage of EDS students achieved lower AC-scores. Whereas results suggest that students at schools with a higher percentage of minority students achieve higher reading AC-scores as compared to students at schools with lower percentage of minority students, %minority was not determined to be a statistically significant predictor of mathematics AC-Score.

These results and their potential implications as well as suggestions for future research in this area are provided in the following chapter.

CHAPTER 5: SUMMARY, DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS

This chapter is divided into five sections. The first three provide a summary of the study, its findings and conclusions. The last two sections discuss the implications of this research, and include suggestions for future research.

Summary of Study

This study focused on the formative use of mathematics and reading assessments. Whereas many different assessments may fall under the formative umbrella, a generally accepted definition of formative assessment is, “frequent, interactive assessments of student progress and understanding to identify learning needs and adjust teaching appropriately” (OECD, 2005, p. 21).

Purpose and Data Collection

The overriding purpose of this study was to examine the effects of formative assessment frequency on student achievement on end of year summative assessments. In order to examine this relationship it was necessary first to determine what formative assessment means and how it is used in an educational setting. Because of this, formative assessments, along with a myriad of subtopics related to assessments, were the primary focus of the literature review. The present study built upon previous research which has suggested that the use of formative assessments may improve student achievement, with good potential to benefit at-risk students in particular (Burns et al, 2010; Bergan et al., 1991; Black & Wiliam, 1998; Foster & Poppers, 2009; Fuchs & Fuchs, 1986; Martinez & Martinez, 1992; Miesels et al, 2003; Peterson & Siadat, 2009; Sadler, 1989; White & Frederiksen, 1998).

Data for this study were obtained from multiple sources. Student achievement data as well as demographic information were obtained from the North Carolina Department of Public Instruction (NCDPI). Student formative assessment data was obtained from a private online formative assessment program (OFAP) vendor.

Restatement of the Research Questions

This study sought to answer one main research questions along with four sub-questions. These were:

1. What are the effects of *formative assessment frequency* on student achievement (represented by student academic change-score or, *AC-Score)* for each subject?
2. Do the effects of the *formative assessment frequency* differ based on *formative assessment cycle-length (short-cycle vs. long-cycle assessments)*?
3. Do the effects of the *formative assessment frequency* differ for students from different student subgroups (*genders, races, EDS, LEP, and SWD)*?
4. Do the effects of the *formative assessment frequency* differ based on school-level at-risk characteristics (%EDS, %Minority)?
5. Do the effects of *formative assessment frequency* differ based on *school-level formative assessment-cycle length* characteristics (*Mean short-cycle assessments, Mean long-cycle assessments*)?

Sample

The sample for this study was drawn from the population of sixth, seventh, and eighth grade public school students in North Carolina from 2010-11. Of this population, 83,799 students who took the mathematics and reading end-of-grade assessments in 2010-11 and took mathematics and/or reading assessments using an OFAP were included.

Limitations

This study had multiple limiting factors. One major limitation of this study was the lack of detailed information regarding the formative assessment administrations. Although the total number of OFAP assessments by subject and assessment cycle-length was included, no information was provided indicating the date and time of the assessment. It would have been helpful to have assessment information regarding the date/time of assessment in order to measure the relative frequency of assessment (i.e. if there were periods that the OFAP was used heavily, or if the OFAP assessment administrations were evenly distributed throughout the year). Another limitation of this study was the lack of detailed information regarding potentially important classroom characteristics such as instructor teaching experience and number of students in each given classroom. These are factors that could potentially moderate the effect of formative assessment on student achievement. In addition, the fact that the schools included in this study were self-selected, in that the OFAP is a commercial product which the schools/districts pay to use, was considered a limitation of this study.

Procedures

A multi-level model was constructed for each content area (mathematics and reading) in order to answer the previously stated research questions. In building each model, student-level variables and interaction effects were added first and tested for statistical significance.

Any statistically significant variables or interactions were retained in the given model. Control variables including *gender, race* (African American, Asian, Hispanic, Multi-racial, Other, and Caucasian)*,* and *at-risk student* variables (EDS, LEP, and SWD) were also included in the initial analysis of each model. Any control variables found to be statistically significant predictors of AC-score for the given content area were also retained in the model. Non-significant control variables were omitted moving forward in the model building process. Once the student-level models were deemed satisfactory, school-level variables and interaction effects were added to the model and tested for statistical significance. Any statistically significant school-level variables or interactions were retained in the final model. The student-level and school-level models were then compared in order to determine which model provided a better fit to the data. The best fit model for each content area was used for final interpretation, conclusions, and recommendations.

Findings

This section provides detailed results specific to the analysis of each research question. The results are organized by research question and content area. Each section provides a brief summary of the research question which the given analysis sought to address. Results for the mathematics analysis are presented first, followed by the results from the reading analysis in each section.

The major findings of this study suggest that formative assessments are positively related to student achievement in reading and mathematics. Results suggest that, short-cycle reading formative assessments, in particular, result in positive gains for students in reading. Both student and school-level short-cycle reading formative assessment frequency were suggested to have a positive effect on student achievement in reading.

