6-2013

The Nature and Predictive Validity of a Benchmark Assessment Program in an American Indian School District

Beverly R. Payne
University of Nebraska-Lincoln, beverlypayne05@gmail.com

THE NATURE AND PREDICTIVE VALIDITY OF A BENCHMARK ASSESSMENT PROGRAM IN AN AMERICAN INDIAN SCHOOL DISTRICT

By

Beverly J. R. Payne

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Fulfillment of Requirements

For the Degree of Doctor of Education

Major: Educational Administration

Under the Supervision of Professor Jody Isernhagen

Lincoln, Nebraska

June, 2013

THE NATURE AND PREDICTIVE VALIDITY OF A BENCHMARK ASSESSMENT PROGRAM IN AN AMERICAN INDIAN SCHOOL DISTRICT

Beverly J. R. Payne, Ed.D.

University of Nebraska, 2013

Adviser: Jody Isernhagen

This mixed methods study explored the nature of a benchmark assessment program and how well the benchmark assessments predicted End-of-Grade (EOG) and End-of-Course (EOC) test scores in an American Indian school district. Five major themes were identified and used to develop a Dimensions of Benchmark Assessment Program Effectiveness model: Professional Development, Assessment Literacy, Data Literacy, Instructional Practice, and Program Effectiveness. The study found that Professional Development, Data Literacy, and overall Program Effectiveness were strengths of the district's benchmark assessment program. Assessment Literacy and Instructional Practice were found to be weaker areas of the district's program. Benchmark assessment scores correlated strongly with the EOG and EOC scores except in two areas. Benchmark assessment scores predicted EOG and EOC scores well.

Dedication

I want to thank my wonderful husband, Dennis Jeffrey Payne, for his love and support during this project. His patience with me while I pursued this dream was nothing short of remarkable. I appreciate the sacrifices that he made in order to help me achieve my goals.

Acknowledgement

Without the tireless support of my advisor, Jody Isernhagen, this project would not have been possible. I appreciate the hours she spent reading draft after draft, providing me with feedback, and always keeping me on course.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

Introduction

Lauren waited patiently in a hallway with the rest of her classmates and a proctor. Inside a nearby classroom, a teacher, who on that particular morning served as a test administrator for the End-of-Course assessment that Lauren and her classmates would complete, scurried from computer to computer, readying each for a particular student in the class. The teacher began to call each student into the classroom individually, indicating which computer would be theirs for the test. Lauren sat down in front of her monitor, feeling confident. Earlier the principal provided each student with a nutritious snack to fuel their brains. Though most students would be done in a couple of hours, the assessment could take as long as four hours. Lauren was also positive this morning because she knew she had been well-prepared for this high-stakes test. Her teacher did well in preparing her class in the content that would be covered as well as providing practice with taking assessments in an online environment. In addition, Lauren's teacher and principal have required that students participate in a benchmark assessment program during the course that provided them with both summative and formative data. Each student completed two or three benchmark assessments which provided a map of their progress, but also supplied their teacher with data about their strengths and weaknesses relevant to the content of the course. With this data, the teacher provided instructional interventions to students, depending upon their individual needs. The same program used to deliver the benchmark assessments was also available to the teacher at other times, so she could re-assess in a quick and timely manner.

Statement of the Problem

Since schools and districts can be labeled as 'failing' under the adequate yearly progress (AYP) requirement of the federal No Child Left Behind act (USED, 2001), or NCLB, finding better and quicker ways of determining student progress toward the curricular goals has received significant focus and attention from administrators. Benchmark assessment programs have been initiated in numerous districts to provide schools with data that is timely, actionable, and that can be used to predict later performance on the high-stakes tests. Some districts develop their own benchmark assessments, using teachers and in-house curriculum experts to write test items. Other districts look to vendors for a commercial program. In some cases, districts are able to use state developed or state sponsored benchmark assessment programs, such as the one that Lauren and her classmates used. Whether a home-grown or commercial product, benchmark assessment programs can be expensive, and districts need reassurance that the funds are being well-spent. Often districts that develop their own items and tests lack the expertise to conduct reliability and validity studies, relying on anecdotal evidence as to the success of their efforts.

Districts that purchase commercial products must depend on research conducted by the vendor, rather than independently conducted studies. Many times the populations of the studies do not match the population of the school purchasing the product, leading to false expectations for the districts. For example, few studies have been conducted with minority populations such as American Indian students. Of the existing studies, many involve research on screening instruments used as part of Response to

Intervention/Instruction (RTI) programs (Atkins & Cummings, 2011; Barger, 2003; Graney, Missal, Martinez, & Bergstrom, 2009; Hintze & Silberglitt, 2005; Nese, Park, Alonzo, & Tindal, 2011; Pearce & Gayle, 2009; Petscher & Young-Suk, 2011; Wood, 2006; Wright, 2010). Educators and researchers (Brown & Coughlin, 2007; Bulkley, Nabors Oláh, & Blanc, 2010) agreed that many aspects of benchmark assessment programs have not been well-researched. However, the lack of reliable research has not prevented schools and districts from purchasing the assessment products. When educators asked for instruments and systems to provide them with student level data prior to the high-stakes state test, the assessment industry responded quickly. Sales in the industry have consistently increased, and the volume of products sold in 2006 was approximately twice the volume sold in 2000 (Burch, 2010). Recently, Pearson (2012), a global education company, reported, "We delivered 13 million secure online tests in 2011" (p. 8), and the company was also awarded contracts from PARCC (Assessment of Readiness for College and Careers) and Smarter Balanced (Smarter Balanced Assessment Consortium) to assist with states transitioning to the Common Core Standards and the accompanying online assessments. Despite the current gloomy economic outlook for state departments of education and local school districts, companies involved in educational assessment are still strong.

Educators need solid information when making such decisions about assessment programs, especially in times of economic austerity. A study on the development of a benchmark program and its predictive nature would provide direction for administrators as they grapple with these decisions. Additionally, schools and districts serving large

populations of minority students would benefit from studies conducted on similar populations. Since many of the studies involve RTI screening measures, a study involving older students and different measures would also add to the literature. Through such a study, educators would be able to identify criteria for the development of a benchmark assessment program as well as better understand how well a specific type of benchmark assessment might predict later student performance on a state End-of-Grade or End-of-Course assessment.

Purpose of the Study

The purpose of this study is to explore the history and nature of a benchmark assessment program and how the assessments are used in relationship to high stakes End-of-Grade (EOG) and End-of-Course (EOC) tests in an American Indian school district. More specifically, the study will look at one benchmark system used by a tribally controlled school system in the Bureau of Indian Education's South and Eastern States Region (SESA). One focus will be to explore the development of the benchmark program in this school system. The study will also examine how well this benchmark program predicts students' subsequent scores on the state assessments given at the end of the year or at the end of the course. The goal is to understand what degree of predictive value such a benchmark assessment program has, especially with a school population of American Indian students.

Though a small minority in most--though not all--public schools in the United States, American Indian students comprise the population of schools either operated or funded by the Bureau of Indian Education (BIE). Typically, American Indian students

have underperformed other subgroups of students, both in public and BIE schools. According to the BIE (2011) for school year 2009-2010, in the 187 schools it oversees, only 30.58% of students were proficient or advanced in math, 39.65% of students in reading were proficient or above, and in science 24.77% of students were proficient. Similarly, National Indian Education Study (NIES), conducted by the National Assessment of Educational Progress (NAEP), found "Twenty percent of AI/AN students at grade 4 and 21 percent at grade 8 performed at or above the *Proficient* level in 2009" in reading (Grigg, Moran, & Kuang, 2010, p. 1). NAEP statistics for students in mathematics indicated that "Twenty-one percent of AI/AN students at grade 4 and 18 percent at grade 8 performed at or above the *Proficient* level in 2009" (Grigg et al., 2010, p. 3).

Research Questions

The qualitative research questions for this mixed method study are

1. What is the benchmark assessment program utilized in a small district, serving a predominantly American Indian population?
2. What are the results of the district's benchmark assessment program?
3. What are the benefits of the benchmark assessment program to the school community?

The quantitative research question for this mixed method study is

4. Do benchmark assessment scores (generated through FABA assessments) predict End-of-Grade (EOG) and End-of-Course (EOC) scores? And what are the implications if the scores predict well or fail to predict well?

Methodology

Qualitative data collection for this mixed methods exploratory study consisted of analyzing artifacts related to funding the benchmark assessment program, such as invoices and contracts. In addition, documents from the FABA (a pseudonym) program delineating the item development process were collected. These artifacts and their subsequent analysis provided information regarding how such a benchmark program was implemented with American Indian students in a small district. Interviews with district personnel were conducted. In addition, the qualitative portion of the study provided rich description of the conditions and situations during implementation of this program, offering useful research on the second research question, focused on the conditions necessary for a benchmark system to predict well.

Quantitative benchmark assessment and End-of-Grade/End-of-Course assessment data was collected for American Indian students in grades 3 – 12 who were enrolled in a tribally controlled school district in the southeastern part of the United States. Scores for reading and math were collected for students in grades 3 - 8, and scores for science were gathered for students in grades 5 and 8. For students in grades 9 - 12, scores were collected for students who were enrolled in English I, Algebra I, and Biology courses. Most students enrolled in these high school courses were in grades 9 and 10, though occasionally students in upper grades also enrolled in the courses. Scores generated from these assessments are ratio data. Correlation statistics—Pearson's correlation coefficient and coefficient of determination-- were run on the data. To account for small sample numbers and multicollinearity, the adjusted R^2 statistics were calculated. Multiple

regression statistics were computed for all of the assessment data, and simple linear regression statistics were applied when multicollinearity was evident. The quantitative data determined how well the benchmark scores predicted the later high-stakes assessment scores.

Definition of Terms

Benchmark assessments—assessments given a few times per year or course and whose data is used in both formative and summative ways by educators. Benchmark assessments may be developed in-house by teachers and curriculum specialists or school districts may purchase commercial products.

Bureau of Indian Education (BIE)—one of the bureaus within the Department of the Interior, charged with serving American Indian and Alaska Native schools and districts.

FABA (Formative And Benchmark Assessment)—an online, formative assessment tool.

Formative assessment—can be formal and informal assessments administered by teachers to generate data that will allow them adjust their instruction according to identified student needs. Formative assessments are low-stakes and occur often throughout the year or the course. Examples of formative assessments that teachers often use include short quizzes, questioning, "clickers," and exit passes.

High-stakes assessment—an assessment which is summative in nature and is used to rate the performance of a school. State end-of-year assessments are considered high-stakes assessments.

Interim assessment—used interchangeably with benchmark assessments.

Summative assessments—high-stakes assessments that occur at the end of the year or course, and determine a student's proficiency in the subject matter. Summative assessments can also occur at the end of a unit of study (e.g., chapter test) to determine acquisition of knowledge.

Assumptions

An assumption of the study was that all the benchmarks assessments are given in the same manner by each test administrator. Test administrators received training from the FABA trainers, and teachers received school and district level support. In addition, it is assumed that each test administrator for the End-of-Grade or End-of-Course assessment participated in test administration procedures training prior to administering the tests, a requirement by North Carolina's testing program. Another assumption is that students delivered their best efforts when completing the benchmark assessments and the End-of-Grade or End-of-Course assessments.

Delimitations

The sample was delimited to American Indian students from the state in grades 3 through 12, as these are the grade levels for which a state assessment might be administered. Most of the high school students were enrolled in grades 9 and 10, although a few older students were enrolled in the courses which have assessments (i.e., English I, Algebra I, or Biology). The study was also delimited by the assessments used by the school district, namely the FABA benchmarking tool and the North Carolina End-of-Grade and End-of-Course assessments. The study was further delimited to students

who are administered regular benchmark and End-of-Grade or End-of-Course assessments. Data from students who participate in alternate assessments was not collected.

Limitations

A limitation to this mixed methods study was that the benchmark assessments are used formatively and are considered low-stakes. Students who understood the low-stakes nature of the benchmark assessments may not have given the assessments their best effort, resulting in scores that were not representative of their level of mastery. In addition, all the benchmark assessments administered through the FABA benchmarking tool were delivered online for all students (grades 3 – 12). The North Carolina End-of-Grade assessments for students in grades 3 – 8 are administered as traditional paper and pencil assessments. The high school End-of-Course assessments for English I, Algebra I, and Biology are delivered online.

Another limitation considered was the small sample size, n, with the caution regarding the utility of results with such a small sample. This study also looked specifically at benchmark assessments created through FABA, which is a program available only in one state. In addition, the results from benchmark assessments given earlier in the year may have lacked utility due to the small number of objectives covered by the early assessments.

Finally, the investigator works in the assessment department within the school district. The interviewees may have disclosed more or less information to the investigator based their knowledge of her role in the school district.

Significance of the Study

This study was significant because it contributed to the literature base in multiple areas. It informed the literature on assessment practices, providing educators with more information about whether benchmark assessment programs are effective in improving student learning and if the benefits are worth the costs. Schools and districts are currently expending scarce funds for these programs, and knowing the quality of the predictive validity of a benchmark assessment system will inform future decisions regarding the purchase of the programs. In addition, the study shed light on other, non-financial costs associated with a benchmark assessment program such as time or autonomy of teachers.

Another area of significance regarded the sample population. American Indian students are typically under represented in the literature, especially regarding assessment and improving student learning. While the results from this study may not be applicable to other minority populations, the results are useful to tribally controlled school systems and Bureau of Indian Education operated schools and districts, as well as public schools with significant numbers of American Indian students. Typically, American Indian students scored lower in reading, math, and science than did their majority peers, and BIE operated or funded schools have difficulty in achieving Adequate Yearly Progress (AYP) targets. This study provided information for these schools in adopting and funding certain types of assessment programs.

Summary

In essence, this mixed method exploratory study investigated how a specific benchmark program was implemented in a particular district serving American Indian

students in the Southeast. It focused on how effective the benchmark assessment program was at predicting End-of-Grade and End-of-Course assessment scores in that school district.

Chapter 2

Literature Review

This review of the literature begins with a broad perspective of assessment history and narrows to benchmark assessments, ending with literature on their predictive validity. Figure 1 indicates the progression of the review of literature.

History of Assessment	>>>>>>	Definition and Purpose of Benchmark Assessments	>>>>>>
Benchmark Assessment Implementation	>>>>>>	FABA (Formative and Benchmark Assessment) Tool	>>>>>>
Use and Quality of Assessments	>>>>>>	Universal Screening in Response to Instruction (RTI) Models	>>>>>>
Predictive Validity of Benchmark Assessments			

Figure 1. Progression of literature review.

History of Assessment and Its Uses Prior to *No Child Left Behind*

With the advent of No Child Left Behind (USED, 2001), schools focused their attention even more on increasing overall test scores and closing achievement gaps with their minority population (e.g., economically disadvantaged, students with disabilities, minorities). No Child Left Behind basically eliminated the "sorting and ranking" (Stiggins, 2005, p. 325) purpose of schools by requiring that all students succeed at mastering certain standards. This law, perhaps more than any other catalyst, spurred the "data-driven" phenomenon that many schools and districts have embraced today. Educators found and created ways to produce data that would provide them with information about where students were at various points during the year, in order to better

direct classroom instruction and practice. The use of benchmark assessments became one way to gather data on students' performance throughout the year. Scarce research (Bulkley, Nabors Oláh, et al., 2010; Herman & Baker, 2005; Shepard, 2010) exists on the use of benchmark assessments. According to Bulkley, Nabors Oláh, et al. (2010) no aspect of benchmark testing has been well-researched. Though some districts create their own assessments, others purchase commercial products, but according to Shepard, commercial products are not often supported by studies either.

If the No Child Left Behind Act (USED, 2001) promoted an environment that led to the need for more assessments in public schools, then vendors certainly responded to this need, as noted by Burch (2010):

> In 2006, the top vendors in the testing industry reported annual sales in the range of $200 to $900 million. Firms show a pattern of increasing sales since the adoption of NCLB. Sales for 2006 were on average double the sales for 2000. (p. 152).

Burch (2010) believes this boom in the assessment technologies industry reflect both business practices and public policy. Most districts that have instituted benchmark assessment systems have done so with the goal of improving student learning. However, Burch suggested other reasons for districts to adopt benchmarking practices, stating "Schools institute practices and adopt policies because they hope it will give them an edge in looking institutionally legitimate" (p. 149). In their seminal piece, *Inside the Black Box: Raising Standards Through Classroom Assessment,* Black and Wiliam (1998) argued that though many mandates have been given in the effort to improve student learning, none of them have been particularly effective because those mandates do not support what happens in the classroom. According to Black and Wiliam,

classrooms are “black boxes” into which things are put (e.g., mandates, programs) with the expectation that certain other things (e.g., increased student learning) will emerge. The authors pointed out that no one is paying attention to or supporting what actually happens inside the box. Benchmark assessments are not a panacea, but Burch believes they are a part of the process toward increasing student learning, if each step is implemented with fidelity.

Definition and Purpose of Benchmark Assessments

For many districts that decide to implement a benchmark assessment program, building assessment literacy is a hurdle for staff. Many teachers do not know or distinguish between the various types of assessments or programs that a school might use. Schools often provide inadequate professional development for teachers on classroom assessment practices (Stiggins, 1995). As instructional leaders, principals must ensure that their teaching staff are assessment literate by providing professional development and support, however, most principals have not been formally trained in assessment literacy, either (Stiggins & Duke, 2008). Stiggins and Duke suggested that principal preparation programs will need to make changes to their program of studies in order to ensure that principals leave their programs able to provide the assessment support that teachers will need.

Most assessments fall within one of two categories – formative or summative. According to Bulkley, Nabors Oláh, et al. (2010), “Formative assessments occur in the natural course of teaching and learning,” (p. 117) and are frequent checks of student learning. McTighe and O’Connor (2005) agreed that formative assessment occurs

simultaneously with instruction and is used to direct teaching. It is this readjustment of instruction by the teacher that can enhance student learning (McTighe & O'Connor, 2005). Formative assessment can be short, tightly-focused quizzes, but it can also consist of teacher observations and questioning. The use of "clickers" can also provide teachers with immediate feedback on whether students have grasped a concept. Black and Wiliam (1998) agreed that formative assessment should be frequent, but brief. While all students involved in formative assessment processes will benefit, struggling students will realize the most benefit from the process (Black & Wiliam, 1998). Schools that understand this characteristic of formative assessment often begin their implementation of it as a school improvement strategy (Stiggins, 2005). Stiggins (2005) indicated that through the use of the formative assessment process, students can realize "achievement gains of one-half to two standard deviations on high-stakes tests" (p. 328). In a report for the Council of Chief State School Officers (CCSSO), McManus (2008) emphasized that formative assessment is "a process" (p. 3), not a one-shot type of test, and that students must be active participants in the process, beginning with establishing goals for learning and subsequently tracking their paths to the goals.

Summative assessment, on the other hand, is typically not used to adjust instruction, simply because it generally occurs too late in the instructional cycle for adjustments to happen. Stiggins (2005) stated that these late occurring types of assessments "lack sensitivity to instruction" (p. 326). State end of grade assessments or other types of high stakes testing are examples of summative assessments. These assessments are used "to measure students' performance against district or state content

standards" (Bulkley, Nabors Oláh, et al., 2010, p. 117) and so are not relevant to classroom instruction. McTighe and O'Connor (2005) offered a somewhat broader definition of summative assessment by characterizing it as assessment that "summarize[s] what students have learned at the conclusion of an instructional segment" (p. 11), indicating that summative assessments can occur at any time during the school year when a class has finished a unit of learning. For example, a chapter or unit test provides data on whether a student has mastered the content in the unit or chapter. A teacher typically does not use the data from the test to alter the course of the instruction, and the class moves on to the next unit or chapter of study.

Benchmark assessments, sometimes called interim assessments, occupy a somewhat murky place between formative and summative assessment. Benchmarks typically occur two or three times during a course or school year, and the data are used to measure a student's progress toward mastery of state standards. This characterization seems to put benchmark assessments squarely in the summative camp. However, while the data are used for summative purposes, most schools and districts use the data to adjust instruction and provide interventions to students, a formative characteristic. Stiggins and Duke (2008) stated that the formative information from benchmark assessments can direct educators' improvement efforts. Because the benchmarks occur before the end of the semester or year, teachers still have time to adjust their practice, and students still have time to master the content before the high stakes test. Bulkley, Nabors Oláh, et al. (2010) agreed that no definitive separation exists between the types of assessments, and interim assessments fall somewhere between formative and summative because they offer

data for prediction, for program evaluation, and for identifying student learning needs. In another study, researchers (Bulkley, Christman, Goertz, & Lawrence, 2010) indicated that benchmark assessments have many purposes, some of which include "instructional, evaluative, and predictive" (p. 187) purposes, and they are used "to inform classroom instruction" (p. 200). Olson (2005) agreed that multiple reasons exist for schools to use benchmark assessments including gauging student learning, providing actionable information for teachers, predicting high stakes scores, and pacing of the delivery of standards. Schools recognize the need for data that is both summative and formative in nature. State results arrive too late to influence instruction or increase student learning (Herman & Baker, 2005). Schools need to know where students are performing at different points during the year while they can still adjust instruction. For this reason, many high performing schools utilize benchmark assessments (Olson, 2005). To mitigate the limitations of end of year summative assessment, states, districts, and schools are beginning to test more often with administration of benchmark assessments, use the benchmark data to adjust instruction, and most importantly, utilize multiple types of assessment in the classroom with student participation (Stiggins, 1995). To meet accountability goals, schools must "link everyday classroom practices with schoolwide outcomes" and "develop data-driven practices" (Halverson, 2010, p. 130). Halverson, Prichett, and Watson (2007) stated, "Summative feedback describes the *results* of processes, while formative feedback is used to *inform* and *adjust* the process as it unfolds" (p. 4), which describes what most schools need a benchmark program to do. The North Carolina Department of Public Instruction promotes the use of formative and

summative type assessments, including the use of benchmark assessments in their "vision for 21st century assessments" (NCDPI, n.d.).