The results from this study also suggest that long-cycle mathematics formative assessments may result in positive gains for students. The interaction between student and school-level long-cycle mathematics assessment frequency suggested that students attending schools that administer a greater number of long-cycle mathematics formative assessments, who take a greater number of mathematics formative assessments experience positive gains in mathematics achievement. In addition, short-cycle mathematics formative assessments seem to have a positive effect on EDS student achievement. Table 5.1 provides a summary of the major findings for each subject.

Table 5.1 – *Summary of Findings – Outcome: Achievement (AC- Score)*

	Reading		**Mathematics**	
Independent Variable	***Coefficient Estimate***	***p-value***	***Coefficient Estimate***	***p-value***
Research Question 1				
Formative Assessment Frequency	0.0052	*p* < .0001	-	ns
Research Question 2				
Short Cycle Assessment (SCA)	0.0048	*p* = .0060	-	ns
Long Cycle Assessment (LCA)	-	ns	-0.0331	*p* = .0083
Research Question 3				
Economically Disadv. Student (EDS)	-0.0146	*p* < .0001	-0.0416	*p* < .0001
Limited English Proficiency (LEP)	-0.0229	*p* = .0099	0.0285	*p* = .0054
Student With Disabilities (SWD)	-0.0118	*p* = .0327	-0.0257	*p* = .0002
SCA x EDS	-	ns	0.0019	*p* = .0130
SCA x SWD	-0.0070	*p* = .0315	-	ns
Research Question 4				
%EDS	-0.1582	*p* < .0001	-0.1386	*p* = .0007
%Minority	0.0459	*p* = .0327	-	ns
Research Question 5				
SCA School Mean	0.0063	*p* = .0327	-	ns
LCA School Mean	-	ns	-0.0257	*p* = .0028
LCA x LCA School Mean	-	ns	0.0186	*p* = .0002

Note: All results taken from Final Full Contextual Models (Tables 4.13 & 4.14) with the exception of Formative Assessment Frequency; ns = not statistically significant; SCA School Mean and LCA School Mean predict the AC-Score intercept for the given subject;

Results for Research Question 1

Research question one sought to investigate the effectiveness of formative assessment frequency on student achievement.

Mathematics results. The initial mathematics model analysis suggested that mathematics formative assessment frequency is not a statistically significant ($p = 0.5319$) predictor of student mathematics achievement. However, the random effects estimate for mathematics formative assessment frequency was determined to be statistically significantly different from zero ($p <.0001$), indicating that the relationship between mathematics formative assessment frequency and student mathematics achievement (mathematics AC-Score) varies across schools and, therefore, the addition of level-2 variables may reveal such a relationship.

Reading results. The initial reading model analysis suggested that reading formative assessment frequency is a statistically significant ($p < .0001$) predictor of student reading achievement. The fixed effect estimate of 0.004 for reading formative assessment frequency suggested that the more reading formative assessments that a student takes, the greater gains he/she will show in terms of reading achievement. As was seen with the mathematics analysis, the random effects estimate for reading formative assessment frequency was also found to be statistically significantly different from zero, indicating that the relationship between reading formative assessment frequency and student reading achievement (reading AC-Score) varies across schools. In other words, this effect is likely to vary from school to school depending upon school-level characteristics. Therefore, the addition of level-2 variables may help to further explain this relationship.

Results for Research Question 2

Research question two sought to investigate if the relationship between formative assessment frequency and student achievement varied depending on the cycle-length of the formative assessments. Instead of using the total formative assessment frequency variable as was done in addressing research question one, research question two tested two student-level formative assessment frequency variables – one representing total number of short-cycle assessments (SCAs) taken and the other representing the total number of long-cycle assessments (LCAs). Because the dataset only included SCAs and LCAs, the sum of these assessment counts was equal to the total number of assessments variable used in answering research question one.

Mathematics results. The analysis of the formative assessment frequency by cycle-length mathematics model suggested that, similar to the findings in research question1, formative assessment frequency was not a statistically significant predictor of mathematics AC-Score regardless of assessment cycle-length. Also in line with what was found in the analysis addressing research question 1, the random effects estimates for both mathematics SCAs and LCAs were statistically significantly different from zero, suggesting that the addition of school-level variables could be helpful in explaining the relationship between these variables and student mathematics achievement.

Although the model used to address research question 1 produced similar results as the model addressing research question 2, comparing the model fit estimates (-2LL, AIC, and BIC) for each model, the formative assessment cycle-length specific model was determined to provide a significantly better fit for the data. Based on this finding, the cycle-length specific mathematics model was used as the basis for the remaining analyses. In addition, even though the formative assessment frequency variables were not found to be statistically significant predictors of student mathematics AC-score, they were retained in order to allow for the testing of interactions in the subsequent analyses.