Guidelines and frameworks (Marshall, 2006, 2008; Perie, Marion, & Gong, 2009; Stiggins & Duke, 2008) have been offered for schools and districts that are eager to implement benchmark assessment systems. Perie et al. suggest that benchmark assessments should be one component of a balanced assessment system. They suggest that districts have only a few purposes for the benchmark assessments because no test "can serve more than two or three purposes well and they tend to work best when the various purposes have been prioritized explicitly" (p. 7). As for the purpose of prediction, Perie et al. advised that prediction should be only one aspect of a well-rounded, balanced assessment system, and they further cautioned that if a test used for prediction offers high quality diagnostic data, then the tests' capacity for prediction may diminish. Accordingly, schools and districts should ground their benchmark assessment systems with a Theory of Action that provides answers to questions ranging from who is to use the information to professional development needs of the users of the system, according to Perie et al.

Benchmark Assessment Implementation

Marshall (2006) offered 23 conditions for succession implementation of a benchmark or interim assessment system. Each of the conditions falls into one of four categories: Antecedents, Assessments, Analysis, and Action (p. 5). Marshall believes that the system should offer a pretest with subsequent benchmarks occurring at nine week or shorter intervals. Data analysis and data team meetings are vital to the success of the

program, according to Marshall. Stiggins and Duke (2008) believed that three questions should be asked regardless of the level (teacher, school, or district) at which the data analysis occurs. Those three questions are: (a) What instructional decisions are to be made based on assessment results? (b) Who will be making those decisions?, and (c) What information will help them make good decisions? (p. 286).

Feedback to students is also an integral component in most frameworks. All students should receive feedback, not just the "bubble" students, a practice Marshall (2006) believes is an ethically gray area for educators. Students should be involved in the data through goal setting and data tracking. Other researchers (Black & Wiliam, 1998; Stiggins, 2005) agreed with the necessity of involving students.

Marshall (2008) cautioned schools that they may sometimes encounter issues when implementing a benchmark assessment system. He identified several common obstacles such as some teachers not understanding why the assessments are necessary, others believing that the results of the assessments will be tied to their yearly evaluations. He also reported that if teachers are not analyzing and acting upon the data, then the program will not be successful. To prevent some of these issues, Marshall (2008) provided several guidelines for schools, such as providing exemplars, setting SMART goals, and holding data meetings. Ultimately, Marshall (2006) believes that the root cause for a failed benchmark assessment system is that teachers erroneously believe that if they teach a concept, then all their students learn it.

The Council of Chief State School Officers (CCSSO) has studied assessment systems and developed a workbook for schools and districts to use when embarking on a

benchmark or interim assessment system (Crane, 2010). The workbook offers a definition of interim assessment that aligns with other established definitions, especially that of Perie et al. (2009). The workbook offers several components that schools and districts must consider when developing an interim, or benchmark, assessment system, but the CCSSO believes that "Goals and Vision" are the most essential and significant of any of the other components (Crane, 2010, p. 4). Districts should set their purposes and then implement the appropriate foundational and methodological work in developing their system. The workbook also advised that districts must know the types of data and the levels of specificity of the data that the interim assessment system will offer to them. One consideration for schools is whether the data generated by the system can be used as part of its RTI process.

FABA-A Formative and Benchmark Assessment Tool

In the state where the study was conducted, many schools and districts use FABA, a program developed by an organization that works closely with the state's Department of Public Instruction (CUACS, 2008). FABA utilizes online delivery in providing assessment items that are aligned to the state's standards. Its purpose is to assist teachers in recognizing specific objectives that students have and have not mastered. The program can be used for formative, common, and benchmark assessments, though its primary purpose is for formative assessment. According to information from a FABA Overview presentation document, "formative assessments [are] based on the needs of the classroom and students" (CUACS, 2011, slide 6) "Common assessments [are] used to generate talking points for data meetings" (slide 6) and a common assessment "provides school

level data" (slide 6). Benchmark assessments are summative and "provides district level data" (slide 6). According to CUACS (2011), as of January 2011, FABA housed 77,000 items in its database, and all of the items undergo a rigorous development process. FABA usage in January 2011 involved 1,010 individual schools, 62 districts with 59 districts using the benchmark tool. FABA had catalogued over 503,000 students in its system (CUACS, 2011).

Little research is available on the FABA program, but one preliminary study available online (CUACS, 2008) analyzed End-of-Grade (EOG) and End-of-Course (EOC) results from schools that used the FABA system for math and compared those results to the EOG/EOC results from schools who did not participate in the FABA system during 2007-2008. The findings demonstrated "that on average more students in schools that give assessments using FABA pass the end-of-grade summative mathematic tests" (p. 4) compared to students not utilizing the system. It is important to note that this analysis focused on whether FABA improved overall proficiency rates on math EOG scores at the school level. The study did not look at individual student scores on the EOG and whether a relationship exists between a student's performance on FABA assessments and their subsequent EOG score. FABA states that the assessments should not be used for "prediction of future student performance on EOG/EOC assessments" (slide 7). This study focused on FABA as a formative assessment strategy only; it did not analyze the relationship of the benchmarking tool and End-of-Grade/End-of-Course results.

Use and Quality of Assessments

The quality of the benchmark assessment affects the degree of improvement in student achievement. Benchmark assessments should match the subject matter content that is taught in order to provide detailed information for teachers (Olson, 2005). In a study on how teachers from the School District of Philadelphia (SDP) used test results, Nabors Oláh, Lawrence, & Riggan (2010) reported that the SDP benchmark program provided short assessments that took little time to score, and the assessments were given on a six week cycle. Schools typically administered the assessments every six or nine weeks, depending on the course length and grading periods, but some schools provided the assessments monthly (Olson, 2005).

Halverson (2010) proposed that schools develop programs, based on systems theory, that fulfill the "three functions of intervention, assessment, and actuation" (p. 132). Interventions are comprised of two tiers, one of which includes school or district policies and school structures or paradigms and the second of which includes classroom based items or practices such as "textbooks, experiments, worksheets, computer programs" (p. 132). Assessments offer data to teachers for determining what students have learned. Halverson indicates that 'actuation' is the process of analyzing data and changing practice based on the data, so that teachers can connect a strategy or program to the assessment. In an earlier study utilizing the intervention, assessment, actuation feedback system, Halverson et al. (2007) described how the feedback system might look in a school:

> In terms of our formative feedback system model, the reading curriculum is the *intervention*, a battery of commercial exams used by Pearson teachers is the *assessment*, and the regular grade-level meetings for teacher reflection and action are the *actuation space.* (p. 10)

The researchers stress the importance of ensuring that the assessment matches the instruction; otherwise, decisions based on the data will be flawed. In this study, the researchers reported that the reading specialist at the school administered all the assessments in grades one and two for consistency and standardization. The reading specialist and teachers kept binders of student data, using them for longitudinal data as well as for parent conferences. The reading specialist spent half days in teachers' classrooms working with small groups, in addition to holding weekly and monthly meetings with teachers to review their data and assist with making instructional decisions. The researchers reported that the school did see improvement in students' performance on state reading tests, but they believed that the "ongoing attention to how reading is taught to specific students constitutes the heart of the school's formative feedback practices" (p. 21).

The degree to which teachers make use of benchmark or interim assessment data varies from district to district, and little research has been completed on this aspect of benchmark assessments (Nabors Oláh et al., 2010). Indeed, Wayman (2005) noted that schools often have abundant data from a variety of sources, but few tools or strategies to access the data and make it actionable. Wayman urged schools and districts to develop or obtain data warehousing and presentation systems, so that administrators and teachers can access the data they have gathered. Protheroe (2001) stated that using data effectively is

a difficult and complex task, and "typically, it was an evolutionary process that may have included some false starts" (p. 2).

In the Philadelphia district, Nabors Oláh et al. (2010) studied data use in average or above average schools, all making Adequate Yearly Progress (AYP) and who were involved in a larger study of the district's benchmark program. The populations in these schools mirrored the overall population of School District of Philadelphia. The researchers interviewed 25 teachers from third and fifth grade who had participated in the math benchmark assessment. The teachers used the week after the assessments were given to analyze the data and "revisit, reteach, practice, [and] enrich" (p. 28). An analysis of the assessment indicated that the distracters in the test items did not provide information about student misconceptions that teachers could use to focus their subsequent instruction. Instruction during this week involved whole group, small group, and peer tutoring strategies. The teachers utilized other adults such as student teachers and volunteers for small group instructional activities. Alternative instructional strategies involved "visualization or manipulatives" (p. 243). Though teachers analyzed data to see where student learning gaps occurred, the analyses did not provide data about students' general misconceptions of content, and therefore, teachers did not focus instruction on theoretical understandings, but rather the focus was related to "procedural" mistakes. Teachers were analyzing the data and using the information to some extent, but the analysis needed to be stronger. Olson (2005) agreed that what happens after data is made available is the important element, but often it is the weakest link in the process.

In another study of the School District of Philadelphia, Bulkley, Christman, et al. (2010) analyzed the benchmark assessment program from the district level. The district's program utilized interim assessments on a six week cycle. The assessments covered only the topics taught during the preceding five weeks. After students completed the assessment, teachers were to review the results, make instructional changes, and then retest to determine whether students had mastered the content. The district provided teachers with a protocol to use when analyzing the results of the assessments. The protocol contained questions related to student weaknesses, how the teacher might regroup students for interventions, and what the intervention might be. A protocol to identify necessary professional development was also given to teachers. The district office supported the teachers and the program by providing reports and resources online, protocols, professional development, time for data analysis through early release days, and School Assistance Teams for schools in AYP restructuring. Bulkley, Christman, et al. (2010), found that teachers were not as adept at the interventions needed after the data were analyzed, and the deficiency might be related to the lack of constructed response type items on the benchmark assessment. The researchers also discovered that during district data meetings, principals tended to compare the results of the current benchmark assessment to results from an earlier benchmark, an unhelpful practice when an assessment is not cumulative in design since the assessments contain different concepts.

In another study from the School District of Philadelphia project, Blanc et al. (2010) found that teachers needed to improve their use of data in order to see learning

improvements. Blanc et al. (2010) suggested a four step feedback system to regulate data use occurring after interim assessments have been administered. The four steps include "Accessing and organizing data," "Sense-making to identify problems and solutions," "Trying solutions," and "Modifying and assessing solutions" (p. 207). A benchmark or interim assessment system is only as good as the action taken from an analysis of the data. Clearly, Blanc et al. (2010) agreed,

> interim assessment data will contribute to changes in teaching and learning only if it is situated within a feedback system in which practitioners access and organize data, interpret data to identify problems and solutions, try out solutions in their classrooms, and modify their solutions based on new assessments. (p. 233)

The researchers also indicated that specific types of conversations need to occur in learning communities about data. The "strategic" conversation involved the "bubble" students, logistics, and quick growth ideas. The "affective" conversation involved discussions about the profession and pedagogy as well as motivation and encouragement. The "reflective" conversation detailed instructional strategies (Blanc et al., 2010, p. 212). The instructional leaders in the district and school also play an important role in making the data and the benchmarking system useful for student learning. "Data can make problems more visible, but only people can solve them," Blanc et al. stated (2010, p. 222). According to the researchers, various stakeholders possess different understandings about the purpose of benchmark assessments and those understandings influence their analysis of the data. Instructional leaders and principals must provide connections so all teachers are on the same page with regard to the purposes of the assessment program. One suggestion they gave for instructional leaders was to offer support for teachers who

are implementing interventions based on the data analysis by visiting their classrooms and using a protocol for the visits. Instructional leaders should use a protocol during the data analysis meetings, as well as have an agenda with "guiding questions" and a plan for the development of "next steps" (Blanc et al., 2010, p. 218).

Black, Harrison, Lee, Marshall, and Wiliam (2004) studied formative assessment in the King's-Medway-Oxfordshire Formative Assessment Project (KMOFAP). Though the project focused on formative assessment, many of their observations about the use of data are applicable for schools and districts implementing benchmark assessment programs. The researchers suggested that assessment practices do not become formative until teachers act upon the data. Similarly, until teachers act upon the data from a benchmark assessment, then the data will not influence teacher practice or student learning. Black et al. advocated that teachers should give fewer grades, but more feedback through the use of comments. This feedback should include what the student handled correctly, where the opportunities for improvement are, and how to get there. Their study also utilized a traffic light strategy for peer and self assessment. In the traffic light strategy, Black et al. believed teachers should use the yellow and red light areas as those on which to focus instruction. Teachers should support students by teaching them goal setting and helping them to work toward those goals. Conversely, principals should support the effort of teachers through providing time for sharing and collaboration, by incorporating the changes into the school improvement plan, and through policy modifications, if necessary (Black et al., 2004).

In another study, Brookhart, Moss, and Long (2008) researched communication between students and teachers and the effects on student learning. They found that over a two year period, Title I students performing at the below basic level dropped from 22.2% to 7.4%. Brookhart et al. analyzed three stages of teacher development. First, in "consciousness raising," many teachers thought they were already doing formative assessment, and indeed, many were utilizing some elements of formative assessment, but they were not communicating to students what their goal was. In the next, "skill building," the researchers found that teachers' use of formative assessment strategies is more purposeful; they are cognizant at this level that the process is more involved than they were initially doing. The third stage is "intentional," in which teachers engage in purposeful collaboration with students about their progress. This stage echoes the beliefs of Black and Wiliam (1998) that the process of feedback should allow students to explain what and how they understand. For this to be successful (Black & Wiliam, 1998), teachers must be open to conversations that meander and that produce unexpected information, rather than questioning students until given an anticipated response. According to Brookhart et al. (2008), as teachers master the process of formative assessment, their dialogue about it changes, indicating that they understand formative assessment is differentiation. The researchers also stated that engaging in this process with their teachers may motivate students by providing a sense of ownership in the learning process. This idea is echoed in the "Collaboration" attribute of formative assessment promoted by Council of Chief State School Officers (McManus, 2008) which

stated that “a classroom culture in which teachers and students are partners in learning should be established” (p. 5).

One concern related to communication and collaboration between students and teachers that educators have with implementing a benchmark assessment program is whether students are giving their best effort when completing the assessment. If the benchmark results are utilized formatively by teachers, then they need to have good data from which to make instructional decisions. If students are not motivated to take the assessment seriously, then the data from the assessment may not be a true indicator of what material they have mastered or where their difficulties lie. One way to address this concern is through the practice of grading the benchmark assessments. Hunt (2008) conducted a dissertation study analyzing the effects of grading a benchmark assessment. Hunt recognized that “students who are not motivated to do well on benchmark assessments may not take them seriously, thus skewing the results and making them less of a valid and accurate predictor of student achievement” (p. 6). His study focused on math benchmark assessments, and his sample was comprised of students at two different high schools in the same district. Prior to the third benchmark assessment, students at one school were told that their scores would be included as part of their course grade. Hunt found “that the addition of an external motivating factor, grading the benchmark assessment, significantly improved student scores on the third and final benchmark assessment” (p. 51). Grading the assessments did not improve the scores of students in the study who have disabilities in math. Hunt recommended further study on the predictive nature of the benchmark assessments.

In a study conducted to determine whether benchmark assessments are effective at increasing student achievement as determined by scores on the Pennsylvania System of School Assessment (PSSA), Hefflin (2009) found a connection between the benchmarks assessment scores and the high-stakes test, though it was stronger in the seventh grade than the eighth grade. In addition, he found that teachers used the data generated through the administration of the benchmark assessment program. Interestingly, Hefflin reported that teachers felt it was important to involve their students in data analysis, typically through conferencing, and that involvement contributed to improved student performance and learning.

Universal Screening in Response to Instruction (RTI) Models

Closely related to benchmarking programs is the Universal Screening measure found in Response to Instruction or Intervention (RTI) models. RTI is a tiered process that allows schools to identify students who are at-risk for failure and to increase the level and number of academic or behavioral supports that are needed. As students move through various tiers and interventions without significant improvement, screening for special education services may become necessary. However, many students who receive appropriate instruction and interventions will overcome academic deficiencies and not need special education services. In the RTI model, the Universal Screening is given approximately three times per year to students to identify any areas of weakness and to assist teachers in adjusting their instruction and in choosing appropriate interventions. According to Wright (2010), "The purpose of school-wide screening, therefore, is to allow buildings to proactively flag struggling students at an early point and match them to

appropriate interventions." The Universal Screenings utilize Curriculum-Based Measurement (CBM) tools. The North Carolina Department of Public Instruction (NCDPI) provides Responsiveness to Instruction training materials for its districts on its website. The training presentations discussed how RTI and its assessments (Universal Screening, Curriculum Based Measures) fit into a balanced assessment system which includes formative, benchmark, and summative assessments (NCDPI, 2011). NCDPI also states that Curriculum Based Measures have much to recommend them, including how the data can be used to predict performance on a subsequent assessment. According to NCDPI, oral reading fluency is "highly correlated with overall reading achievement .91" (slide 36).

Nese et al. (2011) looked at whether Curriculum-Based Measurements were useful in predicting students' later performance on high stakes tests. In the study, benchmark assessments were given to all students in the RTI program, providing helpful data to teachers early in the school year to target and shape instruction for students with learning needs. The school used easyCBM to deliver the Curriculum-Based Measurements to students. The researchers found that "easyCBM reading measures significantly predicted scores on the state reading test and that the vocabulary measures had the largest effects" (p. 612). They also found that benchmarks, or screenings, were a more sensitive predictor than prior achievement on state tests.

Atkins and Cummings (2011) also studied how well oral reading as well as retell fluency predicted reading proficiency with rural students in Montana. The researchers noted that few studies had been conducted with an American Indian population similar to

Montana, which has 12.5% American Indians in its population. Though IDEA (2004) suggested that RTI be used to assist schools in early identification of students with learning disabilities, Montana uses its RTI program for educational improvement, not for identifying students with potential learning disabilities (Atkins & Cummings, 2011). This type of implementation is aligned with the notion of using benchmark assessments to track students' learning progress on learning goals. Atkins and Cummings found that oral reading fluency predicted performance on MONTCAS (Montana reading proficiency test) and on the Iowa Test of Basic Skills (ITBS) for students in grades three and four. Additionally, they found that using retell fluency measures strengthened the validity of the oral reading fluency measures as a predictor of later reading proficiency.

In another study involving American Indian students, Pearce and Gayle (2009) studied Reading First schools in South Dakota. The sample involved 115 American Indian students from the Great Sioux Nations and 428 white students. The researchers analyzed whether the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Oral Reading Fluency scores could predict later performance on the Dakota State Test of Educational Proficiency. Their analysis focused initially on the oral reading fluency scores, the socioeconomic status, and finally ethnicity. Pearce and Gayle reported that American Indian students scored one standard deviation lower than white students on the state reading comprehension assessment and further that "results for the American Indian cohort indicated DORF accounted for approximately 41% of the variance ($p < .001$) of the outcome variable" (p. 423). Their findings indicated that DIBELS Oral Reading Fluency (DORF) predicted reading comprehension well for both ethnicities, and the

measure appears better suited for predicting who would be proficient on the state test rather than who would fail.

Two other studies (Hintze & Silberglitt, 2005; Wood, 2006) also looked at oral reading fluency as a predictor on later reading tests. Both studies found strong correlations between oral reading fluency and subsequent performance on reading proficiency tests. Hintze and Silberglitt (2005) found that reading curriculum based measurements (R-CMB) could be used for early prediction, stating "R-CBM appears to be an efficient method for predicting performance on high-stakes tests demonstrating the ability to predict those students who are likely to pass reading portions of such tests as far back as first grade" (p. 382). Wood's (2006) study examined whether oral reading fluency and its relationship to reading proficiency might vary with a student's grade level. Significant correlations were found for each grade level (grades 3, 4, and 5), and in addition, Wood found that oral reading fluency is useful as a predictor regardless of whether a student's oral reading fluency level was low or high.

Though many studies indicate that curriculum based measures such as oral reading fluency can be predictors for a student's later performance on a reading comprehension test, Petscher and Young-Suk (2011) found that oral reading fluency only somewhat predicted subsequent reading proficiency. The predictive ability of oral reading fluency of students with lower oral reading scores in grade 1 and the fall of grade 2 was less strong than that of students with higher oral reading fluency scores. The researchers speculated this might be due to the "floor effect" (p. 126) associated with student learning at these grade levels. For students at this point in their academic careers,

Petscher and Young-Suk suggested that oral reading fluency scores be triangulated with other student data when educators make decisions about whether a student is at-risk for reading failure.