Mathematics control variable results. It should be noted here that the analysis of this model revealed that the coefficient estimates for each of the control variables were determined to be statistically significantly different ($p < .01$) from zero. The coefficient estimate for the gender predictor (0.0477) suggested that female students achieve higher mathematics AC-Scores as compared to their male counterparts. The results also suggested that EDS and SWD students achieve lower mathematics AC-Scores as compared to non-EDS and non-SWD students. This was not unexpected. However, the results from this analysis suggested that LEP students achieve higher mathematics AC-Scores as compared to non-LEP students. This outcome aligns with the results suggesting that all ethnic groups included as control variables achieve higher mathematics AC-Score as compared to Caucasian students seems to align with the LEP results. It is important to keep in mind that the AC-score metric is a way of representing growth in achievement and, as such, these unexpected results could be illustrating the gap that historically exists between at-risk students and minority students in comparison to Caucasian students. Where there is a gap, there is also more room to grow, which could potentially explain why these results suggest that student subgroups which typically achieve lower scores seem to show greater gains.

Reading results. The analysis of the formative assessment cycle-length specific reading model indicated that short-cycle reading formative assessments were statistically significant predictors ($p = .0059$) of reading AC-Score, whereas long-cycle reading formative assessments were not ($p = .1039$). The fixed effect of 0.0861 suggested that reading SCAs may have a positive effect on student reading AC-score. The random effect for reading SCAs suggested that the relationship between reading SCAs and reading AC-score had statistically significant variance across schools. Like the fixed effect for reading LCAs, the random effect estimate for reading LCAs was also not statistically significantly different from 0. Based on these findings, the reading LCAs variable was omitted from the model.

Although the total formative assessment frequency model (regardless of cycle-length) produced similar results as the cycle-specific reading model, the latter was determined to provide a better fit based on a comparison of the model fit statistics (-2LL, AIC, BIC) between the two models. Therefore, the cycle-length-specific reading model was used in addressing the remaining research questions.

Reading control variable results. Unlike what was seen with the mathematics analysis, the initial reading analysis revealed that not all of the control variables were statistically significant predictors of reading AC-Score. Gender was not determined to be a statistically significant predictor of reading AC-Score. Whereas all ethnic groups included as control variables were estimated to achieve higher mathematics AC-Scores as compared to Caucasian students, the reading analysis suggest that only Asian and Hispanic students achieve higher reading AC-Scores in comparison to Caucasian students. African American students were estimated to achieve lower AC-Scores than Caucasian students. Students in the ethnic category “Other” (i.e. students who were not African American, Asian, Caucasian, Hispanic, or Multiracial) were not found to be statistically significantly different from Caucasian students in terms of AC-Score. All at-risk student predictors in the reading analysis were found to be statistically significantly different from 0 and negative, suggesting that at-risk students achieve lower reading AC-Scores as compared to non-at-risk students.

Results for Research Question 3

Research question three sought to investigate whether or not the relationship between formative assessment frequency and student achievement varied for at-risk students. The three at-risk categories included in the model were: *economically disadvantaged students* (EDS)*, students with limited English proficiency* (LEP)*, and students with disabilities* (SWD). Interactions between each of these variables and the formative assessment frequency variable(s) for the given subject were tested for statistical significance.

Mathematics results. The results from the mathematics analysis addressing research question three revealed that, although mathematics SCAs alone are not a statistically significant predictor of student mathematics achievement in general, they are a statistically significant predictor (p = .0097) of mathematics achievement for EDS students. The coefficient estimate of 0.002 for the EDS and mathematics SCA frequency interaction suggested that mathematics SCAs have a positive effect on mathematics AC-Score for EDS students. Interactions between mathematics SCA frequency and each of the other at-risk student variables (SWD and LEP) were tested for statistical significance but none was found. The same was done for each of the at-risk student variables and mathematics LCAs, but none of these interactions were determined to have statistical significance.

Reading results. The results from the reading analysis addressing research question three revealed that, although reading SCA frequency is a statistically significant, positive predictor of student reading achievement in general, reading SCAs may have a negative effect for SWD students. The interaction coefficient estimate between reading SCA frequency and SWD (-0.0071) suggested that taking reading SCAs may lower reading achievement for SWD students. This relationship was unexpected. However, it is possible that schools may have administered online formative assessments to already struggling SWD students as a last ditch intervention. If this were the case, then the SCA administrations may not necessarily have been the cause of the lower reading AC-Scores displayed by the SWD students, but rather, the fact that SCAs were administered may serve as an indicator of which SWD students were already struggling in reading. All other interactions between reading SCA frequency and at-risk student predictors (LEP and EDS) were not determined to be statistically significantly different from 0.

Results for Research Question 4

Research question four sought to investigate whether or not the relationship between formative assessment frequency and student achievement varied depending on school-level demographic characteristics. The school-level demographic variables added to the model were: *percentage of EDS* (%EDS) and *percentage of minority students* (%Minority) for the given school. Interactions between each of these variables and the formative assessment frequency variable(s) for the given subject were tested for statistical significance.

Mathematics results. The addition of the school-level demographic variables to the mathematics model illustrated the importance of school context in this analysis. The %EDS coefficient estimate of -.1386 was statistically significant (*p* = .0007), suggesting that students at schools with a high percentage of EDS students produce lower mathematics AC-Scores as compared to students at schools with a lower percentage of EDS students. On the other hand, %Minority was not found to be a statistically significant predictor of student mathematics AC-Score. Interactions between both SCA and LCA mathematics frequency and %EDS and %Minority were tested for statistical significance but none was found. As a result, %EDS was the only school-level demographic variable retained in the model.

Reading results. The addition of the school-level demographic variables to the reading model suggested that both %EDS and %Minority are statistically significant predictors of reading AC-Score. As was seen in the mathematics model analysis, %EDS was a statistically significant predictor of reading AC-Score, producing a coefficient estimate (-0.1582) suggesting that schools with a greater percentage of EDS students show lesser gains as compared to schools with lower %EDS. On the other hand, the coefficient estimate for %Minority (0.0459) suggested that schools with greater minority populations displayed greater gains. Interactions between SCA mathematics frequency and both %EDS and %Minority were tested for statistical significance but none was found. Both %EDS and %Minority were retained in the reading model.

Results for Research Question 5

Research question 5 sought to investigate whether or not the relationship between formative assessment frequency and student achievement varied depending on school-level formative assessment frequency mean. The variable(s) for school-level formative assessments frequency mean were added. Interactions between the variable(s) and the student-level formative assessment frequency variable(s) were tested for statistical significance.

Mathematics results. The analysis of the mathematics model with the addition of the school-level formative assessment frequency mean variables produced some interesting findings. School-level mathematics SCA frequency mean was not found to be a statistically significant predictor of student mathematics AC-Score. However, school-level mathematics LCA frequency mean was determined to be statistically significantly different from 0 (p = .0028) with a coefficient estimate of -0.0257, suggesting that students at schools that give a higher number of mathematics LCAs achieve lower mathematics AC-Scores on average. In addition, the student-level mathematics LCA frequency variable that was not statistically significant in the previous models was now significant (*p* = .0083). The coefficient estimate of -0.0331 also suggested that students who take a greater number of mathematics LCAs achieve lower mathematics AC-Scores. This was an unexpected outcome as well. Although this was not particularly anticipated, given that the random effect for mathematics LCAs was statistically significant in previous models, thus supporting the decision to retain the LCA variable in the model on the chance that the addition of a school-level variable may explain the relationship between student-level LCA frequency and mathematics AC-Score, this outcome should not be surprising. Here, the addition of the school-level mathematics LCA frequency mean was able to explain the relationship between student-level LCA frequency and mathematics AC-Score which was otherwise indiscernible. It appears that the context of the school's level of use of mathematics LCAs matters for the student-level LCA frequency variable. The analysis also revealed that the interaction between school-level mathematics LCA frequency mean and student-level mathematics LCA frequency was also determined to be statistically significantly different from 0, however, the coefficient estimate for the interaction between these variables suggested that students who take a greater number of mathematics LCAs in schools that administer a greater number of mathematics tend to achieve higher mathematics AC-Scores. Given that student and school-level mathematics LCAs each have negative effects on student mathematics achievement, there is a certain threshold which must be met before this interaction begins to show net gains in mathematics AC-Score as compared to students who took no mathematics LCAs in a school which administered no mathematics LCAs. Figure 4.1 in the previous chapter illustrates this relationship. As is shown in Figure 4.1, students attending a school that administers approximately three mathematics LCAs per year, who take four or five mathematics LCAs in the same year begin to surpass the gains of a similar student who took no mathematics LCAs.

Reading results. The results from analysis of the reading model with the addition of the school-level reading SCA frequency mean suggested that school-level reading SCA mean is a statistically significant predictor ($p = 0.0365$) of student reading achievement. The coefficient estimate (0.0063) for reading SCA frequency mean suggests that students at schools administering a higher number of reading SCAs achieve higher reading AC-Scores as compared to students who attend schools with a lower reading SCA frequency mean. This outcome was not entirely unexpected given that the reading model to this point suggested that student-level reading SCA frequency has a similar positive effect (0.005) on student reading AC-Score. The interaction between school-level and student-level reading SCA frequency was tested and found to be a statistically non-significant predictor of student reading AC-Score. Figure 4.2 in the previous chapter illustrates the relationship between both student and school-level reading SCA frequency and student reading AC-Score for a baseline student (white, male, non-EDS, non-SWD, and non-LEP student at school with 0% EDS) in this study.

Conclusions

This section provides a summary of the main conclusions from this study. Observations and hypotheses for the relationships evidenced throughout this analysis are presented first for mathematics and then for reading.

For mathematics, this study found statistically significant effects for both student and school-level mathematics long-cycle formative assessments on student gains in mathematics achievement. Whereas each of these predictors alone suggested negative effects, the interaction between these variables suggested that long-cycle mathematics formative assessments may have positive effects on student gains in mathematics. The results suggested that students at a school that administers a greater number of long-cycle mathematics formative assessments, and who take a greater number of long-cycle mathematics formative assessments show greater gains on mathematics end-of-grade assessments. Given these results, it seems to be that, whereas the assessments alone do not improve student mathematics achievement, a strong commitment to student mathematics achievement at both the student and school-level produces positive gains for the students. These findings align with the assertions made by Stiggins and DuFour (2009) suggesting that school-level assessments are an imperative part of evaluating the current curriculum and instructional practices in the school. The results from the mathematics analysis, however, did not concur with the existing literature that suggests that long-cycle assessments aren't likely to have much of an effect on student achievement (Cowie & Bell, 1999; Looney, 2005; Shephard, 2007; Wiliam, 2010).