Predictive Validity of Benchmark Assessments

In a study on the predictive validity of benchmark assessments, Brown and Coughlin (2007) found that benchmark assessments in the Mid-Atlantic Region did not predict performance on later state tests, although the benchmarks were psychometrically well-constructed. Their findings did indicate that the TerraNova benchmark did provide appropriate predictive information in one state for some grade levels. Brown and Coughlin believed that benchmark assessments created by districts typically are not validated for their intended purposes, but products from vendors should be validated for their stated purposes. Many districts and schools have developed benchmark assessment systems with prediction of student performance on subsequent high stakes tests as a stated, if secondary, purpose of the benchmark assessment system. Brown and Coughlin cautioned that "the predictive ability of an assessment is not a use but rather a quality of the assessment" (p. 4). While they suggested that further research is needed on the predictive validity of benchmark assessments, the researchers recognized that only bigger school systems have the personnel available to conduct predictive validity studies.

In a study of 38 grade three students, Barger (2003) noted that DIBELS oral reading fluency scores could be used to predict students' later performance on the North Carolina End-of-Grade reading assessment, $r = .73$. Barger stated, "100 cwpm [correct words per minute] seemed to be the dividing line in terms of making an accurate

prediction of whether or not a student passes the North Carolina End of Grade Reading test" (p. 4). Graney et al. (2009) also discussed the "predictive validity" (p. 122) of curriculum based measures in their study on growth during a school year. The Graney et al. study demonstrated more growth from winter to spring assessments than from fall to winter assessments, and they cautioned that growth in a school might not occur in a straight line.

How should a district, whether large or small, evaluate a benchmark assessment system? Herman and Baker (in Li, Marion, Perie, & Gong, 2010) believe that the technical characteristics of an interim assessment are secondary to the functional aspects of the assessment. Li et al. offered several criteria for schools to consider when evaluating their benchmark assessment systems. The researchers strongly suggested that schools look to the purposes of their assessment program since validity is strongly linked to an assessment's purpose. Additionally, schools should consider how the test was developed and administered, whether it addressed the needs of subgroups, whether it offers the types of reports and data the school needs, and its general usefulness. Li et al. stated that item quality is of primary importance and items must be tied directly to curricular objectives and be written at the appropriate level of difficulty. They also stressed that reliability, $r = .75$, for a low-stakes test used for adjusting instruction is appropriate, while $r = .90$ should be used for more high-stakes decisions. Marshall (2008) believes that benchmark assessments should be low-stakes, stating "Interim assessments are, by their nature, low-stakes and don't have to be psychometrically perfect. However, they must be good enough and long enough to provide teachers with

real insights for classroom follow-up" (p. 67). Crane (2010) believed that a district must determine the level of technical quality it believes to be appropriate. Some districts may want rigorous scientific studies; others may be satisfied with evidence of success in similar districts (Crane, 2010). Crane also suggested that retesting and item exposure that may result from retesting are also questions that districts must address in designing and evaluating an interim assessment system.

Rudner (1994) also offered advice for districts to use when evaluating assessments. According to Rudner, tests should have stated purposes with documentation that supports those purposes. In addition, reliability must be established using appropriate statistics. Rudner further advised that criterion measures be used to validate the test, and that districts should check the process of test development to determine the content validity of their assessments. Schools should follow the same test administration procedures each time the test is administered.

Summary

Schools and districts have become much more focused on collecting data about their students' learning, analyzing the data, and making decisions about instruction and programming based on the data. The larger role that data plays in education is a direct result of the accountability of schools imposed by No Child Left Behind (NCLB) (USED, 2001) and Individuals with Disabilities Education Act (IDEA, 2004). Districts began looking at various types of assessment to provide much needed data. Benchmark or interim assessment programs became one of the most important programs for which schools and districts spend their budget dollars. Indeed, the assessment industry has seen

rapid growth since the inception of NCLB, as districts implement initiatives aimed at diagnosing student learning needs long before the final assessment for NCLB is administered each year. Benchmark assessments can be a powerful instructional tool for schools. According to Marshal (2008), benchmarks or "interim assessments, if handled well, constitute the most effective single initiative that a principal can implement" (p. 68).

But for any assessment program to provide the data that schools need and be beneficial to students, educators must be assessment literate. Educators at all levels must understand the differences between formative and summative type assessments and how benchmark assessments tend to blend the characteristics of the two. Assessments vary in terms of purpose, frequency of administration, end users, and degree of accountability. Formative assessments occur more often, are used mostly by teachers and students, are low stakes, and have the purpose of improving student learning. Summative assessments, on the other hand, occur less frequently and usually at the end of an instructional year or unit, are generally used by administrators, are high stakes, and serve the purpose of accountability. Although benchmark assessments provide summative data for principals and district leaders two to four times each year for monitoring how students and programs are progressing, the assessments also provide teachers with data quickly enough that they can make changes in their instruction to shore up any areas of weakness that the assessment may have identified.

In addition, administrators must understand the importance of establishing the purposes of any benchmark program they may adopt, and how the benchmark assessments fit in the overall district assessment program or framework. Several

researchers caution districts about many issues they may encounter as they begin implementation of a benchmark assessment system. For example, (a) Will the program provide the type of data that the district needs? (b) How will the district provide professional development? (c) What will the professional development resemble? (d) Who will deliver it? (e) What types of protocols and procedures will be necessary to insure that the program is implemented with fidelity? (f) How do we ensure data use at the classroom level? and (g) What technology is required for successful deployment?

The Council of Chief State School Officers (CCSSO) (Crane, 2010) developed a workbook to assist districts with the development of a program that addresses many of these questions. North Carolina districts have access to an online system, FABA, which provides both formative assessment tools to teachers and a benchmarking tool for districts. Preliminary data indicated higher proficiency percentages on North Carolina's End-of-Grade and End-of-Course assessments for districts using FABA.

FABA utilizes a stringent item development process for the items housed in its databases. The system also sets parameters for teachers and district benchmark builders to ensure that the results from the assessments are valid. The quality of any benchmark program must be established for districts. Some researchers believe that less emphasis is needed on validity and reliability for less high-stakes assessments, such as formative and benchmark assessments. Others stated that vendors of benchmark assessment products should provide districts with rigorous statistical research on their products.

Several studies have looked at how schools utilize the data from formative and benchmark assessments. Providing results to teachers quickly and also providing them

with time to analyze the data and make instructional decisions are essential to a successful program. Many districts provide protocols and procedures to assist principals and teachers in utilizing the data. Re-assessing after providing the instructional intervention is an important step in the process. Another important aspect of data use is the involvement of students in analyzing and tracking their data. Students' involvement in their own learning and to this degree encourages motivation. Schools that do not take the time to develop these aspects of their assessment program often find that they have a plethora of data, but no one knows what to do with it. Becoming a strong data user is important at the classroom, school, and district level. Without a strong background in data and in the content being assessed, teachers often focus on formulae and strategies rather than on the essential questions or major concepts of the content. Without a strong understanding of data practices, principals and district leaders often err into comparing performance from one benchmark to another, although the tests are assessing different content standards and objectives.

Though little research has been done on benchmarking systems as such, several studies have looked at universal screenings and other curriculum based measures utilized in Response to Instruction (RTI) programs. These assessments are typically given three times a year to all students to determine learning needs. The results of the screening provide information to what types of interventions may be needed to improve a student's academic skills. Like benchmark assessments, these screenings are given with the same frequency, and the data are utilized in the same manner and for similar purposes.

Many of these RTI-related studies look specifically at oral reading fluency tests and whether they can be used to predict a student's subsequent performance on a reading comprehension test. Most often oral reading fluency tests are strong predictors of later reading comprehension performance. While several studies involving oral reading fluency measures provided similar results regarding predictive ability, some studies of other benchmark assessment programs did not indicate a strong predictive relationship. However, Herman and Baker (2005) believed "if the benchmark tests are doing their job, there should be a strong predictive relationship between students' performance on the benchmark tests and students' performance on the state assessments" (p. 53). This is, of course, the hope and belief of many school and district leaders as they struggle with choosing the appropriate programs and strategies to increase student learning.

Chapter 3

Methodology

Introduction

The purpose of this study was to explore the history and nature of a benchmark assessment program and how the assessments are used in relationship to high stakes End-of-Grade (EOG) and End-of-Course (EOC) tests in an American Indian school district. This chapter will describe the methods used in the study, and it will be comprised of several sections. The first section explains the type of study for the project, followed by a statement of the study's research questions and IRB approval. The next major sections describe the qualitative and quantitative data collection processes. A description of the target participants of the study is covered in section. Another section is comprised of a description of the school district and the process for obtaining approval for the study. The following sections include a description of the benchmark assessment instruments and the North Carolina End-of-Grade (EOG) and End-of-Course assessments (EOC). These instruments comprised grades three through eight and high school and cover reading, math, and science (grades 5, 8, and 10 only). Additionally, the sections discuss data storage, validity and reliability, and data analysis. The role of the researcher is also described.

Characteristics of a Mixed Methods Design

The mixed methods exploratory approach was chosen because this study probed both qualitative and quantitative questions regarding benchmark assessment programs. Richards and Morse (2007) noted the necessity of utilizing more than one method to

provide a more extensive study of the topic. A qualitative approach was appropriate both for understanding how a benchmark program was implemented at a particular school and for understanding what conditions should exist in order for a benchmark assessment program to meet the expectations of a school or district. Quantitative methods were appropriate for determining how effective this particular benchmark program was at predicting subsequent high-stakes assessment scores. A characteristic of mixed methods research, according to Creswell (2005), is that "quantitative data results . . . refine and extend the qualitative findings" (p. 516). Figure 2 indicates the sequence for the study.

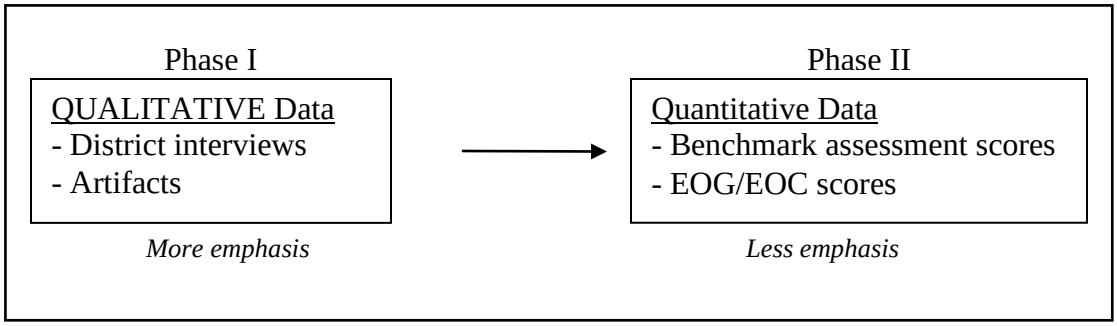

Figure 2. Sequence and emphasis of study activities.

The process denoted by Figure 2 was similar to the process described by Creswell (2005) for an exploratory design. Creswell explained that in an exploratory design the qualitative data carries more weight than the quantitative data that is collected later. This study collected the qualitative data (Phase I) prior to the quantitative data (Phase II), as Creswell suggested, and emphasis was placed on the data that was obtained through interviewing personnel from the district. The quantitative data collection captured the test scores of the sample on benchmark assessments and End-of-Grade (EOG) and End-

of-Course (EOC) assessments. As Creswell explained, "the procedure of first gathering qualitative data to explore a phenomenon, and then collecting quantitative data to explain relationships found in the qualitative data" (p. 516) fit well with the study. Most educators understand that benchmark scores could be used to predict subsequent assessment scores. However, educators need information on the conditions that would or would not lead to a benchmark assessment program having a strong predictive nature on subsequent assessments. The collection of both qualitative and quantitative data provided the type of data necessary to understand the topic. The assessment scores alone cannot provide educators with the information necessary to implement a solid benchmark assessment program. Nor would the data obtained from a purely qualitative study provide educators with information about how well the FABA benchmarking tool predicts later EOG and EOC scores. As Richards and Morse (2007) explained, often a study is best served through mixed methods because a single method "will not provide a comprehensive answer to the research question" (p. 93). This mixed method study sought to shed light on the complex nature of benchmark assessment programs and their predictive nature.

Research Questions

The qualitative research questions (Phase I) for this mixed method study were

1. What is the benchmark assessment program utilized in a small district, serving a predominantly American Indian population?
2. What are the results of the district's benchmark assessment program?

3. What are the benefits of the benchmark assessment program to the school community?

The quantitative research question (Phase II) for this mixed method study is

4. Do benchmark assessment scores (generated through FABA assessments) predict End-of-Grade (EOG) and End-of-Course (EOC) scores? And what are the implications if the scores predict well or fail to predict well?

Institutional Review Board

This study was conducted after approval was granted from the Institutional Review Board of the University of Nebraska – Lincoln.

Qualitative Data Collection (Phase I)

Qualitative data collection focused on artifact analysis (e.g., invoices, training materials) and interviews. Primarily, the researcher interviewed district personnel regarding the implementation of the benchmark assessment program in the district. Interviewees included principals, assistant principals, and teachers. Interview questions involved an exploration of how each user was involved with the benchmark assessment program, how each user perceived the strengths and weakness of the program, and how each user believed the predictive nature of the program affected its implementation. An interview protocol was utilized for all district personnel interviewed (Appendix A). Interviews were recorded, transcribed, and coded. Following suggestions from Creswell (2007), audio files were stored on a computer and back-up files were also maintained.

Artifacts were collected, analyzed and coded. These artifacts included training schedules, enrollment reports, and invoices.

Quantitative Data collection (Phase II)

The quantitative portion of data collection involved benchmark assessment and End-of-Grade (EOG) and End-of-Course (EOC) scores. The school district already had a benchmark assessment program in place, with benchmarks developed for each grade and subject area. Testing windows for each of the benchmark assessments were established at the beginning of the school year. The school district scheduled three benchmark assessment windows for year-long courses in grades 3 – 8 and English I, and it scheduled two assessment windows for Algebra I, Biology, and some English I (semester long) courses. Each year-long benchmark assessment occurred near the end of the quarter for the first three quarters of the school year. For semester long courses at the high school level, benchmark assessment windows fell during the middle of each of the two quarters. Table 1 displays each assessment, corresponding grade level, course length, and number of benchmark assessments administered.

The testing window for the state End-of-Grade (EOG) and End-of-Course (EOC) assessments typically fall within the last three weeks of the school year. The district already required student participation in its benchmark assessment program and state end of year assessments. The district provided a data file containing the benchmark scores and the EOG or EOC scores. A spreadsheet containing student test scores was constructed, containing test scores from both two or three benchmark assessments and the final EOG or EOC scale score. Though identifiers for each student's data were needed initially, once the data set was completed for each student, the identifiers were removed, thus increasing confidentiality of the data. The data was cleaned to eliminate any missing

Table 1

Subject Areas Assessed, Grade Levels Duration of Course, Number of Benchmark Assessments Administered

Assessment	Grade levels	Course Length	Number of Benchmarks Administered
Reading	3 – 8	Year-long	3
Math	3 – 8	Year-long	3
Science	3 – 8	Year-long	3
Algebra I	9-12	Semester	2
English I	9-12	Year-long and Semester	3 for Year-long; 2 for Semester
Biology	9-12	Semester	2

student scores or any scores that did not fit the population parameters (i.e., non-American Indian). The last benchmark assessment was administered in April 2012, and the EOG/EOC assessments were administered in May 2012. The data files and back-up files were stored electronically. The researcher also maintained hard copies of the data files in a locked cabinet. All raw data will be destroyed one year after the completion of the study.

Study Participants

This study utilized nonprobability, convenience sampling. The researcher used test scores from American Indian students attending a tribally controlled school system in the Bureau of Indian Education's South and Eastern States Agency region. Student scores were sampled from grades 3 - 8 and, in high school, scores from English I, Algebra I, and Biology were included in the sample. The school system implemented a

benchmark assessment system with assessments given two or three times during the year, depending on the length of the course. At the end of the year or semester, a state End-of-Grade or End-of-Course assessment was administered. The sample contained students who had scores from all of the benchmark assessments as well as the state assessment.

The population of students whose scores were included was 772 students in grades 3-12 at the time of the study. Most of the students enrolled in English I, Algebra I, and Biology were in grades 9 and 10, which put the sample population closer to 630 student scores. Though a tribal school system, a few students were not American Indian. The data was cleaned to reflect only complete score sets of American Indian students.

Site Identification, Description, and Approval Process

The school district incorporating the target population was a small, Appalachian district in the southeastern part of the United States. It is controlled by a federally recognized American Indian tribe, and the tribe and the Bureau of Indian Education (BIE) funded the school district. Nearly 1,100 students were enrolled in one of the three schools (one PK-5 elementary, one 6-8 middle, and one 9-12 high school) that comprised the district. Seventy-two percent of the student population qualified for the free or reduced lunch program.

The researcher presented the study to the entire school board during its annual retreat in June 2011 and received the board's approval for the study at that time. Later, the school board chair provided a letter of permission for study (see Appendix B). Initially, student scores were identified by name and identification number, though once all test data had been accumulated for students, all identifiers were eliminated from the

data file. Student and parent permissions were not necessary because the study did not require any student participation since the test scores are archival school district data.

Instruments

The school district selected for this study utilizes FABA, an online formative assessment program, with a benchmarking tool. FABA is a program developed by organization which often partners with North Carolina's Department of Public Instruction on various projects, including ones from the testing and accountability division. FABA trains item writers, often teachers in North Carolina, to write multiple-choice test items based on North Carolina's Standard Course of Study goals and objectives. The program was developed specifically for North Carolina schools, so the item banks are not generic repositories usable by any school in the nation. Additionally, each item created is vetted through a meticulous item development process (CUACS, 2011, slide 5). Test items are created at easy, medium, and hard difficulty levels for each goal; additionally, test items are created for various levels of thinking. Item stems and foils are written in the same format as the items on North Carolina's End-of-Grade (EOG) and End-of-Course (EOC) assessments (slide 3).

The actual benchmark assessments were developed by teachers employed in the participating school district. Teachers constructed the benchmark assessments by choosing items from the FABA benchmarking database with guidance from the district. Specifically, assessments were to follow the pacing guides for the grade and course; pacing guides are based on North Carolina's Standard Course of Study goals and objectives. Each assessment is cumulative, containing items for goals and objectives that

have been taught up to that point, so that students are assessed on information recently taught, as well as objectives that were taught earlier in the year or semester. None of the benchmark assessments created by the district contain more than 50 items, and the teachers building the benchmarks try to vary the item difficulty and thinking skill level to provide students with different types of questions. Scores are determined by the percent of the items on an assessment a student answers correctly. Benchmark assessment reports may be generated at student, class, school, and district levels.

Proficiency in reading and math is measured each year, and science proficiency is measured at grades 5, 8 , and 10, through a student's performance on the End-of-Grade (EOG) or End-of-Course (EOC) assessment. These assessments were developed by the North Carolina Department of Public Instruction (NCDPI). NCDPI publishes a technical report for each of the assessments that it develops. The EOG and EOC assessments contain items with varying levels of difficulty (NCDPI, 2008a, 2008b, 2009). In addition, according to NCDPI, the assessments also include Marzano's (Marzano et. al., 1988) thinking skills levels. Test items are generally created by teachers trained as item writers, although the mathematics EOG and EOC and the English I EOC assessment items were written by trained teachers and a vendor (NCDPI, 2008a, 2008b). Item writers created items for each standard in North Carolina's Standard Course of Study. Test items moved through six phases which include tryouts, field testing, pilot testing, and finally, operational testing, a process that can take 44-49 months (NCDPI, 2008a, 2008b). North Carolina utilized multiple forms of each assessment at each grade level. Each form in a grade level is equivalent (NCDPI, 2008a, 2008b).

Reading assessments in grades three through five contain 50 multiple-choice, four foil items. In grades six through eight, reading assessments contain 56 multiple-choice items, each with four foils. Students are asked to read and answer questions about various types of text, e.g., fiction, nonfiction, poetry, content-related, and consumer-related (NCDPI, 2007). The English I EOC required students to analyze literary texts and to analyze student compositions, and the assessment consisted of 80 total items, although only 56 items are operational (NCDPI, 2008a).

According to North Carolina's technical reports (2008a, 2008b, 2009) the testing program converted students' raw scores into scale scores. Scale score ranges are developed for each of four achievement levels. The program also provides percentiles for students.

Reliability and Validity

Alternate form reliability statistics were unavailable for the FABA benchmark assessments because only one form of each assessment was available. Students were not allowed to re-take the test, so test-retest reliability could not be determined. The program produced an item analysis for individual benchmark assessments, but the report did not generate internal consistency coefficients. The district did not have the capacity to perform reliability statistics on its benchmark assessment program.

Instructional validity for the benchmark assessments was determined by the use of teacher benchmark builders who are familiar with the content and who typically teach the content. Content validity was maintained by use of pacing guides to construct the assessments. The pacing guides are aligned to North Carolina's Standard Course of

Study goals and objectives. The district had not performed any studies to determine criterion-related validity, concurrent validity, or predictive validity. The present study sought to establish predictive validity for the benchmark assessments.

Reliability for the EOG reading assessments was calculated with internal consistency coefficient statistics, and "the NC Statewide Testing Program follows industry standards and maintains a reliability coefficient of at least 0.85 on multiple-choice tests" (NCDPI, 2009, p. 44). NCDPI reported that the lowest coefficient alpha for EOG reading was in grade 8, 0.897, and the highest for grade 3 EOG reading, 0.925 (p. 44). Reliability for the English I EOC assessment was 0.91 (NCDPI, 2008a). Likewise, reliability was strong for the mathematics assessments, both EOG mathematics and the Algebra I EOC. According to the technical report (NCDPI, 2008b), "Looking at coefficients alpha for the different groups reveals that across all test forms, in all grades and subjects, 57% of the values were at or above 0.90 and all but 5 (97% of all reliability coefficients) were above 0.85" (pp. 59-60).

Construct validity for the EOG reading assessments and the English I EOC assessment was established utilizing item writers who are familiar with the goals and objectives in the North Carolina Standard Course of Study (NCSCOS). Additionally, the construct validity was established through teacher and curriculum expert reviewers. For instructional validity, NCDPI also provided item review questionnaires to teachers, and each response was carefully analyzed to determine the appropriateness of the test item. Concurrent validity was measured using Pearson correlation coefficients for criterion such as anticipated scores and anticipated course grades. According to NCDPI (2009),

"the correlation coefficients for the NC EOG Reading Comprehension Tests range from 0.50 to 0.69, indicating a moderate to strong correlation between scale scores and external variables" (p. 61). English I EOC correlation coefficients were similar, "0.51 to 0.69" (NCDPI, 2008a, p. 54). NCDPI (2008b) reported strong concurrent validity for its mathematics tests as well, establishing high relationships between EOG/EOC mathematics scores and SAT and NAEP results. Content validity for all EOG and EOC assessments were derived from the goals and objectives of the curricula. Likewise, North Carolina trained teachers from schools across the state to write items for the state assessments, thus ensuring instructional validity of the EOG and EOC assessments (NCDPI, 2008a, 2008b, 2009).

Sample Size

Phase I: Qualitative participants. For the interview phase of the study, the investigator utilized purposeful selection of participants from the district. Participants were chosen based on their expertise and involvement with the program. Specifically, the investigator interviewed four administrators and 10 teachers who were involved in the benchmark assessment program.

Phase II: Quantitative sample. This study looked at the scores of students enrolled at each grade level or in each high school course to determine the sample size. The smallest number enrolled was 64 students in grade seven. Grade 8 had the largest number of enrolled students at 94. The final sample size for each of these grades and courses was smaller than the enrolled numbers for a variety of factors. Student scores

that were not American Indian were not included. Student scores from alternate assessments were not included, and incomplete data sets were eliminated.

According to an online statistical calculator (Soper, 2006-2012), for a study with three predictor variables (benchmark 1, benchmark 2, and benchmark 3), $r^2 = .15$, $\alpha = .8$, $p = .05$, the sample needed to include at least 76 participants. With the enrollment levels at the time of the study, at least two grades/courses did not contain enough participants to have an adequate sample. With some subjects/grades, simple linear regression statistics were also calculated when multicollinearity issues were suspected. With only one predictor variable, the sample size required was 54 participants. The district used benchmark assessments that contained only test items that had been covered to that point in the course. In other words, the assessments were not comprehensive or did not consist of items from all the objectives of the course. Benchmark three (or benchmark two for semester-long courses) most resembled the subsequent EOG or EOC in that it consisted of test items from most, if not all, of the objectives of the course.

Data Analysis

Phase I: Qualitative data analysis. Qualitative data analysis included coding of any artifacts obtained from the district regarding the benchmark assessment program and the transcribed interviews of district staff. The researcher used topic coding for all interviews as a means of determining what information – and possible themes and categories - was available in the data (Richards & Morse, 2007). Analysis began when data collection began. That is, following the recommendations of Richards and Morse,

the researcher initiated analysis of data as it became available. Reflection was a major data analysis strategy.

Phase II: Quantitative data analysis. Quantitative data analysis focused primarily on simple linear regression statistics to determine the predictive nature of the benchmark assessments. Assistance with data analysis was sought from the NEAR Center (Nebraska Evaluation and Research Center). For each data set, the values for the regression equation will be calculated. The equation (Gravetter & Wallnau, 2009) is $\hat{Y} = bX + a$, $b = SP/SS_x$ and $a = M_y - bM_x$.

Multiple regression statistics were performed on all complete data sets. However, as noted earlier in this chapter, the information obtained from the multiple regression calculations may lack utility of small sample sizes or multicollinearity issues. Simple linear regression statistics were calculated in some instances when multicollinearity may have been evident.

In addition to regression statistics, Pearson correlation coefficient, r, the adjusted R^2 was computed. The adjusted R^2, the multiple correlation squared, is a measure of strength of association. The NEAR Center recommended using the adjusted R^2 which adjusts for small sample sizes and multicollinearity errors. The correlation statistics determined the utility of using the earlier benchmarks as predictors.

Role of the Researcher

The researcher serves as the Director of Testing and Data Management and for a short time (December 2011 through mid-March 2012) served as interim superintendent of the district to be studied. As such, she is interested in the research questions posed by

this study. The results of the study could guide future decisions regarding the district's assessment system. The researcher has no vested interest in a particular assessment system, whether it is the FABA tool or another system such as NWEA MAPS assessments.

Summary

This mixed methods study sought to understand the benchmark assessment program (e.g., type, conditions, and appropriateness of the program) used in a particular district (Phase I: Qualitative) and how well or poorly the benchmark assessment scores predicted later scores on a high-stakes assessment (Phase II: Quantitative). Artifacts were collected from the district and district staff were interviewed to determine how the program had been implemented and the conditions that surrounded the benchmark program as well as the high-stakes assessment (Phase I). Additionally, student scores from the benchmark assessment program and the EOG and EOC scores were analyzed to determine the predictive ability of the benchmark scores (Phase II).

Chapter 4

Results

Introduction

The focus of the study is to explore the benchmark assessment program implemented in one school district, serving a tribal population. The study incorporated mixed methods, utilizing participant interviews and analysis of student test scores. The qualitative portion of the study was the primary focus, and this chapter will describe only the qualitative data collection process and analysis. Chapter 5 will discuss the quantitative portion of the study.

Five themes that emerged from the analysis of the participant interviews will be discussed in detail. Results organized by the three qualitative research questions will then be discussed. Quotations from the participant interviews will be used to illustrate both the themes and the research questions.

Qualitative Data Collection

Description of the participants. The qualitative data collection began with a pilot study of the interview protocol. The protocol was sent to four educators, two principals and two directors, who had served on the district's AdvancED Quality Review Team, earlier in the year when the district sought district-level accreditation. Three of the pilot study participants were in the same state as the district and were familiar with its curriculum and assessment programs, including FABA. One participant was from a different state, but worked with a Bureau of Indian Education (BIE) school, and thus had a familiarity with tribal student populations and BIE requirements.

Each of the four participants reviewed the protocol and provided feedback. Three of the participants did not recommend any changes. One participant did recommend a change that was not implemented, as it involved incorporating an unrelated component (i.e., the district's character education program) into the protocol.

After the pilot study was completed, interviews of various school district staff, faculty and administrators, began in late June 2012 and ended in late September 2012. Fourteen individuals were interviewed, with the administrator interviews occurring during June and July. Due to the summer break, the teacher interviews were conducted during August and September. Table 2 displays the participant descriptor information.

Table 2

Participant Descriptors

	Gender	Ethnicity	Average Years in Education
Administrator	M = 1	W = 1	32
	F = 3	W = 3	13.3
Teacher	M = 2	W = 2	8.5
	F = 8	AI = 3	14.3
		W = 5	4.8

Note: W = White; AI = American Indian

The four administrators had worked with the district for a varying number of years. The lone assistant principal participant had only been an administrator for one year. Of the three principals interviewed, one had been a principal for only a few

months, but had served as an assistant principal at one of the schools for five years previously. One principal had served as assistant principal and principal at different schools in the district for one year in each role. The third principal had been an administrator for 25 years, having served one year as assistant principal and two years as principal in the district.

Likewise, the teacher participants in the study had a diverse number of years serving the district. Several teachers were new to the profession, having completed only one or two years at the district. Others were mid-career and veteran teachers. Table 3 describes the number of years participants had been with the district.

Table 3

Years with the District

	0-5 Years	6-10 Years	11-15 Years	16-20 Years	21+ Years
Administrators	3	1			
Teachers	7	1	1		1

The teacher participants were also diverse in terms of school assignment (elementary, middle, or high school), as well as varied in terms of teaching duties. At the elementary school, the teachers worked with students in grade three and grade four. At the middle school, the participants were assigned to specific grade levels, but they were also discipline specific. The high school teachers instructed students from several grade levels, but each taught in a different content area. All teachers were regular program

teachers with the exception of one, who was an inclusion teacher with the special education program.

The teacher and administrator participants were similar to the demographics for their respective groups in the system. The school district employed more female than male teachers, and few of the faculty members were American Indian. While more teachers from the faculty as a whole had worked for the district between 6 and 20 years, most of those teachers were not assigned to one of the tested grades or subject areas. None of the administrators with the district are American Indian. The study included all but two of the administrators, one of which left the district at the end of the school year, and the other had only worked for the district for a few months. Table 4 lists each of the participants, their position, and level. Pseudonyms have been used to maintain confidentiality.

Six of the participants had one year of experience in developing benchmark assessments for the district using the FABA program, and one of the participants had two years of experience in creating benchmark assessments for the district. The district had utilized FABA as a formative and benchmark program for two years in grades three through eight and for three years at the high school level for Algebra I, Biology, and English I.

Interview process. All teacher participants were interviewed in classrooms, and the principals were interviewed in their office or a conference room at the school. Participants signed Informed Consent forms prior to the interview. The researcher

Table 4

Participant Names, Positions, and Levels

Name	Position	Level
Uma	Principal	Elementary School
Wendy	Assistant Principal	Elementary School
Ralph	Principal	Middle School
Irene	Assistant Principal	High School
Patricia	Teacher (Grade 4)	Elementary School
Ida	Teacher (Grade 4)	Elementary School
Hannah	Teacher (Grade 3)	Elementary School
Jaclyn	Teacher (Grade 3)	Elementary School
Xavier	Teacher (Math)	Middle School
Sam	Teacher (Inclusion)	Middle School
Rachel	Teacher (English Language Arts)	Middle School
Wanda	Teacher (Math)	High School
Sarah	Teacher (Science)	High School
Roxane	Teacher (English Language Arts)	High School

followed a standard interview protocol with each participant (Appendix A). Each interview was recorded by the researcher and later transcribed by an University of Nebraska - Lincoln administrative assistant who had received Institutional Review Board (IRB) training prior to serving as a transcriptionist.

The transcribed interviews were read several times in order for the researcher to gain a more holistic understanding of the content. The researcher conferred via telephone with a qualitative and mixed methods consultant at the University of Nebraska-Lincoln's

NEAR (Nebraska Evaluation and Research) Center at the beginning of the data analysis process to discuss possible approaches to coding and analysis. An initial list of possible codes and the transcripts were loaded into the data analysis software program Dedoose (2013, version 4.5), a web application for analyzing qualitative and mixed methods research data. Some of the codes that were generated were in vivo codes, but most of the labels were common terms within the education field. Descriptor characteristics for each participant were also uploaded to the program.

Using Dedoose (2013, version 4.5), the researcher coded excerpts from each of the transcripts. A second consultation with the NEAR Center consultant occurred during the coding and memo-writing to ensure that the codes were being applied appropriately. A coding matrix (appendix C), listing all the initial codes used with each participant, was generated. Through the reading, re-reading, and coding of transcripts, five core themes emerged from the study: professional development, assessment literacy, data literacy, instructional practice, and program effectiveness. All of these themes were codes generated during the initial coding process. All of the other code labels identified initially related to one of these five identified themes.

Qualitative Themes

Theme 1: Professional development. The professional development theme covered a broad array of topics (e.g., training, introductory training, professional development access). All participants discussed their FABA training, which varied somewhat among participants. Some of the comments from administrators and teachers about the FABA training are discussed below:

Administrators. Ralph, a middle school principal, discussed how "we practiced pulling up test results." He also shared that "I learned how to view each teacher's results, each class or even individual results on the benchmark test."

An elementary assistant principal, Wendy, stated, "We were able to go in and learn how to build things [assessments and reports], and learn how to design things, and how to really use it." And she felt, "it was nice to have someone straight from the program developer providing us with the training."

Uma, elementary principal, believed, "It [the training by the FABA consultants] was very functional training."

Teachers. Sarah, who taught high school science, shared:

> We had the training with you [the researcher] and weren't there some people that came from Raleigh when we first started it? That was . . . it was informative, but it wasn't until I was able to get on there [the FABA program] and start playing with it and making my own assessments and quizzes that I was able to understand it better. I'm more of a hands-on person and you can tell me how to do it, but I want to be able to try it and figure it out on my own.

Fourth grade teacher Patricia also shared her perspective on the FABA training:

> Two years ago now, we did have somebody come in and go over an overview with us, and they did kind of show us how to build things. It was a one-time thing done at the beginning of the year. I think it was like three hours long, and they showed us a lot of stuff that didn't necessarily apply to the younger grades, that are more helpful for middle school and high school.

Sam, a middle school inclusion teacher, also remembered the training sessions, "We had two times when somebody came in from FABA and did formal trainings for a few hours a piece. Then you've [the researcher] done refresher courses throughout the time."

Not all teachers felt as positive about the training. Rachel, who teaches English Language Arts in the middle school, shared one of her frustrations about the training:

> We wouldn't know how to print those [the reports] out. I don't know how she [the principal] knew, but she would print those out and bring them to us. We'd look at it and say, 'Well, this would have been helpful weeks ago.' If they could teach me, in our training, what she knows how to do, that would be helpful.

However, a high school math teacher, Wanda, indicated that locating reports was a part of the training:

> Yes, they [the trainers] showed us how to access the reports and what it is that we're looking at. How to access it by student, by the class, by the school, to see how our students are doing compared with other students within the school.

In addition, many of them reviewed the professional development they had received regarding how to use assessment data. Administrators and teachers shared their experiences with data training.

Administrators. Elementary assistant principal Wendy explained her formal and informal data training:

> Formal training, I guess most extensively, has been in my Master's work. We spend a lot of time in various courses in my Master's in School Administration [program], talking about, …. from a data analysis standpoint, from a statistics standpoint, to looking at it 'ok, that's what the numbers say, but what does this really mean?' [to] if you're advising or working with a teacher to improve. So that's my most formalized training in using data. I have, since being an administrator, I've learned a whole lot on the job, informally, and not always informally, but through little workshops here and there, or projects, or activities that we as an administrative staff in our district participated in. I've learned a lot through those experiences, but, I would say, predominantly, my on-the-job training is day to day working with principals who work with data who've had experience with data and just spending time with it and learning from them.

Teachers. High school math teacher Wanda remembers receiving data training sponsored by the district, "We had a data workshop that focused on actually using the

EOG [End-of-Grade] scores for eighth grade, for placement of students, to actually see where our students are coming into the ninth grade."

Xavier related an important point from the FABA data workshop that he attended,

> I do actually remember the person [the trainer] saying that she used it [the strategy], and her goal is for her kids to get to 80% mastery of whatever skill that was. She had a little transparency and used it every year, and that helped her feel comfortable going into the state assessment.

While some participants, especially administrators, acknowledged the role their graduate work played in developing their knowledge and skill with data, other participants, such as Sarah, a high school science teacher, related different experiences. Sarah shared, "We didn't cover it [data use] in college in any of my classes. What I've had has been the things we've done here with our staff development time." She continued:

> InformEd, they came and did the workshops with us over the course of the whole year. We started at the beginning of the year, and we had some more in December when we came back from Christmas. They showed us how to analyze and compare.

Hannah, a third grade teacher, shared her experience with the InformEd training, "She [the trainer] showed us how to categorize our kids into what they didn't know and what they did based on objectives. We were then supposed to go back and re-teach the objective they did not get."

Five codes emerged within the professional development theme, totaling 115 times in the transcripts.

Theme 2: Assessment literacy. Most of the participants' comments, especially those from the teacher participants, indicated only a rudimentary understanding of assessment literacy topics. Comparability of scores from benchmark assessment to

benchmark assessment was confusing for many participants. Ida, a fourth grade teacher, had grasped the difficulty with attempting to compare the results of one benchmark to the results from the previous benchmark. She shared, "Each quarter is kind of different. You want to see growth, but there's different material [on the assessment]. I'm not really sure we're comparing what we want to compare. It's like comparing two different things."

Another topic within the assessment literacy theme that the participants discussed was the advantages and disadvantages of the types of test items, specifically constructed response items versus multiple choice items. For example, teacher and administrator participants asked for a definition or clarification of the phrase 'constructed response test item.' One administrator, Uma, asked "Tell me what a constructed response would be." Both administrator and teacher participants asked for clarification, and both had ideas about the benefits and weaknesses of those types of test items. Irene, another administrator, believed that constructed response items would be better, and stated "I think that constructed responses can show what students know. Often, multiple guess shows what a student doesn't know." Hannah, a third grade teacher, would have preferred to have a benchmark assessment containing constructed response items:

> Well, my students do better with these kinds of responses anyway because they're not getting tricked. I feel like whenever we're discussing [in] class, and the kids are writing their own sentences about answering the questions, they will give me at least partially the right answer. So I can say, 'What do you think about that?' But when they're doing A, B, C, or D, it's like there are two that are right and you have to decide which one really is right. So I'm like, 'did they get tricked?' Or did they totally not get it? Or did they guess right?'

Sam, an inclusion teacher with the special education program, was not as enthusiastic about using constructed response or short answer questions. He shared that using these

types of questions for students with writing disabilities "might actually get less of a response of their knowledge."

Another teacher, Wanda from the high school, noted that the addition of constructed response items to the benchmark assessment would be helpful because of the adoption of the Common Core State Standards. She said, "With the Common Core . . . they're going to be required to have those types of questions on their test [End-of-Course], or to answer those types of questions. I think it would be helpful to see how they would score it." The scoring concern of such types of questions was readily apparent to the teachers. Xavier, a middle school math teacher, discussed the issue, "I think the briefly constructed responses might be more time consuming, and I don't know how . . . how do you grade it? Then certain people grade it differently."

Hannah, a teacher, indicated that she understood other types of test items and assessments as well. She was a proponent of computer-adaptive testing, a type of test that modifies the difficulty level of subsequent test items based on the responses that a student provides to initial questions. She would have been happier if the benchmark assessment program was designed to be computer-adaptive. She related she would prefer to do:

> more differentiated testing for students not on grade level, so I can really see what they don't understand. I don't really know if it's because they didn't understand [the] author's purpose, or they weren't really able to understand what they read.

Jaclyn, another teacher, also discussed computer-adaptive testing for her students:

> If the kid gets the question correct, it [the computer-adaptive program] will bump them up a month, or a grade level, or whatever. So if they're answering more complete answers, they'll get more difficult questions, and they can go up the ladder of grade level equivalency. With the benchmarks, it's set in stone.

Participants who developed some of the benchmark assessments indicated through their comments the importance of developing assessments that tested higher-level thinking skills. These comments supported the Assessment Literacy theme. "A lot of time, too, the questions that are in our testing are straight knowledge-based. So, to really understand what students understand, you have to get away from just knowledge-based [questions]," related Wanda, a teacher, indicating that she understands the value in test items that penetrate to a higher level thinking skills. Sarah, a high school teacher, echoed this understanding by saying that working on the district benchmark assessments "made me more aware of the different levels of knowledge, the organizing [of] those different skills that they had to do. I was finding, as I was doing it, I was looking at those things to make sure that I had a good mix." Third grade teacher Jaclyn's comments about critical thinking skills and testing were more pointed:

> I think the EOG [End-of-Grade assessment] needs to be revamped so there are short answer responses, because it's not enough, I think, to have the right answer. In terms of Bloom's Taxonomy, you need to be able to explain it, critical thinking. So many kids can't do that. I think it's a really important skill to have.

Elementary teacher Patricia revealed through her comments that she understood that assessment was more than testing what students should know:

> It's [the benchmark assessment program] made it more clear what assessments are for. You know you assess, but the FABA system makes it really easy to go back and look at their scores and see exactly where they're missing. You can give a test, you grade it, and you hand it back out. Yeah, you've done an assessment but are you using it to drive your teaching and to go back and hit those things that they're missing. So, that's been really helpful for me as a beginning teacher.

Sam from the middle school believed that teachers needed "continuous encouragement to use FABA, not just as [a] benchmark, but as weekly reviews to look at that data as well to see if they're understanding that objective, or the objectives for that week."

Administrators also recognized the necessity of utilizing formative assessments. Elementary principal Uma understood that formative assessment was important, believing that the district's benchmark assessment program "highlights the need for formative assessment." Later, she returned to the topic, "That's [formative assessment] the goal. It's hard to get teachers – some teachers – to buy into that. I think the teachers that we currently have, have seen the need to do that."

The researcher condensed four codes (i.e., test item literacy, formative assessment, assessment literacy, teacher developed benchmarks) into this category. The codes were identified a total of 65 times in the transcripts.

Theme 3: Data literacy. The participants' comments demonstrated foundational knowledge of data literacy concepts. Many of the comments were concerned with the data analysis process. Ralph, an administrator, said that he could "interpret the data, and I can compare that to the objectives and generally see if the teacher's being successful or not." Wendy, assistant principal, acknowledged that prior to the implementation of the district's current benchmark assessment program, she had not seen "a whole lot of really effective data analysis going on," but that the current program "give[s] very clear guidance in the discussions [of the data]." She explained the basic process that she has seen teachers engage in:

> We looked at it [the data], we pulled together, we had a discussion, made changes based upon what the data was saying, and then went back again and looked at it

> for a third time, and saw a change. That's been my experience with how we've used it. We pull it down, and then we meet, so at least quarterly, we come together as grade blocks, or departments, I would say.