In addition, although short-cycle mathematics assessments were not statistically significant predictors of student mathematics achievement in general, the results suggested that mathematics short-cycle formative assessment frequency is a positive predictor of gains in mathematics for economically disadvantaged students (EDS). This finding suggests that more frequent (short-cycle) mathematics formative assessments may be particularly helpful for EDS students - a population which often struggles in academics. This finding is consistent with the current cycle-length literature (Cowie & Bell, 1999; Looney, 2005; Shephard, 2007; Wiliam, 2010), the current literature regarding formative assessments and economically disadvantaged students (Burns et al, 1991; Meisels et al, 2003), and the current literature on formative assessment frequency (Martinez & Martinez, 1992; Peterson & Siadat, 2009).

For reading, this study found statistically significant effects for both student and school-level short-cycle reading assessments. Unlike the results from the mathematics analysis, the results from the reading analysis suggested that a greater number of short-cycle reading formative assessments, both at the student and school-level, is likely to produce positive gains for student achievement on end-of-grade reading assessments, supporting the current literature regarding assessment cycle-length (Cowie & Bell, 1999; Looney, 2005; Shephard, 2007; Wiliam, 2010), formative assessment frequency (Martinez & Martinez, 1992; Peterson & Siadat, 2009) and school-level assessment (Stiggins & DuFour, 2009). Whereas the importance of both student and school-level commitment to formative assessments was more apparent in the mathematics analysis, the results from the reading analysis also suggest that a greater gain can result from a greater commitment by both the student and the school.

Implications

The results from this study suggest that in order to increase student achievement on mathematics summative assessments there must be a strong commitment to formative assessment by both the student and the school he or she is attending. It appears that schools, in which there is a culture of commitment to formative assessment for all students and not just as an intervention or "quick fix" for struggling students, are more likely to see positive gains in student summative assessment achievement for mathematics.

In addition, it also appears that more frequent formative assessment may be particularly important for producing gains in mathematics for economically disadvantaged students. This could be particularly relevant given that the analysis suggests that economically disadvantaged students typically show lesser gains in mathematics as compared to non-EDS students.

Results from the reading analysis suggest that short-cycle reading formative assessments have the potential to increase student gains on reading summative assessments regardless of the level of commitment to formative assessment at the school-level. However, although positive gains may be had without a strong culture of formative assessment at the school-level, the results suggest that the presence of such a commitment would only serve to further increase student gains. Use of short-cycle reading formative assessments could be particularly pertinent for at-risk students (EDS, LEP, and SWD) as students in each of these at-risk subgroups typically produce lesser gains as compared to non at-risk students.

Recommendations

The following recommendations are offered for related research in the field of education – specifically regarding formative assessment in the areas of mathematics and reading.

1. Given the increasing use of formative assessment in the classroom, a series of longitudinal studies, based on the models used in this study, would provide the opportunity to evidence long-term trends.
2. As technology becomes integrated into classroom-level instruction and assessment, tools designed to increase the ability of educators to formatively assess students will undoubtedly become more prevalent. Examination of the differences between the

various tools available would likely provide results which would be useful to schools and districts in deciding upon technological tools to fund in their classrooms.

3. Whereas this study examined the effects of formative assessment for only mathematics and reading, it would likely be beneficial to explore the effects of formative assessment among other subject areas. Based on what was found in this study, the possibility that the effects of formative assessment vary from subject to subject seems likely.
4. Whereas this study focused only on students in grades six through eight, exploration of the effects of formative assessment on student achievement in other grades, both lower and higher than the span used for this study, would provide additional context for the current findings.

The following recommendations are offered for practitioners in the field of education – particularly in reference to the areas of mathematics, reading, and formative assessment.

1. The findings of this study support the use of short-cycle formative mathematics assessments for economically disadvantaged students.
2. The results from this study also support the use of long-cycle formative mathematics assessments but with one caveat – there must be a strong school-wide commitment to these assessments.
3. The use of short-cycle formative reading assessments is supported by the findings in this study. Positive outcomes are suggested for all students and student subgroups.
4. A strong school-level commitment to short-cycle formative reading assessments is also supported by the findings in this study.