Irene, another administrator, believes that the district's administrators would benefit from a brief data retreat to help them with the analysis process. She shared, "We know the process, but we don't do a very good job of leading [the process] ourselves. If we had a facilitator, a couple of hours and a facilitator, we'd get a lot done that way."

Teachers also revealed their data analysis process. The teachers reported frequently that they would focus on the objectives with which students struggled or on the specific test items that stymied their students. Wanda, a high school teacher, disclosed:

> The first thing I look at is the questions . . . identify questions that are troubling for most students. Generally, if, I'd say, 40% or more miss the question, I'll go back and look at it, and see if it's how the question was written or if there's something in there that's getting them.

Ida, an elementary teacher who is less comfortable with computers, would apply more traditional analysis strategies. She would,

> Print out the whole thing, cut it, and tape it together, and highlight the ones they missed, so you can readily see, and line it up. You can see right off the bat which questions the majority of students are missing. I'm a very visual person, so I have to lay it all out and it's about four feet long. Tape it together and highlight the ones they missed.

After pulling her data together in this manner, Ida would meet with another teacher in her grade block on weekends to review the data. Sometimes, she says, "We may decide to reteach certain areas together."

Some grade blocks would analyze data together. Teachers would examine the benchmark score reports and objective reports, comparing each groups' performance.

Hannah articulated, "We would pull up the data for each class, and see how we're averaging with the state and district, and see if we're super far below or on the right track. There are little graphs you can get."

For some participants, involving students in the data analysis process was crucial. Teachers would present data to their students and have the students work with it as they felt was appropriate for their students. Xavier, a math teacher in the middle school, shared data with his students and involved them in the data analysis process. After reviewing the class percentage correct for each item, his students will choose which items to examine as a whole group. He related:

> We go over the benchmark together. I'll show how they did as a class, and I'll say 'Which one should we go over?' Well, only 16% got this one right, so let's go over that one. They'll choose it, and I want to know what happened because apparently we all missed it. More of them think it's a competition. 'Did I get it right when the others got it wrong?' They're analyzing, but from a personal standpoint.

Patricia, an elementary teacher, also involves her students in the data analysis process. When her students take the benchmark assessment, she has them use a worksheet to show their work. She collects these and hands them back when it's time to review the assessment. She explained:

> The FABA system has this really great thing where you print off a paper, and each problem has a little box. Since I teach math, I have them record something written in every single box. They have to record every single answer that they get. So when I go over it on the SMART Board, we go through and we work out every single problem, and they can see exactly where they missed….Did I divide instead of multiply? Did I subtract instead of divide? So they are really seeing, 'Oh, this is what I've done wrong.' That's really helpful to see, then I can look at their work, too, and when I see that they've missed a problem, did they just make a silly little arithmetic error or are they way off in left field and don't have a clue what they're doing.

Participants articulated how data was analyzed and then was used to make instructional decisions. Assistant principal Wendy wanted to "provide them [teachers] with data that helps them evaluate where they go next in their instructional practices, as well as pinpoint those pocket areas where there's severe gaps in the learning, and they can adjust accordingly." Irene, another administrator, believed her teachers often made instructional decisions based on their data:

> Our teachers are very quick to say, 'Wait a minute, the way that I've taught concept 1 apparently is not effective. I'm going to have to reteach it, think of a new way to look at it, try this project and see if the students can grasp it that way.' I think they're very quick to do that every time they get their benchmark scores.

Teachers also viewed the data as the vehicle for making decisions. Patricia was emphatic in her response, "Basically, any time you have data, you need to look and see how it can drive your teaching." Her colleague, Ida, provided an explanation, "If the majority of the class misses a certain type of division problem or something, I can focus on that more." Sam offered this illustration, "In math, for example, [I] make a more concrete homework project to try to see if they can understand it better."

Not all teachers were as comfortable with making decisions based on data. Sarah, a high school teacher, initially related that she did not know the next step after analyzing the data, but later in the interview, she contradicted herself by giving examples of how she made decisions. Early on she stated, "I know how to read it, and I know how to interpret it, but I don't know what to do with it after that." Later, she revealed, "It's helped me plan review sessions and know which students need to attend tutoring. Who needs the additional time, and who doesn't need the additional time. It helps me know if I'm doing a good job." Sarah's latter comments indicated that she does, indeed, make

decisions about her student's learning and her instructional practice. Her statements also indicated that she engages in self-reflection and evaluation.

Xavier, a middle school teacher, was identified by two administrators as a teacher who understood data and used it in his classroom. Xavier combined data, goal-setting, and competition to motivate his students during a difficult time of year. He described in depth how he utilized data near the end of the school year to help prepare students for the End-of-Grade assessment:

> I created boy and girl spreadsheets with each standard, and this was basically our review near the end of the year. This is not specific, necessarily, to the benchmarks, but FABA and using the quizzes that they had prepared, or I would make my own. Often times, there was one category that I saw a lot of people missed, so at that point I said, 'You know what? Let me remake another one, and let's talk about what happened and where things went wrong.' I made a goal for them, called 'Angry Birds.' Each skill they got, they got a little piece of the puzzle, and they created little scenes. Then there was an 'Angry Birds' party. So, it was kind of like, the last nine weeks, this is what we did. Some of them started getting frustrated with it, but they pushed through. I think it helped a bunch of them near the end to be successful because they were seeing these questions over and over.

In addition to instructional decision-making taking place in the classroom, the district uses data to make school-level decisions. In the elementary school, the administrators and teachers in a particular grade block had decided to implement specialized classes for the students. Some teachers were responsible for teaching reading to all students in the grade block, while others were tasked with teaching all students the math curriculum. Wendy, an assistant principal, explained how the school used the data to evaluate that decision.

> We had implemented something very new in the fifth grade, and after looking at that data for a semester, and comparing the two quarters, we met with the group of teachers to see what they thought had happened. It basically came out that they

> felt like maybe we didn't need to be blocking their classes, [and] that they wanted to go back to keeping their own kids and being responsible for all subjects as opposed to specializing and blocking. So, they changed at that point and started back to the traditional classroom structure after Christmas. From what I gather, their scores improved.

For the high school, Irene, assistant principal, related that the administrators and teachers had used data to determine that foundational and support courses were needed in grade nine. The school determined that grade nine students were not transitioning well from year-long courses in middle school to semester-block courses in high school. Foundation math courses were added for first semester, and an English Language Arts support course was also added. Irene commented, "Those are some of the changes that we've made course-wise, curriculum-wise, to accommodate some of the needs that we saw through the numbers."

Both Wendy and Irene remarked on the importance of using data with the school improvement process. Irene stated, "In those processes in creating [a] strategic plan or school improvement plans or restructuring plans, we have to start using more data in those things as well. We have to start having data-related goals." Assistant principal Wendy shared that the elementary uses various data, including benchmark assessment data, to set goals and create plans. Wendy indicated that her school's School Improvement Team (SIT) had already incorporated benchmark assessment data in to their work:

> We're identifying what our needs are and setting our goals for the following school year. Then, as we're monitoring our progress, on our school improvement plan, we use that data formatively to contribute to the plan, and to tweak it.

Thirteen codes ranging from ‘analyzing data’ to ‘team planning’ and ‘data led decisions’ were combined to create this theme. The transcripts provided 280 total code responses, shaping the Data Literacy theme.

Theme 4: Instructional practice. Participants spoke often and at length about what they do – their strategies and techniques – in their classrooms, or in the case of the administrators, what they see or want to see occurring in their schools’ classrooms. Wendy, assistant principal at the elementary school, believed that benchmark assessments should be used “to gain information into their [teachers’] instructional practices” and to “evaluate where they go next in their instructional practice.” At the high school, assistant principal Irene was more specific about the instructional use of the benchmark assessment data, stating, “We also want to use it as a gauge for what we need to reteach, what we need to enhance, what we need to enrich.” Wendy and Irene’s comments indicated they have an understanding about the instructional benefits of a benchmark assessment program, but their responses did not indicate if the reality in their schools matched with their ideas on the purpose of the program. Uma, an elementary administrator, spoke more directly about the degree that the data is influencing instruction. Recognizing that teachers should be using data to focus their instruction, Uma expressed her doubt as to the degree this was occurring in this way, “I think that’s part of something that we’ve missed here – is ‘what now?’”

Administrators have a much broader perspective on what happens in the school than teachers do, but teachers in the study revealed that they are paying attention to the data and using the information to modify their practice. Sarah, a high school science

teacher, used the benchmark assessment data to make changes to the units she taught in subsequent semesters. She explained her changes, "The unit that we're doing right now, I've added maybe four or five new activities this year in the unit, based on things that I've seen from testing in the last year."

Teacher comments on instructional practice often centered on re-teaching and remediation efforts. "I think the biggest thing is the re-teaching. Especially if it's something that I thought that the students had," explained Wanda, a high school math teacher. Benchmark assessment data was used to structure one school's remediation program, according to one elementary teacher, Ida, who shared:

> Last year we also had a remediation program and that [benchmark data] was . . . very helpful in helping us determine what we needed to remediate on as opposed to tutoring which is more specific skills for each individual. Remediation was basically what the whole class seemed to be missing. We could also group them so if one class was having a hard time with multiplication, we could put them all in one group for remediation.

Hannah, another elementary teacher, also discussed the remediation time the school had set aside to address student needs. She related, "Last year we did remediation time . . . so the kids that were having certain issues with maybe graphing . . . I would pull those . . . kids to a group and have something [an activity] on the internet." One teacher from the middle school, Rachel, was less enthusiastic in discussing how the benchmark data influenced her instructional practice. She expressed her procedure for adjusting her instruction, "When you get your results, reteach the objective."

Teachers often acknowledged the need to differentiate instruction to address individual student needs or a small group of students' needs. "I've gone over certain standards more heavily because a bunch of the kids didn't get it. I've targeted certain

students and pulled them from another class, and we've gone over it individually or with a small group," shared Xavier, a math teacher. Patricia, an elementary teacher, provided this example of how she differentiated for her students, "If they're getting all of these wrong, we need to go back and have a small group of these students who are missing this part of subtraction and not regrouping correctly." Patricia also related how she would differentiate for a small group of students by utilizing the inclusion teacher:

> If it is only like four or five students who have missed whatever part it is, I'll get the other kids going on an independent activity that they can do. I'll pull them over and work with them in small group. Or, I have utilized [the inclusion teacher] for that too because a lot of times it's her students who are missing that [concept], who are in inclusion. I'll say, 'We're really not getting this. I've got your three and I've got two in my regular class. Can they come over to you for a couple of days while you guys hit that hard?' Then they come back over.

However, teachers also struggled with the best way to implement a differentiated classroom. Sarah, a science teacher, wrestled with how best to assist students who had mastered the material:

> When I look at it [the data], I know what I am looking at, and I know that if they're not meeting a certain point, then they need more instruction. But, if they're over, and I can tell that they understand the concept, I don't know what to do with them. Do I keep working with them on that concept because the rest of the class still needs it? Or do I move them on to the next thing?

A few teachers mentioned reviewing the test questions once they had the results of the benchmark assessment. Teachers Xavier and Patricia reviewed the assessment with their students, furthering the idea of assessment for learning. Patricia shared, "We go over it as a whole class so the kids can see what they got wrong and why they got it wrong." Though Xavier reviewed the test with his students, he also believed the FABA program inhibited his ability to do that efficiently:

> You're not allowed to print things off really. . . . It's challenging to get the whole benchmark and to go over that with them where they can manipulate it and try it again versus me going over it on the screen. Then, after a while that's tough. The benchmark's kind of long so some point you're not really getting the return on it. You spend a day and maybe they understand a couple more questions.

Rachel, another middle school student, echoed Xavier's frustration, "That's the part we need--easy access to the questions so we can go over them." Despite the difficulty with accessing the questions for reviewing the benchmark assessments, teachers were reviewing them and using them as an instructional tool. Ralph, an administrator, understood the value in that instructional practice. Ralph related that he had a teacher who did that, and his expectation was that all of his teachers would utilize that instructional strategy. He related, "He [the teacher] goes over each question with the students in his math class. . . . So the expectation as the administrator would be to get teachers to that point, or get all teachers to that point, so it's a process."

The instructional practice category identified 10 codes in total, e.g., differentiation and remediation, totaling 198 times in the transcripts.

Theme 5: Program effectiveness. Participants peppered their responses to the interview questions with comments that related to the effectiveness of the district's benchmark assessment program. Though some limitations were identified, generally, teachers and administrators spoke about the effectiveness of the benchmark assessment program and the benefits derived from it. The benchmark assessment program afforded Wanda, a math teacher, the opportunity to delve deeper into the weight that the state placed on particular goals and objectives in the curriculum:

> It [the benchmark assessment program] gave me a chance to really look at the concepts to see [pause] I did a lot of research on the high, medium, low importance that the state ranks certain concepts, and just being able to see a different variety of questions and wording.

Patricia from the elementary school also appreciated how the program provided her with test items and data on specific objectives from the curriculum, "I like the FABA program because it does break it down by objective. I think that is a really strong thing to know exactly when I look back at my data."

The objective breakdown that Patricia references is part of the alignment of the FABA assessment program with the state curricula, a strength that several of the participants discussed. Uma, an administrator, was enthusiastic about this aspect of the program, saying:

> I think FABA is wonderful for the [state] End-of-Grade test because it's perfectly aligned with the [state's] End-of-Grade test. That's what it was developed for, specifically, not any other state. So the data that we get from that is very accurate regarding our state tests.

This belief resonated with Wendy, another administrator, who said of the program:

> It simulates the EOG [End-of-Grade] type questions which often times our teacher-made assessments may not do, or any other book assessments provided to them through the [textbook] resources they use may not do. This is a tool that's more closely related to EOG, which is the way we assess at the end, so it's more authentic in the long run.

While discussing the benefits of the benchmark assessment program with teachers and administrators, Irene, an high school assistant principal, also revealed that students see the FABA assessments as an effective tool for them as well. She imparted that the older students "realize how important the EOC [End-of-Course] itself is. They're starting to realize about halfway through the year, it's [the benchmark assessment] an indicator of

how well they'll do on the EOC." However, not all of the teachers believe students see the efficacy of the assessments, as Irene does. While acknowledging the importance of the program for teachers, Xavier, who teaches younger students, verbalized:

> I think with the EOG [End-of-Grade] testing, we're talking about it constantly. Constantly talking about the importance of what it means, so I think they do better with that. Whereas, I think with the benchmark, they're just kind of…they don't see the importance that we see in it.

It is the constant pressure of the high stakes state assessments mentioned by Xavier that compels teachers to seek tools—such as a benchmark assessment program--to assist them with preparing students for those assessments. The district's program provides these tools, according to Sarah, a high school teacher:

> It [the benchmark assessment program] really does give a good view of how they're going to be in the end. The students that are putting forth the effort, and they're trying, and they're understanding, they are doing well on the benchmarks and, at the end, on their EOC [End-of-Course]. So, it's given me kind of a heads up of what to look for.

Ida, who teaches students younger than the students that Sarah teaches, concurred with Sarah's evaluation of the program:

> The teachers can realize how their students are progressing. So we know what to work on, what weaknesses they have, what strengths they have. It gives the students a chance to see their weaknesses and strengths also. It helps prepare for the EOG. It shows us, also if there's any growth to an extent.

Not every teacher was as positive about the district's program, however. While admitting that it provided her data with where to go next in her instruction, Rachel felt that the assessments did not align well to her teaching:

> It's good for showing what we need to work on, but a lot of times it was out of order to what I was teaching. It didn't really align to what I was teaching. There was a lot of stuff I didn't get to that it was hitting on. So, it really didn't mesh. It wasn't the end all, be all, for me.

The benchmark assessments were designed to be aligned to the pacing guide for the subject. All teachers of the subject were to use the district's pacing guide. Rachel may not have been using the district's established pacing guide, or because of disruptions to the school schedule (e.g., inclement weather, last-minute assemblies) her teaching may not have been aligned with assessment.

Generally, the participants believed the district's benchmark assessment program to be effective for them, but Ralph tempered his response, "It's very useful and dependent on how much we require the teachers to use it, or [dependent upon] how important from the administrative level that we deem it to be."

Because of the frequency and depth of discussion around it, the Program Effectiveness theme became this study's core phenomenon, a facet of qualitative research discussed by Creswell (2007). Program effectiveness encompassed 22 codes which were identified a total of 343 times in the transcripts. Consistent with Creswell's Grounded Theory description, determining the effectiveness of the program, however, involved the conception and analysis of the other themes mentioned previously.

Qualitative Data Summary by Research Question

Introduction. The five themes discussed previously shed light on the three qualitative research questions explored by the study. This section will provide answers to the research questions given the themes that emerged from the interviews.

Research question one: What is the benchmark assessment program utilized in a small district, serving a predominantly American Indian population? The first research question was concerned with the processes--not just the FABA product--

involved in the school district's benchmark assessment program. The benchmark assessment program was comprised of FABA, as well as what occurred before and after the benchmark assessments were administered. The district's benchmark assessment program consisted of the pacing guide, training, the FABA program, data analysis, and follow-up instruction, which includes formative assessment. This section will discuss each part of the district's program and will include how the previously identified themes help to answer the research questions.

In order to bring consistency within a subject across the grade level, the district had established pacing guides developed with input from teachers. The district had engaged in curriculum development for several years for the core subject areas, generating curriculum guides that contained a pacing guide. The benchmark assessment program utilized the pacing guides as assessments were built. Each assessment was aligned to the content that had been taught up to that point on the pacing guide. Though each assessment was different because of the new content that was taught since the administration of the previous benchmark, each assessment also contained items from objectives that were assessed previously. This allowed teachers to ascertain whether students were retaining information from previous quarters.

Though the district had provided benchmark assessments to its students for a few years, in 2010-2011 it purchased the FABA product for both the formative and benchmarking tools. The decision to make this move involved several factors. Administrators and teachers wanted a benchmark assessment tool that was aligned to the state's End-of-Grade and End-of-Course assessments. Having been developed by an

organization that often assists the state's department of public instruction with test item development, the FABA system was tightly aligned with the state curricula and assessments. In addition, the district was looking for a program that could be delivered via the web. Online state assessments for the high school students had already been developed, and online assessments for the younger students were in development. The district preferred an assessment system that would prepare students for the shift to the online testing environment, which FABA did. The online delivery system also meant automated scoring, speeding up the turnaround on grades, and it also freed up district resources in terms of paper and ink costs, as well as time spent copying assessments. Financially the purchase of FABA made sense for the district, too. The cost for three schools (approximately 800 students) in 2011-2012 was $8100, which included both the formative assessment tool ($3600) and the benchmarking tool ($4500).

In addition to pacing guides and the FABA product, the district's benchmark assessment program also consisted of professional development for the FABA program. As a member of the administrative team in the area of testing and accountability, the researcher was involved in the professional development provided by the district. The district provided a typical roll-out of professional development when it first implemented the FABA system. The high school staff attended this initial training one year prior to the other schools because they used the system for formative assessment one year prior to the district switching to FABA's benchmarking tool. Trainers from the program were on site for two days to provide the basic, introductory training. Each teacher was provided with a half day of training on the system. Principals and teachers who would deliver

technical and administrative support were provided with an additional half day of training. Xavier, a middle school teacher, recalls his FABA training:

> I remember sitting in a room with computers and being taught how to log in, how to use the specific [pause], how to navigate the website, how to create a quiz of my own, or to schedule a quiz. I felt very confident in how to use it after the training.

Other participants had similar recollections of the initial training by the district. Hannah, who does not teach in the same school as Xavier, recounted her training experience, which is comparable to Xavier's experience:

> It was a long time ago. Well, we had one this year, we did a follow up. Two years ago, my first year teaching here, I think a person from Raleigh had come in and taught us how to set our classes up, how to make quizzes, how they are premade, and how to have the kids take the quizzes. She kind of showed us how they were aligned and what purpose they were for our school.

This initial training allowed teachers and administrators to begin using the system immediately for formative assessment, and it allowed the district to begin building its benchmark assessments. Later in the year, a trainer also provided staff with an additional half day of training on using the data to improve achievement. During the second year of implementation, the district did not contract with the program for additional training sessions. New teachers were trained by colleagues or by the researcher in one-on-one sessions. Invariably, some newer staff members or teachers who had been moved from a non-tested grade or subject to a tested grade or subject did not receive any training or support. One administrator participant acknowledged this problem:

> Somebody new may slip in under the radar, be expected to use it, and because they're new, may not have the confidence to ask or admit they don't know what they're doing or they'll go to a colleague and ask for it, and they're going to get just a very quick exposure kind of training to it, whereas we got a lengthy training.

The lack of formal professional development for these teachers put them at a disadvantage in implementing the benchmark assessment program with fidelity and skill. This lack of training might be more problematic for the formative part of FABA, but may not impact the actual administration of the benchmark assessments as strongly.

Theme 2, Assessment Literacy, informs the first research question through two of its components – formative assessment and teacher-developed benchmark assessments. Many of the participants discussed their benchmark assessment program as an aspect of formative assessment. In fact, many of the participants identified the primary purpose of the benchmark program as formative. Wendy, an elementary administrator stated, "I believe that our benchmark assessment program's purpose is to provide teachers with a tool that is controlled and in a way, as far as the development, to be very objective and also formative." In her five years as an assistant principal, Wendy had witnessed the district using different types of benchmark assessment systems. One of those systems combined teacher made test items with commercially developed items, but the district built its own assessments from the combined item bank. Another program was a total commercial product (i.e., test items and assessment).

Teachers also viewed the benchmark assessment program in formative terms or as one component of formative assessment. One high school teacher, Sarah, remarked, "I try to use the FABA as much as possible, not just for the benchmarks but I like to try to do it once a week." Roxane, another secondary teacher with only one year of teaching experience, had a slightly different perspective on the use of the system. She shared:

> I guess it's really been one of the few types of formative assessments I've seen so just by its existence it helps me, I guess, think about assessment as a way to figure out what you need to teach rather than a way to figure out if you taught it already.

Roxane arrived at the school having worked in the business field prior to moving into education. As a somewhat older first year teacher, Roxane pinpointed one of the tenets of formative assessment, determining what one needs to teach.

As the Chapter 2 Review of Literature indicated, benchmark assessment occupies a nebulous place in the assessment continuum because it can be both summative and formative in nature. Many of the study participants' remarks substantiated the dual nature of benchmark assessments in their district. While many of the participants recognized that benchmark assessments could be formative in nature, Uma, an elementary administrator "think[s] it highlights the need for formative assessments; for formal formative assessments." This statement is an indication that Uma understands that assessment for learning is a process, that benchmark assessments can be used formatively as a component of the process, but that other formative assessments should also be utilized by classroom teachers.

Six of the teachers interviewed for the study had participated in creating the benchmark assessments that the district administered by pulling items from the FABA benchmarking database. Generally, teachers who created benchmark assessments for their grade levels or departments tended to view the experience favorably. Sarah, a science teacher who had built several assessments for her department, related:

> It made me understand how to work the tool better. We had done the quizzes and things the year before, but being able to go through and do more than 10 questions or 20 questions made me more aware of the different levels of knowledge, the

> organizing those different skills that they had to do. I was finding, as I was doing it, I was looking at those things to make sure I had a good mix.

The process of creating benchmark assessments for the district provided Sarah and other teachers with the opportunity to develop their technical skills in developing assessments. In addition, developing the benchmark assessments enabled teachers to relate their instructional practices with their assessment practices, as Ida, a veteran elementary teacher conveyed, "We get to see some of the questions, and we can realize, 'Gee, maybe we need to add a few of these kinds of problems into our lesson plans.'" More than one participant acknowledged that since the implementation of the FABA benchmark assessment program, they developed their teacher-made classroom tests differently, with a greater focus on test format and on test item level of difficulty.

Teachers also reported greater attention to their pacing guides since the advent of the FABA system. Hannah, a young and energetic elementary teacher, who had created benchmark assessments the previous year for her grade block, enthused:

> I knew exactly what was going to be on the test, especially for math. I was able to say . . . 'make sure that they know how many sides a hexagon has,' or whatever. I can't tell her [another teacher] everything, but I can tell her, 'make sure they know the shapes,' or 'make sure they know what line segments are.'

The comments of participants who engaged in the teacher-developed benchmark assessments demonstrated a deeper understanding of the curriculum standards and their grade block or department pacing guide.

The theme of Data Literacy is one of the more significant themes to provide insight to research question 1. In describing the district's benchmark assessment program, the participants often discussed topics related to using or analyzing data. Most

of the participants related that they felt relatively comfortable with analyzing data. Teachers, with the exception of the special education teacher, tended to feel slightly more comfortable with data than administrators. The special education teacher scored higher because of his constant use of various types of data to monitor the students on his case load. The researcher used Dedoose's coding weight function to assign each participant to a number on the scale. If a participant's transcript indicated that he or she felt comfortable with data, a five was recorded. If a participant reported higher levels of comfort an eight or nine was recorded, and a two or three was recorded for lower levels of comfort with using data. Associating a weight to a particular code allowed the researcher to more seamlessly blend the qualitative and quantitative aspects of this mixed methods study. Figure 3 indicates the data comfort level by participant role.

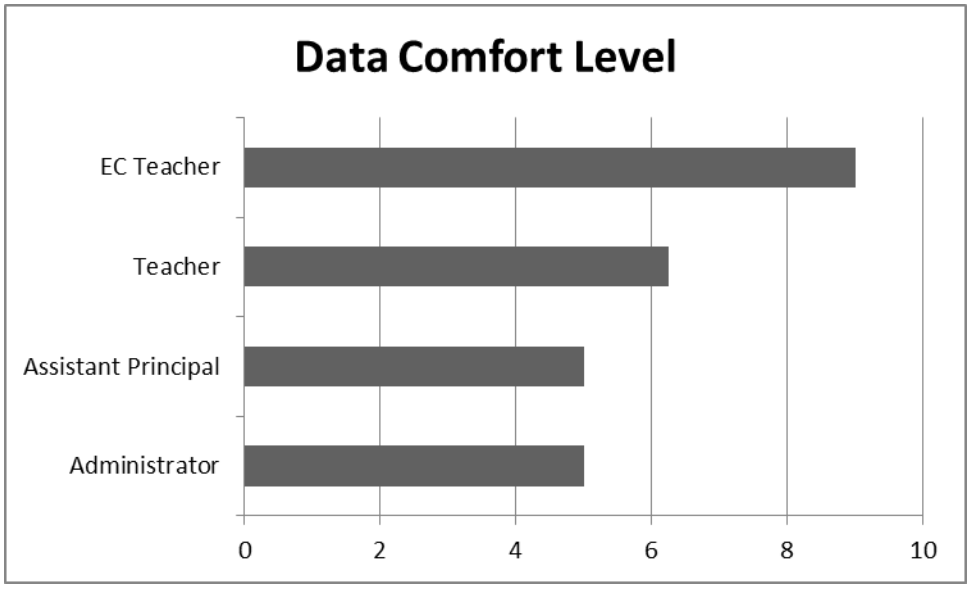

Figure 3. Data comfort levels by participant role.

The administrators reported having received training in using data as part of their Master's program preparing them as school administrators. Most reported additional professional development they had received at the district. Two of the administrators related that much of what they learned resulted from being on the job or came via a mentor relationship. Uma, an administrator who has been in education for 15 years, talked about a superintendent in her graduate program back in the mid-1990s who told her class that data would drive schools, a shift that she saw when she graduated and returned to work. Uma also shared that she had a mentor that taught her a great deal about using data to improve schools. She remarked:

> I was fortunate enough to work with one principal – a very, very successful high school principal in the state – and he got it. He got it clearly and was pushing his teachers. You've got to look at this. You've got to look at this.

Teacher participants did not relate any type of mentoring stories, probably due to their relative youth in the profession. With more experience and more opportunities to network, these relationships may develop.

Although teachers, on the whole, had comfort levels slightly higher than the administrator group, some teacher participants reported not having received any formal training in their teacher education programs. Most of the teacher participants discussed district-provided professional development on data use. Roxane, who came to education via the business field disclosed:

> I don't think I've had any training in the education system. Before this, I was a claims manager for a Fortune 500 company branch manager. We collected a lot of data so I 'm used to manipulating data and trying to reach performance goals based on data. I think my background helped me with that. I don't think I've had any training in data in education, either in the school or in teacher preparation.

Roxane was hired at the district the year after it focused primarily on data. Many of the participants who were working with the district during its data focus recalled the training the district provided on data through InformEd, consultants who worked with the district's teachers and administrators on data and school improvement. According to Sam, a middle school teacher:

> Two years ago we had those . . . ladies that came in. Yes, they came in and broke down test scores and benchmark scores and showed us what needs to be focused on. They were here for a year and half, two years.

The district contracted with InformEd for an entire school year, beginning their work with the administrative staff during the summer months. At the beginning of the school year in August, the consultants worked with teachers from each of the schools and with the administrators. After the initial workshop with the teachers, an administrative decision allowed the elementary school to opt out of the remaining workshops. The consultants focused the remainder of their time that year with the middle and high school teachers. All teachers in the district, regardless of elementary or secondary level, received the basic training in analyzing test data. Some participants saw the benefit in the workshops more than others. As with many professional development opportunities, the participants possessed different skill levels with the topic, so pacing was problematic.

One administrator revealed the truth that many schools and districts face when evaluating their programs. Fidelity is an issue that all districts face. Irene, who had been with the district for many years as a teacher prior to becoming an assistant principal, articulated this idea well:

> A couple years ago, we had the training here and just the simple fact that in a way it seemed more general, it was very simple in nature. Let's group these kids and

> look at the ones who have done well. Let's look at the ones who are borderline and as we were sitting there, a few teachers were, 'We could have done this ourselves.' My question was then, I think I even posed it, 'Are we though? We can, but are we?' I don't think we had been doing a very good job of that. We hadn't actually looked at the data in that perspective. We were very,'well, 75% passed an EOC'. Then we move on. Seventy-five percent were proficient, 25% weren't. We just look at it and move on. We can't move on if we look at it that way.

Irene's comments indicated her awareness that teachers should have been doing this type of analysis all along, but they had not been. Her comments also revealed that teachers often find school and district mandated professional development an ineffective use of their time. Irene's experience indicated that teachers have a difficult time maintaining a positive attitude in these situations.

On the whole, the participants use the data from the benchmark assessments to determine where student strengths and weakness lie. Teachers focus primarily on the areas showing the most need. This process follows the one that the InformEd consultants shared with the district's teachers and administrators previously. It also follows the FABA data training provided to teachers. Most teachers at the very least noted the areas of concern. Several reported conducting more sophisticated and deeper analyses to aid them. Several used color-coding techniques to help with organization, and others looked at specific items from the assessments to determine where students' understanding of the material was lacking. Wanda, a high school teacher, looks at her data to determine "if it [an objective] is something that the whole class needs to work on or if it's something that a few students are missing." In addition, Wanda shares her data with other members of her department and elicits their ideas on how to re-teach a concept. She spoke to her

colleagues about "how they got students to get something my students were still struggling with, to get different ideas from them."

Most teachers reported a process similar to Wanda's story. A few teachers, notably Xavier, a middle school teacher, discussed other data analysis techniques such as creating color-coded spreadsheets and posters for his students with current performance and mastery performance indicators. On the other end of the spectrum, one teacher reported that she did not do much with her benchmark data. "Just look at your benchmarks. When you get your results, reteach the objective," said Rachel, describing her data analysis process.

The third dimension, Instructional Practice, also sheds light on research question one, as to the type of benchmark assessment program the district has implemented. Though most of the participants understand the basic strategies of data analysis, whether they are superstars such as Xavier or less enthused adherents such as Rachel, most expressed some frustration with what to do once the data had been analyzed. For some, a state of almost paralysis had been reached. This stage of the process is the data analysis cliff. Teachers have their data, they have analyzed it, they know where the deficiencies are, and the question is now, what? The 'now, what' is the cliff that teachers and administrators must manage as they begin to make decisions that impact instructional practice. This is a difficult stage to maneuver on two fronts. One is the problem of knowing a different way to re-teach or present information. Another is the issue of managing a classroom in a manner that may not be comfortable for the teacher.

In this district, teachers often talked about ‘re-teaching’ or ‘remediating’ an objective or concept. Wanda expresses it this way, “I think the biggest thing is the re-teaching. Especially if it’s something that I thought that the students had.” Ida echoed this idea, “Remediation was basically what the whole class seemed to be missing.” Few teachers discussed how their re-teaching or remediation efforts looked different from their original teaching of the concept or objective. The impression that many of the participants left was that the re-teaching was a review of the same lesson previously taught. Sarah is forthright in her self-assessment of re-teaching, “As far as re-teaching, I probably don’t do as good a job as I need to do, going back and presenting it in a different way so that they can try to understand it.”

Though some teachers discussed small group instruction, and re-grouping students after a benchmark assessment had been administered, few teachers or administrators related examples of its occurrence. Sarah from the high school expressed her frustration at this component of the program:

> I feel like I need more instruction on what to do with it [the data] after I look at it. So, I have a hard time with differentiation. I don’t know if it’s the thought of trying to do more than one thing at one time that bothers me or if I just don’t know exactly how to properly execute it. So, I’m still working on that, been to a couple of workshops. [Her mentor] has been helping me do some different things. I’m getting better; we’re not sitting in our seats all of the time. If I had a little more instruction on how to properly use the data that would help me with my differentiation as well.

Sarah’s comments indicate that she reflects on her teaching practice, and she has identified her own needs. The frustration is the lack of access for appropriate professional development to assist her with the skills she needs to be effective with her students.

Teachers spoke of an instructional paralysis resulting from benchmark assessment data that indicated students were struggling with many objectives. How will this data affect a teacher's instructional practice? How does a teacher decide on which objectives to focus? A common refrain throughout the transcripts is that the benchmark assessment questions are more difficult than the state End-of-Grade (EOG) or End-of-Course (EOC) assessments, and the students' scores on the benchmark assessment are typically low on all objectives. Ida expressed her discouragement, "you see that a kid is missing a variety of questions. It's almost disheartening to think, 'I have to re-teach everything.'"

Teachers at the elementary and high schools were responsible for administering the assessments to their students. The middle school scheduled their assessments differently, due to the lack of a general use computer lab and the desire to protect instructional time. Ralph explained the decision:

> In the middle school, the biggest weakness I can see is--and it was my decision to do it this way, but it's protecting class time--was having a computer teacher to administer the benchmark test, and I think students do better when their regular teacher is there. But because of a time thing, we didn't do that.

Students were administered the benchmark assessments by their computer skills teachers during their computer skills class time, creating a disconnect between the content teacher and the benchmark assessment. Some teachers navigated this well, accessing scores and beginning their analysis immediately. For others the disconnect was more difficult to negotiate.

In summary, the district benchmark assessment program involved professional development on data use and the FABA product. Teachers typically developed the benchmark assessments based on the pacing guides. Teachers analyzed the data once the

assessments were given, but the consistency of the data analysis across all schools was questionable. Teachers typically re-teach or remediate students based on the data from the benchmark assessments. Much of the post-assessment instruction is whole group, although some teachers provided small group or individual instruction.

Research Question Two: What are the results of the district's benchmark assessment program? Research question two focuses on the outcomes of the district's benchmark assessment program described in research question one. The Data Literacy, Professional Development, and Program Effectiveness themes run strongly throughout the exploration of research question two. Although the district was struggling to meet its Annual Measureable Objectives (AMO) for NCLB accountability, the participants believed that the benchmark assessment program provided them with good results and was integral to their schools. Several participants realized that the school was not reaping all the potential benefits of the program, but few articulated why this was the case. Ralph, principal, alluded to the situation, "It's [the benchmark assessment program] very useful, and dependent on how much we require the teachers to use it, or how important from the administrative level that we deem it to be." Ralph understood that without administrative support and formal structure from the district, the schools may not reap the full benefits of the program.

Generally, participants were happy with the type of results they received from the benchmark assessment program. Most felt that the data was beneficial and provided them with the direction for their instruction. Teachers appreciated knowing where their

students were in regard to their curriculum early enough in the school year to help students who still needed to master specific goals or objectives.

Though most participants were comfortable with the results of the benchmark assessment system, some English Language Arts (ELA) teachers were frustrated with the program. Some of the elementary teachers believed the ELA data might not be particularly helpful to a classroom teacher. Hannah, a third grade teacher whose class was mostly reading below grade level stated:

> Now if my kids were all on grade level and could read those [text passages], I think it would be wonderful because I could see what they didn't get. But I feel like I could pull up a college textbook online, and they would do the same thing as they would with the FABA third grade reading. They're not going to try; they're going to get overwhelmed.

FABA reading items are written on grade level, and the program is not computer adaptive. Students who cannot read on grade level may struggle with the reading passages, as well as the test items.

The dissatisfaction with English Language Arts (ELA) benchmark assessments was not confined to the lower grades. High school teacher Roxane expressed her concern:

> I'm not sure if this is a weakness of the system. I teach English. The English goals are very broad in that they don't [pause] I've seen the biology goals and what the social studies teachers are supposed to teach under the previous course of study, and they're very specific. It would say something like 'Students can name five causes of the American Revolution.' FABA could come back and say 'they [students] couldn't do this,' so I can go back and teach this one thing. English is more a set of skills that is more broadly applied, and I found it very difficult to look at the benchmark assessment and say, 'Ok, this means I need to do this particular thing some more.' Even the grammar things, which are more specific and more particular than the others. There's still a huge category of types of grammar there are in each of the goals. Even if I know that I did poorly on this

> particular goal, I don't know which thing in the goal they did poorly on. It makes it hard for me to go back. . . .

Roxane's words gave voice to her frustration and her realization that the non-specificity of the goals was a characteristic of the state's curriculum, rather than a weakness of the benchmark assessment system.

Another result of the implementation of the benchmark assessment program was a focus on data. In addition to professional development on the FABA system itself, the district began a data initiative by providing its administrators and faculty with workshops on how to use data for school improvement. Ralph, who has spent much of his career as an administrator, understood the importance of building a data culture at his school:

> I support any workshop, or any staff development that we can bring to school, or when feasible, send teachers to take appropriate staff development. It's very important and if, as we evolve as a school, we need a core group of teachers who are really very good with data and how to apply that.

To begin working toward this type of data culture the district contracted with a consulting group, InformEd, to provide a series of onsite workshops. The training began in the summer for administrators, and the consultants provided hands-on training to teachers at each school at the beginning of the school year. The training for teachers focused on the EOG and EOC test scores from the previous school year. The consultants returned later in the year to work with middle and high school teachers, with a focus on using the same data analysis strategies with benchmark assessment scores. The elementary administrator chose to focus the school's professional development time in another area, so the elementary teachers were not a part of the follow-up data training. Administrators also continued with follow-up training and consultations during the remainder of the year.

The initial data training was fairly well received by the participants, though some expressed frustration with the pacing and simplicity of the content. The frustration was expressed as the weariness many teachers feel when confronted with another workshop that they do not see as being relevant to their work. This is especially true at the beginning of the school year when teachers long to be in their classrooms preparing for the new school year. Rachel expressed her dissatisfaction with the workshop by saying, "I haven't had that much training on how to use data here. We've had some workshops, but nothing I can remember of any use."

The content of the workshops did contain basic data analysis strategies (e.g., organizing objectives/goals by performance, organizing student performance by objective/goal, color-coding and grouping), and because of the size of the groups, the pace moved slowly. According to Patricia, an elementary participant, "They [the district] never got them [the consultants] back in and that was only with EOG data from the year before, and it was very dry and long and boring. People tend to tune out with things like that." Being an elementary teacher, Patricia may not have realized that the consultants did return for follow-up workshops at the other schools. She did, however, identify the need for continued support and a brisk pace during professional development.

At no point during the training or its follow-up was an expectation formally stated by the district that its teachers should engage in this type of data analysis with each of its benchmark assessments. A protocol was not instituted for teachers and principals to utilize with benchmark assessment data, and thus, an effective monitoring tool was not

available. This is not to say that some teachers were not implementing the strategies from the data workshops. Indeed, Xavier shared his post-workshop classroom activities:

> They [the data consultants] had us all look at our data and our information and we made posters. I used my poster with my kids and said, 'Here's your benchmark score. This isn't where we want to be. Let's color-code it for a certain percentage. This is what mastery looks like. This is what a passing score would be.' I had that in my room and every benchmark, I'd stick it up there. That was some more training on the data.

Most of the participants spoke of using data, both benchmark assessment data and other data, but none mentioned a continued use of the exact strategies taught in the InformEd data workshops.

The data training did not contain strategies for addressing the learning deficiencies found through the analysis of the benchmark assessment data. Nor did it address what teachers might implement for students who were excelling. Uma, an elementary administrator, spoke of the need to go beyond data analysis strategies. She suggested:

> I also think we need even more to branch off of 'what do you do with the data now that we have it?' I think the teachers maybe had some data before, and didn't know what to do with it. For me, it's all about, now that I have this, what do I do with it? It's all about, what do I do next? And how do I re-teach this? And how do I re-group?

Teachers and principals recognize that their benchmark assessment data should drive the instruction being delivered in the classroom. A few teachers in the district reported they use data consistently to drive their instruction. Most teachers recognized the need for differentiation based on their data, but they a need for help in making differentiation work for them. The interviews revealed reliance on re-teaching and remediation, but without an emphasis on how that instruction differed from the initial instruction. In

addition, they expressed needs for more support and training on strategy instruction so that they could re-teach the content in a different way. Roxane, a high school teacher, was very specific with the type of support she needed to allow data to drive her instruction:

> Let me sit down with someone who knows how to use this effectively, look at actual results and have that person lead me through how I can follow up on those results to do better instruction. I'm not sure sitting in a classroom with a bunch of teachers from a bunch of different disciplines would be helpful. I think it's probably something that would need to be at least department by department.

Many teachers echoed Roxane's frustration and expressed a similar desire for professional development on strategy instruction.

Another result of the district's benchmark assessment program is its accuracy. Participants from each of the schools believed the FABA product to be helpful and effective in their efforts to improve student achievement. Both administrators and teachers believe the product to be completely aligned with the state's curriculum. The strong alignment is due to the product being developed specifically for the state's schools by an organization working with the state's Department of Public Instruction. Most benchmark assessment products are developed by national companies who must satisfy the demands of many states, so although the products may align, the alignment may not be as strong as what the participants found with the FABA product.

In addition, the FABA developers also work closely with the Department of Public Instruction in test item development. FABA test items undergo a rigorous process before becoming a part of either the formative database or the benchmarking database. Participants believed the test items to be high quality and formatted similarly to the type

of question found on the state assessments. Several participants remarked that they believed the FABA test items to be more difficult than the items students faced on the state assessments. Wanda from the high school explained one reason the questions were harder for students:

> The questions require more than just basic knowledge. A lot of times they're two parts, sometimes three parts that they have to do. I think it allows the students to see that just because they get an answer that's on there, they have to make sure it's the answer that they're actually looking for.