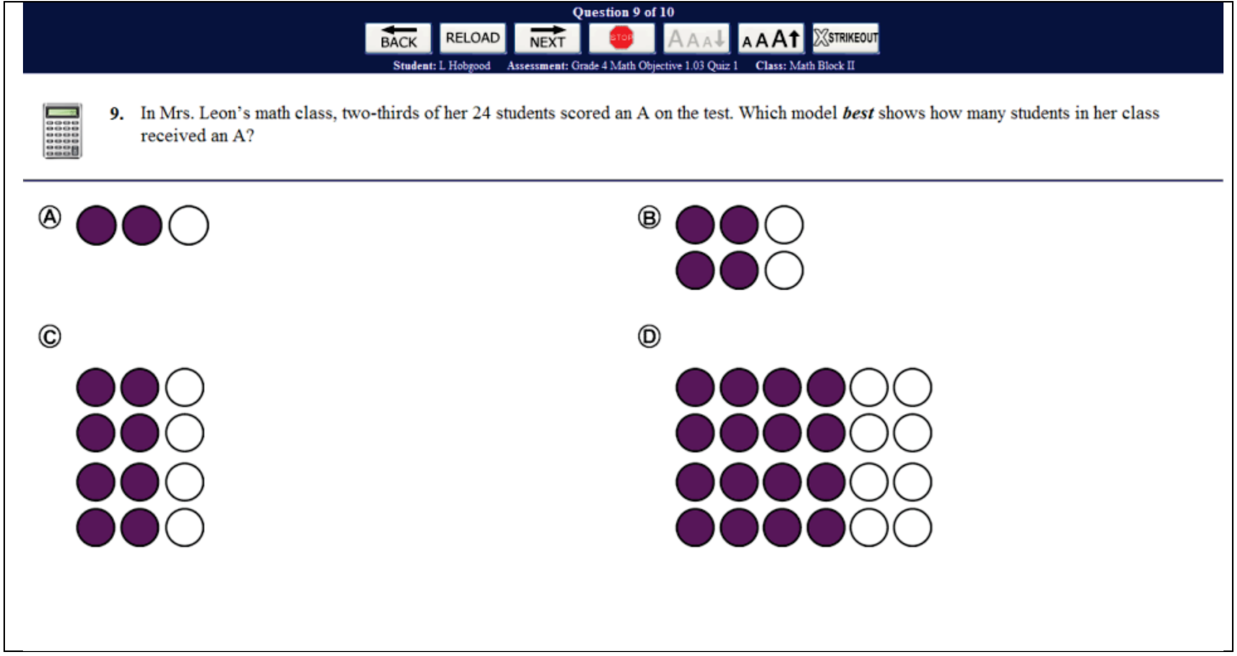
Considerations for Student Achievement

The results from this study bring to light a few important considerations regarding formative assessment and student achievement. It appears that formative assessments have the potential to increase student performance on summative assessments. Whereas increasing summative assessment scores is not the end-all, be-all for improving student learning, summative assessments continue to be the best current accountability method for upholding educational standards on large-scale basis. It seems, however, that with the potential to increase student achievement and assist in classroom instruction, all while working within the current accountability model, a comprehensive assessment system which incorporates formative assessment into the accountability model may be beneficial. An assessment system in which students are not simply measured by one sample of their ability after a semester or school-year of instruction has passed but, instead, a more dynamic system in which students are measured on a more regular basis, providing more opportunities for remediation closer to the time of initial instruction as opposed to later on down the road. The potential for formative assessments to increase student achievement is considerable and, with the assistance of the boom in educational technology, this endeavor seems to be increasingly more attainable.

APPENDIX A

This appendix provides Mathematics and Reading item examples from the Online Formative Assessment Program used for this study as well as a sample report.