Participants believed this difficulty accounted for the low FABA scores. Students could score low on the benchmark assessments, but still perform at the proficient level on the state assessment.

Exploration of research question two indicates that the benchmark assessment program provides quality data (Program Effectiveness) that teachers can utilize to direct their instruction. The program has provided professional development on data analysis, giving teachers the necessary skills to conduct the analyses. Less differentiation is occurring because teachers feel less confident about its implementation, expressing a need for professional development on the topic.

Research Question 3: What are the benefits of the benchmark assessment program to the school community? Research question three focuses on the benefits and value of the district's benchmark assessment program, and is where the Program Effectiveness theme is found most strongly. Overall, the participants viewed the benchmark assessment program positively and understood the benefits of the program to the students and to themselves. A major benefit to the district—and a topic covered earlier in the chapter—is the alignment of the assessments to the state's curriculum. Uma

was enthusiastic about this aspect of the program, "That's what it was developed for specifically—not any other state. So the data that we get from that is very accurate regarding our state tests." Faculty and administrators have also benefited by receiving technical and data analysis professional development, topics discussed previously.

The program in its current state is providing benefits to the system, according to the participants, and most feel that it is effective for them. Several of the participants discussed technical aspects of FABA that provided them and their students with direct and indirect benefits. For example, most participants believed that FABA is easy to use and provides results in a timely manner. Teachers appreciate that the benchmark assessments are developed for them and are scored electronically which saves them time. Jaclyn, a teacher in the elementary school, was especially positive about the computerized scoring, "I like the fact that it's on the computer and that it's graded in the computer. The teachers don't have to deal with sorting through all of that paperwork." Students benefit by completing assessments that are aligned to their curricula and state assessments. Test items are comparable to what they will encounter on the end-of-year or end-of-course assessment.

One area that participants identified as both a benefit and disadvantage is the online format of the benchmark assessments. High school students benefit greatly by having their benchmark assessment delivered in the same format as their state assessments, which are online assessments. One high school teacher, Roxane, believed that the online delivery was a significant boon to students. She expressed, "I think the activity of taking the test is a lot more useful than the results it generates." Though the

state assessments at the elementary and middle school levels are not yet online, the state plans to implement online assessments at those levels in a few years. Having students complete their benchmark assessments online now will prepare them for the state assessments' move to the online format.

However, some of the teachers believe that the online testing is a distraction from the content being tested. Wanda, a high school teacher, said of her students, "a lot of times, especially for the first one, students seem to just rush through it because they think they're going to get to play games or something afterward." Elementary participants feel they need to teach their students about computers and how to take a test online before they can administer the benchmark assessment. Patricia shared this about her students:

> It's really hard for fourth graders to take it on the computer sometimes. They're not used to doing that, and so, I think sometimes they get more easily distracted than if they had pen and paper in front of them.

Though she views the online delivery and scoring mostly as a benefit, Jaclyn did relate one particular difficulty she and her students experience, "A weakness would be the logging in for, again, the younger grades. It's just really difficult for them to get that all in. I know it's not the benchmark program itself that's the weakness." Several of the elementary participants believe the results from the early benchmark assessments may be skewed simply because students have difficulty navigating in the online environment or students are distracted by the opportunity to use a computer. In addition, computer access at the elementary school is not as available as it is at the middle and high school levels. Despite this downside to the program, most participants preferred the online medium.

Possibly the most significant benefit that the participants mentioned is that the benchmark assessment program provides them with a good, early indication of how students will perform on the EOG or EOC, allowing them the necessary time to make adjustments to their instruction. Wendy, an assistant principal, explained that the assessments "provide them with data that helps them evaluate where they go next in their instructional practices, as well as pinpoint those pocket areas where there's severe gaps in the learning, and they can adjust accordingly." Irene, an assistant principal, also recognized this benefit to the school, saying, "It helps determine what's going to happen between that benchmark and the next benchmark."

Though a cut score has not been available to them for making EOG and EOC predictions, most teachers have developed a 'feel' for what range of scores represents a proficient score on the subsequent state assessment. Wanda, a high school teacher, put forth her approximation:

> I've generally found if students are 60% or above with the FABA material, they generally have the concept, understand it. I've talked with a science teacher that has used FABA a lot more before she came here. That's generally what she's found too.

A teacher at the elementary school uses the same estimation. Hannah shared:

> Someone told me if they make above 60%, that's a good indicator they're going to make a[n achievement level] three on the EOG. If not, then they are having problems. Now, in my experience, there are a few kids who made below 60% who still passed, and I've had a few kids who did not make that grade, but didn't pass the EOG. Overall, I thought that was a good indicator.

Teacher participants obviously attempted to make a connection between students' performance on the benchmark assessments and their subsequent state assessment scores,

but administrator participants were also interested in this information. Wendy expressed the need this way:

> But when we're looking toward achievement levels as the end goal, that is still the one that, that we all kind of go, well, we have to figure out roughly what that would be if that were an achievement level. That creates some subjectivity with it.

Although teachers do not have a prediction model available to them, the program still provides them with solid data regarding what students know and do not know.

A related benefit is the ability of teachers to use this early performance data to identify and assist those students whose learning struggles might be masked. Some students who fail to grasp a new concept may be able to cover it up through effort or cheating. With performance data on individual students made available at designated times of the year, student learning deficiencies are less likely to go unnoticed by the teacher. Fourth grade teacher Patricia believed that the assessments:

> Tell me what's going on in my classroom because there are those kids who fall through the crack, and that you think they've got it, but when push comes to shove, maybe they're looking on their neighbors paper and you're working with somebody independently over here. You don't necessarily see that so you think, 'Oh, they've got this, so I'm good.' I find it really useful.

Because the district's benchmark assessment program provides three assessments to students throughout the school year, if a student were to "fall through the crack," it is likely that the teacher would catch the student on one of the successive assessments.

The district's benchmark assessment program created an environment that allowed teachers to become more reflective about their teaching practice, which is both an individual benefit as well as a school benefit. Uma, a principal, enthused, "I also think it's a very useful tool for teachers to see where they are and what they need to go back

and look at and reflect upon." Patricia, a teacher, explained how she used the benchmark assessment data for reflection:

> I have gone back and rethought how I taught something. Did I use enough manipulatives? Did I use the SMART Board enough? Were my lessons more engaging or do they need to be more engaging? Are the students having enough time to practice these skills or are we just hitting them quickly and moving on? Then if I get my benchmark data back and as a whole the class has really messed up some part of whatever objective, I put whatever I'm supposed to be teaching on hold, and we go back and hit that. So that way I'm not looking at it, 'Oh, they're missing this? Oh, well.'

This self-reflection, coupled with the functionality of the FABA program, has moved participants to engage in formative assessment, another benefit of the benchmark assessment system. FABA provides teachers with a database of test items that can be used to generate brief quizzes on specific objectives from the curricula. Pre-made quizzes are also available, allowing teachers to administer pre- and post-assessments to their students. Sarah, a science teacher, used the FABA system formatively:

> I try to use the FABA as much as possible, not just for the benchmarks, but I like to try to do it once a week. If we're still on the same topic or concept from one week to the next I will not repeat it. I probably should, to see if there's any growth but in the past I haven't done it. I'm not sure how to get the data out of the tests that I give them that are paper/pencil. With the FABA, I can use their tools to do it.

When teachers formatively assess their students, as Sarah does, then tracking mastery becomes much easier. Jaclyn, an elementary teacher, feels strongly about tracking her students' learning:

> I think a teacher is only successful if they are tracking their student data to show growth and to show what the students have mastered and what they need help with. I use it as an assessment tool, and I think that all teachers need to be assessing their kids so they have a better understanding of where the students are and where they need to be.

Jaclyn's statements indicated that she has grasped the power of formative and benchmark assessments for improving student learning and for demonstrating her effectiveness as a teacher.

The district benefits from its benchmark assessment program because of its ease of use and its formative assessment component. These elements provide teachers with the opportunity and the data to reflect upon their practice, and through that reflection, focus more on their students' learning. All of these aspects of the district benchmark assessment program are woven into the Program Effectiveness theme.

Summary

The district has implemented a benchmark assessment program that delivered the assessments online and allowed teachers to access their student data through the site. The district has also afforded its administrators and teachers with professional development on the benchmark assessment system itself and on data analysis strategies. Several of the participants expressed a need for additional professional development on differentiation. Although the participants were mostly satisfied with the type of results they obtained through the implementation of the benchmark assessment program, several of them acknowledged, whether overtly or tacitly, that the district was not realizing the full promise of the program. Ralph, a principal, phrased it succinctly, "I guess I would just like to see all of the teachers using what we have right now and become proficient at that. I think that would be a monumental thing for us here."

The coded transcripts revealed a key category or theme of Program Effectiveness. This category became the central phenomenon and is supported by other themes gleaned

from the data. Three themes that support and build Program Effectiveness are Assessment Literacy, Data Literacy, and Instructional Practice. Undergirding these categories is Professional Development. These five categories are instrumental in discovering responses to the three qualitative research questions. Though each of the themes might be found in each of the questions, some of the themes were more prevalent in answering specific research questions. The matrix in Table 5 indicates which themes were predominant in the exploration of each of the research questions.

Table 5

Themes Associated with Research Question

	Research Question 1	Research Question 2	Research Question 3
Assessment Literacy	X		
Data Literacy	X	X	
Instructional Practice	X	X	
Professional Development	X	X	
Program Effectiveness		X	X

Chapter 5

Quantitative Results

Introduction

This chapter will consider the quantitative data collection and analysis of the mixed methods study, focusing on the final research question: Research question four: Do benchmark assessment scores predict End of Grade and End-of-Course assessment scores, and what are the implications if the scores predict well or fail to predict well? This chapter will discuss the preliminary analysis (i.e., data screening, descriptive statistics) and then the main statistical analysis for each data set.

Data Collection

The initial step in data collection involved gathering many different test score files from the school district. Grades three, four, six, and seven each produced four data files for reading (i.e., benchmark assessment 1, benchmark assessment 2, benchmark assessment 3, and End-of-Grade (EOG)) and four similar files for math. Thus, for each of these four grades, eight original data files were consolidated into two files, one for reading and one for math. Grades 5 and 8 included the same reading and math files, and these two grades levels produced four additional files for science (i.e., benchmark assessment 1, benchmark assessment 2, benchmark assessment 3, and an EOG), resulting in three consolidated files for each grade level, one each for reading, math, and science. For the high school subjects of Algebra I and Biology, only two benchmark assessments were administered, and those two files were consolidated with the End-of-Course (EOC) score file to produce one file for each of those subjects. Three benchmark assessments

were administered for English I, which was taught as a year-long course, producing three data files. These files were consolidated with the End-of-Course file.

Student identifiers were removed from the consolidated files, and incomplete records were deleted. Deleting the incomplete records lowered the sample size for each grade, but doing so provided a degree of control for an external validity risk. Table 6 lists the sample size for each grade/subject. Using only records with a complete data set (i.e., all benchmark assessment scores and an EOG or EOC score) increased the likelihood that the results could later be generalized.

Table 6

Sample Size by Grade and Subject

Grade	Subject	Number in Sample
3	Reading	59
	Math	40
4	Reading	56
	Math	61
5	Reading	73
	Math	51
	Science	62
6	Reading	61
	Math	61
7	Reading	52
	Math	57
8	Reading	79
	Math	76
	Science	64
High School (primarily Grade 9 & 10)	Algebra I	44
	Biology	68
	English I	52

Statistical Analysis

The researcher utilized the services of the Nebraska Education and Research (NEAR) Center for help with computing descriptive statistics, correlations, and regression statistics for each of the 18 consolidated data files. These statistics allow for the examination of the relationship between the criterion (EOG or EOC score) and its predictors (benchmark assessment scores). As stated previously, after removing incomplete records from the data files, the sample size decreased. To compensate, the adjusted R^2 will be reported which is often used when dealing with small sample sizes and multicollinearity issues (Newsom, 1999-2007, ¶ 1).

Elementary Results

Grade 3 reading and math. The results from the statistical analyses conducted on the Grade three reading and math data from the district's elementary school are shown in Table 7. Mean, standard deviation, and correlations were calculated for each subject area. Additionally, the statistical analyses included multiple regression calculations for reading and math scores.

The initial analyses of Grade 3 reading indicated that the three benchmark assessment scores collectively are significant predictors of the reading End-of-Grade (EOG) score. Benchmark 1 and 2 are moderately correlated with the EOG, and Benchmark 3 strongly correlated with the EOG. Table 7 displays the results of the descriptive statistics, the correlation, and the multiple regression weights of each of the benchmark assessments with the EOG for grades 3 - 5. The correlations for grade 3 reading indicate that students with higher scores on the benchmark assessment 3 variable

Table 7

Descriptive Statistics, Correlations, and Beta Coefficients for Grade 3

					Multiple Regression Weights	
	Variable	Mean	Standard Deviation	Correlation with EOG	b	β
G3 Reading	EOG	335.97	9.193			
	BM 1	36.25	12.363	.566***	.107	.144
	BM 2	44.46	15.799	.609***	.084	.145
	BM 3	54.34	16.104	.791***	.352***	.617***
G3 Math	EOG	341.18	8.80			
	BM1	47.38	16.19	.645***	.101	.185
	BM2	51.88	13.19	.719***	.177	.265
	BM3	53.20	11.12	.745***	.313	.395

Note. G = grade level; BM = benchmark assessment; EOG statistics were obtained from state score reports.
*** $p < .001$

tended to have higher EOG scores. The multiple regression model with all three reading predictors collectively resulted in adjusted $R^2 = .639$, $F(3, 55) = 35.219$, $p < .001$. The benchmark assessment program, as a whole, predicts well for reading EOG in grade 3. The regression weights indicate that students with higher benchmark assessment 3 scores were expected to score higher on the reading EOG assessment. Benchmark assessments 1 and 2 did not contribute to the multiple regression model.

The analysis of the grade three math data set indicated that the three benchmark assessments collectively are significant predictors of the subsequent math EOG score. Benchmarks 2 and 3 demonstrated fairly strong correlations with the EOG, but

benchmark 1 indicated only a moderate correlation. The correlations for grade three math, located in Table 7, indicate that students with higher scores on the benchmark assessment 2 and 3 variables were inclined to have higher math EOG scores. The multiple regression model for grade three math indicates that the three benchmarks collectively resulted in adjusted R^2 = .571, F(3, 39) = 18.328, p < .001. However, none of the three benchmarks contributed significantly to the model, according to the regression weight statistics. This could be due to multicollinearity error. Table 8 summarizes the correlations of the benchmark assessments to each other.

Table 8

Correlations between Grade 3 Math Benchmark Assessments (BM)

	BM1	BM2	BM3
BM1	1.000	.684***	.706***
BM2		1.000	.829***
BM3			1.000

Note. *** p < .001

To manage the multicollinearity issue, simple linear regression statistics were also computed on each of the predictors. The simple linear regression model demonstrated that each benchmark assessment predicted well for the math EOG assessment. For benchmark assessment 1, the linear model produced an adjusted R^2 = .401, F(1, 39) = 27.098, p < .001. The linear model produced an adjusted R^2 = .504, F(1, 39) = 40.689, p < .001, for benchmark assessment 2. For benchmark assessment 3,

the linear model produced an adjusted R^2 = .544, F(1, 39) = 47.523, p < .001. Because of the stronger adjusted R^2, benchmark 3 is the strongest predictor for grade 3 math.

Grade 4 reading and math. The results from the statistical analyses conducted on the Grade four reading and math data from the district's elementary school are shown in Table 9. Mean, standard deviation, and correlations were calculated for each subject area. Additionally, the statistical analyses included multiple regression calculations for reading and math scores.

Table 9

Descriptive Statistics, Correlations, and Beta Coefficients for Grade 4

					Multiple Regression Weights	
	Variable	Mean	Standard Deviation	Correlation with EOG	b	β
G4 Reading	EOG	341.68	7.561			
	BM 1	41.32	13.145	.564***	.056	.097
	BM 2	46.23	13.850	.718***	.169**	.309**
	BM 3	52.11	13.394	.758***	.276***	.488***
G4 Math	EOG	346.52	7.762			
	BM1	44.05	14.511	.744***	.158**	.295**
	BM2	49.92	14.719	.732***	.122*	.232*
	BM3	53.54	12.250	.791***	.267***	.421***

Note. G = grade level; BM = benchmark assessment; EOG statistics were obtained from state score reports.
*p < .05, **p < .01, *** p < .001

Table 9 indicates that the grade 4 reading benchmark assessments correlated significantly with the grade 4 reading EOG, with benchmark assessments 2 and 3 demonstrating strong correlations. Students with higher benchmark assessment scores could be expected to produce higher scores on the reading EOG. The multiple regression model with all three grade 4 reading benchmark assessments scores produced an adjusted $R^2 = .626$, $F(3, 55) = 31.715$, $p < .001$. Both benchmark assessment 2 and 3 contributed significantly to the model, according to the regression weights in Table 9.

The district's benchmark assessment program provides solid predictors for the grade 4 math EOG. The correlations for grade four math demonstrated moderately strong correlations between each of the benchmark assessments and the math EOG. Table 9 presents the descriptive statistics for the grade 4 math data sets, as well as their regression weights. The multiple regression formula for all three benchmarks produced an adjusted $R^2 = .708$, $F(3, 60) = 49.449$, $p < .001$. All three benchmark assessments provided a significant contribution to the prediction model.

Grade 5 reading, math, and science. The results from the statistical analyses conducted on the Grade five reading, math, and science data from the district's elementary school are shown in Table 10. Mean, standard deviation, and correlations were calculated for each subject area. Additionally, the statistical analyses included multiple regression calculations for reading, math, and science scores.

The grade 5 reading predictors were positively and moderately correlated with the reading EOG. The correlations indicate that students who scored well on benchmark assessments 2 and 3 were more likely to perform well on the reading EOG. Table 10

Table 10

Descriptive Statistics, Correlations, and Beta Coefficients for Grade 5

					Multiple Regression Weights	
	Variable	Mean	Standard Deviation	Correlation with EOG	b	β
G5 Reading	EOG	344.84	13.497			
	BM 1	43.27	14.687	.452***	.052	.056
	BM 2	43.67	15.764	.554***	.262*	.306*
	BM 3	5.84	15.795	.547***	.257*	.301*
G5 Math	EOG	349.43	26.776			
	BM1	50.22	14.795	.285*	.389	.215
	BM2	44.47	13.693	.165	-.032	-.017
	BM3	50.24	15.938	.255*	.219	.130
G5 Science	EOG	146.45	8.077			
	BM1	47.68	12.840	.529***	.085	.135
	BM2	37.79	11.401	.793***	.362***	.511***
	BM3	41.52	13.631	.737***	.199***	.337***

Note. G = grade level; BM = benchmark assessment; EOG statistics were obtained from state score reports.
*p < .05, *** p < .001

displays the descriptive statistics and analysis of the grade 5 reading data sets. The multiple regression model with all three predictors produced an adjusted R^2 = .332, F(3, 72) = 12.939, p < . 001. Table 10 shows that benchmark assessments 2 and 3 had significant regression weights, indicating that students with higher scores on these

benchmark assessments were expected to have higher reading EOG scores. The model signifies that the benchmark assessment program predicts well for grade 5 reading.

The grade 5 math statistical analysis produced anomalous results. Unlike the other grades and subject areas in the district's elementary school, grade 5 math data yielded much weaker correlations (Table 10) of the benchmark assessments with the math EOG. The multiple regression model produced an adjusted $R^2 = .034$, $F(3, 50) = 1.582$. The predictors explained very little of the variance in the EOG scores.

Grade 5 science benchmark assessments 2 and 3 correlated significantly with the science EOG (Table 10), meaning that students with higher scores on those predictors were expected to have higher science EOG scores. Benchmark assessment 1 produced a moderate correlation with the science EOG. The grade 5 science multiple regression model with all three predictors produced an adjusted $R^2 = .710$, $F(3, 61) = 50.849$, $p < .001$. According to the regression weights in Table 9, benchmark assessments 2 and 3 contributed significantly to the model, meaning that higher scores on benchmark assessments 2 and 3 were expected to produce higher science EOG scores. Benchmark assessment 1 did not contribute to the model.

Middle School Results

Grade 6 reading and math. The results of the statistical analyses performed on the reading and math data from the grade 6 data sets are shown in Table 11. Sample sizes can be found in Table 6. Mean, standard deviation, and correlations were calculated for each subject area. Additionally, the statistical analyses included multiple regression calculations for reading and math scores.