Figure A1 – *Mathematics OFAP Item Sample*

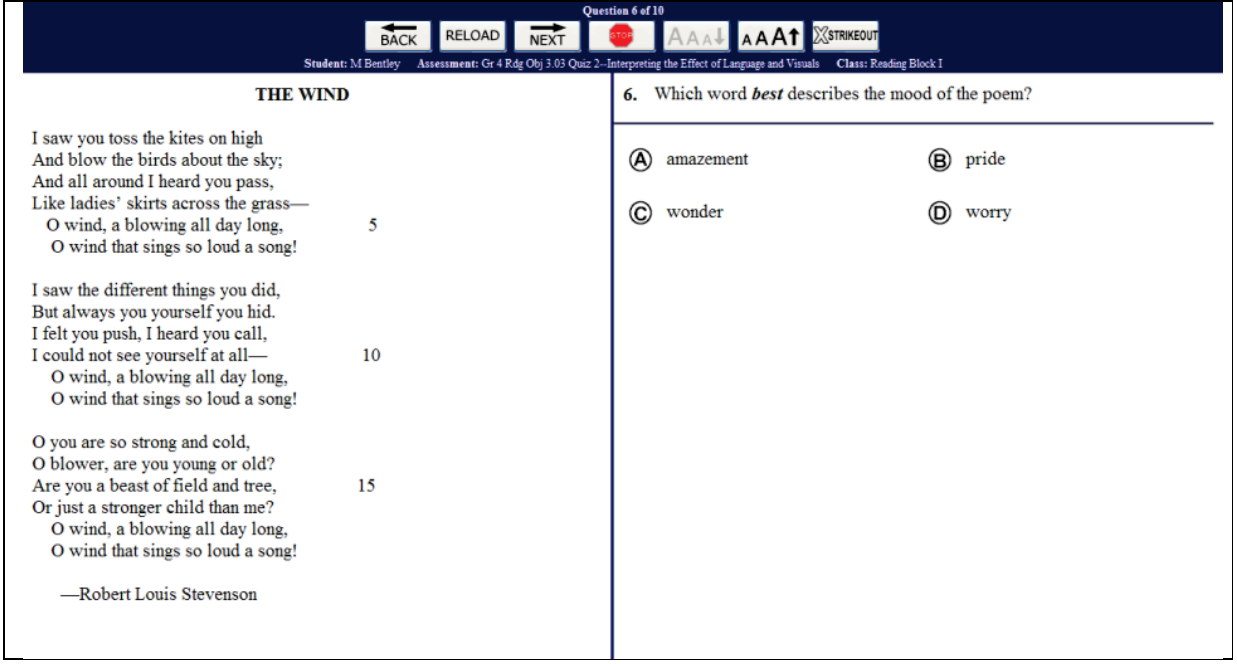


Question 9 of 10

Student: L Hobgood **Assessment:** Grade 4 Math Objective 1.03 Quiz 1 **Class:** Math Block II

9. In Mrs. Leon's math class, two-thirds of her 24 students scored an A on the test. Which model ***best*** shows how many students in her class received an A?

Figure A2 – *Reading OFAP Item Sample*



Student: M Bentley   Assessment: Gr 4 Rdg Obj 3.03 Quiz 2--Interpreting the Effect of Language and Visuals   Class: Reading Block I

THE WIND

I saw you toss the kites on high
And blow the birds about the sky;
And all around I heard you pass,
Like ladies' skirts across the grass—
O wind, a blowing all day long,
O wind that sings so loud a song!

I saw the different things you did,
But always you yourself you hid.
I felt you push, I heard you call,
I could not see yourself at all—
O wind, a blowing all day long,
O wind that sings so loud a song!

O you are so strong and cold,
O blower, are you young or old?
Are you a beast of field and tree,
Or just a stronger child than me?
O wind, a blowing all day long,
O wind that sings so loud a song!

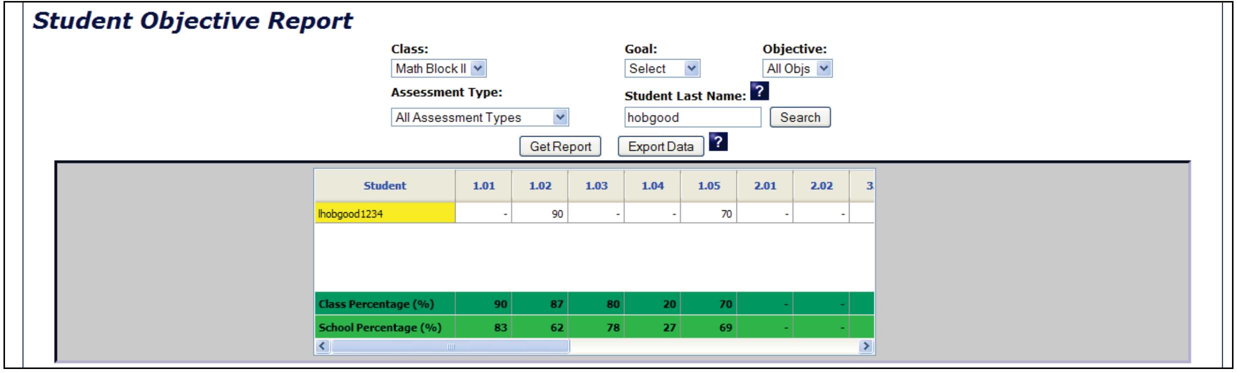
—Robert Louis Stevenson

6. Which word ***best*** describes the mood of the poem?

Ⓐ amazement Ⓑ pride

Ⓒ wonder Ⓓ worry

Figure A3 – *OFAP Student Objective Report Sample*



Student Objective Report

Class: Math Block II

Goal: Select

Objective: All Objs

Assessment Type: All Assessment Types

Student Last Name: hobgood

Student	1.01	1.02	1.03	1.04	1.05	2.01	2.02	3
lhobgood1234	-	90	-	-	70	-	-	
Class Percentage (%)	90	87	80	20	70	-	-	
School Percentage (%)	83	62	78	27	69	-	-	

References

Bergan, J. R., Sladeczek, I. E., Schwartz, R. D., & Smith, A. N. (1991). Effects of a measurement and planning system on kindergarteners' cognitive development and educational planning. *American Educational Research Journal, 28*(3), 683-714.

Bill & Melinda Gates Foundation. (2010). *What's next? The assessment challenges facing states.* Retrieved September 4, 2012, from http://www.gatesfoundation.org/

Black, P. & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Brynes, D. & Yamamoto, K. (1984). Academic retention: An inside look. Unpublished Paper. Logan, UT: Utah State University.

Burns, M. K., Klingbeil, D. A., & Ysseldyke, J. (2010). The effects of technology-based formative evaluation on student performance on state accountability math tests. *Psychology in the Schools, 47*(6), 582-591.

Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics, and challenges. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 3-17). New York: Routledge.

Common core states initiative. (2011). Retrieved from http://www.corestandards.org

Cowie, B., & Bell, B. (1999). A model of formative assessment in science education. *Assessments in Education: Principles, Policy, and Practice, 6*(1), 32-42.

Firestone, W., Schorr, R., & Monfils, L. (Eds.). (2004). *The ambiguity of teaching to the test: Standards, assessment, and educational reform*. Mahwah, NJ: Lawrence Erlbaum.

Foster, D. & Poppers, A. (2009). *Using formative assessment to drive learning*. The silicon valley mathematics initiative: A twelve-year research and development project. Retrieved October 25, 2012, from http://www.svmimac.org/

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

Furtak, E. M. & Ruiz-Primo, M. A. (2005). Questioning cycle: Making students' thinking explicit during scientific inquiry. *Science Scope,* 22-25.

Grunwald, H. & Peterson, M. W. (2004). Factors that promote faculty involvement in and satisfaction with institutional and classroom student assessment. *Research in Higher Education, 44*(2), 173-204.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Kingston, N. & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28-37.

Looney, J. (Ed.). (2005). *Formative assessment: Improving learning in secondary classrooms*. Paris: Organisation for Economic Cooperation and Development.

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks: Sage.

Martinez, J. G. R. & Martinez, N. C. (1992). Re-examining repeated testing and teacher effects in a remedial mathematics course. *British Journal of Educational Psychology, 62*, 356-363.

McCoach, D. B. (2010). Hierarchical linear modeling. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 123-140). New York: Routledge.

Miesels, S. J, Atkins-Burnett, S., Xue, Y., Nicholson, J., Bickel, D. D., Son, S. (2003). Creating a system of accountability: The impact of instructional assessment on elementary children's achievement test scores. *Education Policy Analysis Archives, 11*(9), 1-18.

Monfils, L., Firestone, W., Hicks, J., Martinez, M., Schorr, R., & Camilli, G. (2004). Teaching to the test. In W. A. Firestone, R. Y. Schorr , & Monfils, L. (Eds.), *The ambiguity of teaching to the test: standards, assessment, and educational reform*. (pp. 37-63). New York: Routledge.

National Education Association. (2012). *Students affected by achievement gaps*. Retrieved September 4, 2012, from http://www.nea.org/home/20380.htm

Niklason, L.B. (1984). Nonpromotion: A pseudo solution. *Psychology in the Schools, 21*,485-499.

No Child Left Behind Act, 20 U.S.C. 6301. (2002)

North Carolina Department of Public Instruction. (2011). *Guide to Career and Technical Education's Special Populations – Challenge Handbook*. Retrieved October 16, 2012, from http://www.ncpublicschools.org/docs/cte/related-services/support/special-populations/challenge-handbook.pdf

North Carolina Department of Public Instruction\Accountability Services. (2011). *The ABCs of Public Education Academic Change for Schools 2010-11*. Retrieved October 23, 2011, from http://www.ncpublicschools.org/docs/accountability/reporting/abc/2010-11/academicchange.pdf

Organisation for Economic Co-Operation and Development [OECD]. (2005). *Formative assessment - Improving learning in secondary classrooms*. Paris: Centre for Educational Research and Innovation, OECD.

Perie, M., Marion, S. & Gong, B. (2009) Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28*(3), 5-13.

Peterson E. & Vali Siadat, M. (2009). Combination of formative and summative assessment instruments in elementary algebra classes: A prescription for success. *Journal of Applied Research in the Community College. 16*(2), 92-102.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioural Science, 28*(1), 4-13.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: searching for instructional sensitivity. *Journal of Research in Science Teaching. 39*(5), 369-393.

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

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P.R., Furtak, E. M., Ruiz-Primo, M. A. (2008). On the impact of curriculum-embedded formative assessment on learning:

A collaboration between curriculum and assessment developers. *Applied Measurement in Education, 21*, 295-314.

Shephard, L. A. (2007). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of formative assessment: Shaping teaching and learning* (pp. 279-303). Mahwah, NJ: Erlbaum.

Singer, J. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24*(4), 323-355.

Stiggins, R. & DuFour, R. (2009). Maximizing the power of formative assessments. *Phi Delta Kappan, 90*(9). 640-644.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. (2009). *The nation's report card: Long-term trend 2008*. Washington, DC: U.S. Government Printing Office.

U.S. Department of Education. (2009). *Race to the top program executive summary*. Washington, DC: U.S. Government Printing Office.

White, B. Y. & Frederiksen, J. R. (1998). *How children think and learn: The social contexts of cognitive development*. Oxford, UK: Blackwell.

Wiliam, D. Lee, C., Harrison, C., & Black, P. J. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education, 11*(1), 49-64.

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of formative assessment: Shaping teaching and learning*. (pp. 53-82). Mahwah, NJ: Erlbaum.

Wiliam, D.. (2010). An integrative summary of the research literature and implications for a new theory of formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment*. (pp. 18-40). New York: Routledge.