Table 11

Descriptive Statistics, Correlations, and Beta Coefficients for Grades 6

					Multiple Regression Weights	
	Variable	Mean	Standard Deviation	Correlation with EOG	b	β
G6 Reading	EOG	350.62	6.711			
	BM 1	53.41	20.465	.743***	.141**	.430**
	BM 2	43.84	17.129	.677***	.012	.032
	BM 3	41.51	15.125	.731***	.064*	.380*
G6 Math	EOG	353.66	28.559			
	BM1	44.56	12.618	.355**	-.017	-.008
	BM2	39.26	15.265	.312**	-.207	-.111
	BM3	3.30	13.501	.505***	1.250**	.591**

Note. G = grade level; BM = benchmark assessment; EOG statistics were obtained from state score reports. *p < .05, **p < .01, *** p < .001

As indicated in Table 11, correlations for the grade 6 reading benchmark assessments and the reading EOG were strong. Each benchmark assessment was a significant predictor for the reading EOG assessment. The model produced an adjusted $R^2 = .619$, $F(3, 60) = 30.859$, $p < .001$. The regression weights displayed in Table 11 indicate that students with higher scores on benchmark assessments 1 and 3 were expected to demonstrate higher reading EOG scores. However, each of the benchmark assessments also demonstrated a stronger correlation with each other than to the reading EOG scores. Table 12 summarizes the correlations of the benchmark assessments with each other.

Table 12

Correlations between Grade 6 Reading Benchmark Assessments (BM)

	BM1	BM2	BM3
BM1	1.000	.797***	.758***
BM2		1.000	.797***
BM3			1.000

Note. *** $p < .001$

The high degree of correlation between the benchmark assessments in grade 6 reading indicates a multicollinearity error, meaning that all three of the predictors are explaining the same amount of variance. If the benchmark assessments were too similar in design, then none of the assessments would contribute enough new information to be useful for prediction in a multiple regression model. Simple linear regression statistics were performed on each of the benchmark assessment data sets due to the multicollinearity. The simple linear regression for benchmark assessment 1 produced an adjusted $R^2 = .553$, $F(1, 61) = 72.855$, $p < 001$. The linear regression for benchmark assessment 2 resulted in an adjusted $R^2 = .445$, $F(1, 61) = 49.918$, $p < .001$. The benchmark assessment 3 linear regression model produced an adjusted $R^2 = .528$, $F(1, 61) = 69.367$, $p < .001$. All three reading benchmark assessments are good predictors of the reading EOG for grade 6.

Correlation and multiple regression analyses were also conducted with grade 6 math assessments. Benchmark assessment 3, as indicated in Table 11, has a moderately strong correlation with the math EOG. The regression model produced an adjusted

R^2 = .222, F(3, 60) = 6.707, p < .001, which would indicate the benchmark assessment program is not a strong predictor for grade 6 math EOG scores. However, similar to grade 6 reading, the benchmark assessment scores for math correlated highly with each other. Table 13 specifies the benchmark assessment correlations with each other for grade 6 math.

Table 13

Correlations between Grade 6 Math Benchmark Assessments (BM)

	BM1	BM2	BM3
BM1	1.000	.723***	.750***
BM2		1.000	.725***
BM3			1.000

Note. *** p < .001

Because of the multicollinearity within the grade 6 math benchmark assessments, linear regression statistics were conducted on each of the benchmark assessments separately. The model for math benchmark assessment 1 resulted in an adjusted R^2 = .111, F(1, 60) = 8.528, p < .05. Math benchmark assessment 2's model produced adjusted R^2 = .082, F(1, 60) = 6.365, p < .05. Though significant, benchmark 2 is only explaining 8% of the variance in the EOG scores. The model for math benchmark assessment 3 produced an adjusted R^2 = .242, F(1, 60) = 20.171, p < .001. Though benchmark assessment 3 explains more of the variance (24%) in the grade 6 math EOG scores than

the other two benchmark assessments, it does not provide enough information about the math EOG to make it a practical tool for prediction for an educator.

Grade 7 reading and math. The results of the statistical analyses performed on the reading and math data from the grade 7 data sets are shown in Table 14. Mean, standard deviation, and correlations were calculated for each subject area. Additionally, the statistical analyses included multiple regression calculations for reading and math scores.

Table 14

Descriptive Statistics, Correlations, and Beta Coefficients for Grades 7

					Multiple Regression Weights	
	Variable	Mean	Standard Deviation	Correlation with EOG	b	β
G7 Reading	EOG	351.8	7.976			
	BM 1	47.54	16.205	.711***	.256***	.520***
	BM 2	38.38	14.010	.537***	.109	.191
	BM 3	32.56	18.667	.500***	.060	.140
G7 Math	EOG	351.26	6.137			
	BM1	38.26	11.828	.739***	.269***	.518***
	BM2	36.77	10.884	.680***	.217***	.384***
	BM3	32.53	9.623	.458***	.023	.037

Note. G = grade level; BM = benchmark assessment; EOG statistics were obtained from state score reports.
*** $p < .001$

Correlations for the grade 7 reading assessments produced unexpected results. As Table 14 indicates, all three reading benchmark assessments correlated fairly strongly with the reading EOG in grade 7. However, the strongest correlation occurred with benchmark assessment 1, although the expectation was that benchmark assessment 3 would have provided the strongest correlation. The regression model indicated that the benchmark assessments, collectively, predicted well for the grade 7 reading EOG, producing an adjusted $R^2 = .514$, $F(3, 51) = 18.948$, $p < .001$. The regression weights indicate that benchmark 1 was the only significant predictor.

Analyzing each reading benchmark assessment individually indicated that each benchmark assessment was statistically significant as a predictor for the grade 7 reading EOG, but benchmark assessments 2 and 3 accounted for little of the variance in the EOG scores (27% and 24%, respectively). The model for benchmark assessment 1 produced an adjusted $R^2 = .495$, $F(1, 51) = 51.040$, $p < .001$. The model with benchmark assessment 2 singly resulted in an adjusted $R^2 = .274$, $F(1, 51) = 20.291$, $p < .001$, and the model for benchmark assessment 3 found an adjusted $R^2 = .235$, $F(1, 51) = 16.710$, $p < .001$.

The grade 7 math benchmark assessments performed similarly to the reading assessments in relation to the subsequent EOG. As Table 14 indicates, moderate to strong correlations were found between the benchmark assessments and the math EOG. Similar to grade 7 reading, math benchmark assessment 1 demonstrated the strongest correlation (.739) of the three tests. The multiple regression model with all three predictors resulted in an adjusted $R^2 = .642$, $F(3, 56) = 34.487$, $p < .001$. The regression

weights in Table 14 indicate that grade 7 math benchmark assessments 1 and 2 are significant predictors of the math EOG, and benchmark assessment 3 does not significantly contribute to the model.

Grade 8 reading, math, and science. The results from the statistical analyses conducted on the Grade eight reading, math, and science data from the district's middle school are shown in Table 15. Mean, standard deviation, and correlations were calculated for each subject area. Additionally, the statistical analyses included multiple regression calculations for reading, math, and science scores.

The grade 8 reading data sets produced moderate to strong correlations between the reading benchmark assessments and the reading EOG, as evidenced in Table 15. The multiple regression model resulted in an adjusted $R^2 = .515$, $F(3, 78) = 28.585$, $p < .001$. The regression weights, located in Table 15, indicate that benchmark assessments 1 and 3 are significant predictors for the reading EOG. However, a closer examination of the correlations indicates a multicollinearity issue. Table 16 displays the correlations between the benchmark assessments.

Simple linear regressions were computed for each of the reading benchmark assessment. The linear regression model for benchmark assessment 1 produced an adjusted $R^2 = .487$, $F(1, 78) = 75.185$, $p < .001$. Benchmark 2 linear regression analysis resulted in an adjusted $R^2 = .359$, $F(1, 78) = 44.740$. $p < .001$. The linear regression model for benchmark assessment 3 generated an adjusted $R^2 = .399$, $F(1, 78) = 52.762$, $p < .001$. Though each of the benchmark assessments are statistically significant as

Table 15

Descriptive Statistics, Correlations, and Beta Coefficients for Grades 8

					Multiple Regression Weights	
	Variable	Mean	Standard Deviation	Correlation with EOG	b	β
G8 Reading	EOG	356.68	7.667			
	BM 1	46.33	17.327	.708***	.199***	.449***
	BM 2	49.1	17.809	.606***	.046	.106
	BM 3	49.08	15.317	.638***	.121*	.241*
G8 Math	EOG	354.16	23.090			
	BM1	39.80	14.765	.276**	.154	.099
	BM2	47.68	17.271	.317**	.284	.212
	BM3	38.61	11.384	.284**	.147	.072
G8 Science	EOG	149.20	6.636			
	BM1	38.88	13.424	.605***	.191**	.386**
	BM2	36.69	13.784	.332**	.026	.054
	BM3	31.50	12.917	.595***	.179**	.349**

Note. G = grade level; BM = benchmark assessment; EOG statistics were obtained from state score reports.
*p < .05, **p < .01, *** p < .001

Table 16

Correlations between Grade 8 Reading Benchmark Assessments (BM)

	BM1	BM2	BM3
BM1	1.000	.751***	.724***
BM2		1.000	.677***
BM3			1.000

Note. *** $p < .001$

predictors, benchmark assessment 1 explains the largest portion of the variance (49%) on the subsequent reading EOG.

Table 15 demonstrates that the correlations between the grade 8 math benchmark assessments and the grade 8 math EOG are significant, but weak correlations. However, a closer analysis of the correlations between the benchmark assessments themselves indicates the presence of multicollinearity. Table 17 summarizes the correlations of the benchmark assessments with themselves.

Table 17

Correlations between Grade 8 Math Benchmark Assessments (BM)

	BM1	BM2	BM3
BM1	1.000	.584***	.739***
BM2		1.000	.653***
BM3			1.000

Note. *** $p < .001$

The correlations between the math benchmark assessments themselves are much stronger than the correlations between them and the math EOG.

The multiple regression model with all three grade 8 math benchmark assessments produced an adjusted R^2 = .078, F(3, 75) = 3.118, p < .05. As Table 15 displays, none of the benchmark assessments have significant regression weights, indicating they do not contribute to the multiple regression model. While the simple linear regression statistics for the benchmark assessments individually produce statistically significant results at the p < .05 level, the benchmark assessments account for very little of the variance in the math EOG scores. The linear regression model with benchmark assessment 1 produced an adjusted R^2 = .064, F(1, 75) = 6.099, p < .05. The benchmark assessment 2 linear regression model resulted in an adjusted R^2 = .088, F(1, 75) = 8.266, p < .05. The linear regression model with benchmark assessment 3 generated an adjusted R^2 = .068, F(1,75) = 6.481, p < .05.

The descriptive statistics and analysis results for grade 8 science are summarized in Table 15. The science benchmark assessments are positively and significantly correlated with the science EOG scores. These correlations indicate that students who score higher on the benchmark assessments are expected to produce higher science EOG scores. The multiple regression model with all the benchmark assessments resulted in an adjusted R^2 = .432, F(3, 63) = 16.957, p < .001. The regression weights found in Table 15 indicate that benchmark assessments 1 and 3 are significant predictors for the science EOG, and benchmark assessment 2 does not contribute to the multiple regression model.

High School Results

Only three areas in the high school have state assessments for NCLB accountability: Algebra I, Biology, and English I. These courses are not grade level dependent, although most of the students in Algebra I and English I are enrolled in grade 9, and most of the students in Biology are enrolled in grade 10 in this school district. Table 18 summarizes the descriptive statistics and the analysis results for the high school subjects.

Table 18

Descriptive Statistics, Correlations, and Beta Coefficients for High School Subject Areas

					Multiple Regression Weights	
	Variable	Mean	Standard Deviation	Correlation with EOG	b	β
Algebra I	EOG	145.57	9.754			
	BM 1	45.23	17.996	.523***	.176**	.325**
	BM 2	33.68	12.177	.659***	.433***	.540***
Biology	EOG	149.01	6.666			
	BM1	49.53	15.090	.673***	.181***	.410***
	BM2	41.91	13.064	.690***	.230***	.450***
English I	EOG	149.69	6.236			
	BM1	42.81	13.215	.576***	.230**	.488**
	BM2	41.35	14.459	.454***	.061	.142
	BM3	37.73	13.237	.330**	-.003	-.007

Note. G = grade level; BM = benchmark assessment; EOG statistics were obtained from state score reports.
**p < .01, *** p < .001

Notice that Algebra I and Biology have only two benchmark assessments because they are taught by semester. English I, however, is taught as a year-long course, and so it has three benchmark assessments.

Algebra I. As Table 18 indicates, the Algebra I benchmark assessments are positively and significantly correlated with the criterion, Algebra I EOC. Students with higher scores on the benchmark assessments are expected to produce higher Algebra I EOC scores. The multiple regression model with both predictors produced an adjusted R^2 = .503, $F(2, 43) = 22.772$, $p < .001$. The regression weights located in Table 18 indicate that both of the Algebra I benchmark assessments were significant predictors for the Algebra I EOG.

Biology. The science benchmark assessments also correlated positively and significantly with the Biology EOC scores, as summarized in Table 18. Higher science benchmark assessments scores tended to have higher EOC scores. The multiple regression model with the two predictors generated an adjusted R^2 = .573, $F(2, 67) = 45.999$, $p < .001$. The regression weights found in Table 18 indicate that both benchmark assessments are statistically significant contributors to the model.

English I. The English I benchmark assessment predictors are positively and significantly correlated with the English I EOC. Though significant, only benchmark assessment 1 demonstrates strength with the correlation and that only moderately so. The multiple regression model with the three predictors produced an adjusted R^2 = .302, $F(3, 51) = 8.354$, $p < .001$. As Table 18 summarizes, only benchmark assessment 1 had a

significant positive regression weight. Benchmark assessment 3 actually generated negative weight, but it was not significant.

Summary

With a few exceptions, the district's benchmark assessments generally correlate positively and significantly with the subsequent state EOG or EOC assessment. All of the multiple regression models predict well, except in grade 5 math. Fourteen of the 17 subjects had an adjusted R^2 equal to or greater than .300. The three areas with lower R^2 were grades 5, 6, and 8 math. Although multicollinearity occurred in some areas, simple linear regression analysis performed on the benchmark assessments individually indicated at least one strong predictor. Excluding the three grade levels in math previously noted, the benchmark assessment program predicts well for the district.

Chapter 6

Discussion

Introduction

This chapter will provide a discussion of the findings from the study. A model will be presented to indicate the significant components of a benchmark assessment program. This chapter will also provide an assessment of the significance of the findings, including implications and limitations of the study. Additionally, the chapter will include recommendations for future research.

Discussion

The purpose of the study was to explore the benchmark assessment system implemented in one school district, predominantly serving American Indian students. Four research questions (three qualitative, one quantitative) were posed:

1. Research question one: What is the benchmark assessment program utilized in a small district, serving a predominantly American Indian population?
2. Research question two: What are the results of the district's benchmark assessment program?
3. Research question three: What are the benefits of the benchmark assessment program to the school community?
4. Research question four: Do benchmark assessment scores predict End of Grade and End-of-Course assessment scores, and what are the implications if the scores predict well or fail to predict well?

Five themes were extracted from the teacher and administrator interviews which informed the first three research questions. The five themes included Professional Development, Assessment Literacy, Data Literacy, Instructional Practice, and Program Effectiveness. The interrelationships of these five themes were used to form the Dimensions of Benchmark Assessment Program Effectiveness model depicted in Figure 4, which evolved as suggested by Heppner and Heppner (2004) in their discussion of Straus's and Corbin's analysis strategies.

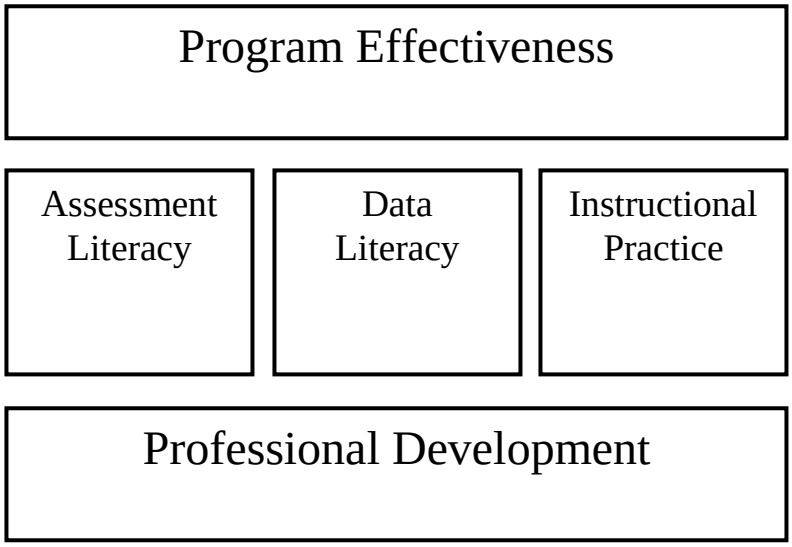

*Figure 1***.** Dimensions of benchmark assessment program effectiveness.

Professional development – Tier 1. Professional Development forms the foundation of the model. Without adequate professional development on the Tier 2 dimensions (Assessment Literacy, Data Literacy, and Instructional Practice) of the model, a district may realize a diminished Tier 3 (Program Effectiveness) or a less effective benchmark assessment program. All of the participants related taking part in

some of the professional development that the district provided to its teachers. With a few exceptions, teachers reported having training on how to use the FABA system, which focused primarily on the formative assessment part of the program. This training included basic functionality of the system (e.g., how to set up a class, how to create and schedule assessments), as well as information on how to access the reports and data. Initially, the district brought in trainers from the FABA program. Subsequent training was provided by district staff (the researcher) with a support person, who received additional training, available at each school for teachers to utilize for assistance. Unfortunately, the district did not always identify new teachers or teachers moving from an untested grade to a tested grade for training. Thus, a few teachers were left to learn the system through trial and error or through assistance from colleagues.

Teachers indicated familiarity with the FABA system and remarked on its ease of use. Most teachers used the benchmark assessment tool, and some used the formative assessment item banks as well. Fidelity to the use of FABA's benchmarking tool was moderately high.

The district's roll out of professional development on data analysis was targeted for all staff, so teachers were not overlooked as was the case with the FABA training. In adhering to the tenets of high quality professional development (Hassel, 1999; Learning Forward, 2012), the district provided the data analysis training systemically, and it was on-going. According to the Standards for Professional Learning found on the Learning Forward (2012) web site, "Learning designs that occur during the workday and engage peers in learning facilitate ongoing communication about learning, develop a

collaborative culture with peer accountability, foster professionalism, and support transfer of the learning to practice." Teachers were introduced to the concepts in their grade block or departmental teams, and then follow-up occurred at planned intervals throughout the year. However, not all of the district's teachers were provided with the follow-up training because of an administrative decision to spend professional development time in another area. Because of this change, the elementary teachers did not receive ongoing support for analyzing benchmark assessment data. The initial training had focused on using the analysis strategies with the previous year's End-of-Grade (EOG) and End-of-Course (EOC) scores, with subsequent sessions focusing on the benchmark assessment scores. There was no continued use of the analysis methods, nor was there a district-wide plan indicated that teachers were required to continue with the analysis methods. Though successful in two of the three schools, from a district-wide perspective, implementation was faulty.

Assessment literacy – Tier 2. Assessment Literacy is a Tier 2 dimension and is a vital component in an effective benchmark assessment program. This is an area in which the district appeared to be lacking. Participant interviews revealed a lack of understanding about several of the aspects of assessment literacy. Few teachers or administrators demonstrated a thorough understanding of the concept of assessment for learning (Popham, 2008; Stiggins & Duke 2008), of which formative assessment is an integral component. While most teachers understood the importance of looking at data and addressing needs instructionally -- other Tier 2 dimensions to be discussed later in the chapter -- few articulated the myriad of strategies through which this could be

accomplished, including the use of the formative component of the FABA program. Additionally, some participants, both teacher and administrator, did not initially understand the test item term 'constructed- response.' Once defined, though, the participants were able to discuss whether the addition of those types of items might be beneficial.

Interviews with the participants indicated that neither the schools nor the district had provided training on assessment literacy. The FABA and data analysis professional development provided by the district presupposed that the teachers and administrators were knowledgeable about assessment literacy. Often, districts rely on principals and other administrators to provide assessment literacy support to teachers, but as Stiggins and Duke (2008) related, few school administrator graduate programs provide the training that principals will need to later support teachers in this area.

The findings indicated the benchmark assessment data may not have been used formatively as often as the administrators felt it should be, suggesting the teachers' belief that the benchmark assessment program's purpose was summative in nature. Indeed, the district may not realize the "transformative" (Popham, 2008) nature of their benchmark assessment program because its teachers have not completely grasped the idea of formative assessment. Popham believes that typical commercial benchmark products are not formative because the data from them is not used to adjust instruction. However, in this study, each of the participants related a formative purpose for the benchmark assessments. Teacher remarks also indicated a belief that assessment is not part of instruction, but an additional requirement imposed from above. Often teachers who are

not comfortable with assessment being a part of instruction and learning believe that benchmark assessments are another program that is being 'done' to them by the district or school, and in turn, the benchmark assessments become something that teachers 'do' to their students. The teacher comments indicating a disconnect between what they understand is the purpose of the benchmark assessment program and what they actually do in the classroom emphasizes the need for additional training on assessment literacy, which would include the concept of assessment as part of a teacher's instructional practice.

Data literacy – Tier 2. The district recognized that if it wanted to promote a culture of data use in its schools, then it would need to provide teachers and administrators with training on its use. As indicated earlier in the chapter, all teachers and administrators were required to participate in the data training, but only the middle school and the high school received the follow-up training. Most participants acknowledged the helpfulness of the strategies given to them by the consultants, although some felt the training was tedious and not the best use of their time. Without training in how to use the data that results from a benchmark assessment program, teachers will find themselves drowning in numbers, but unable to make sense of them.

According to Marshall (2008), for a school district to experience success, its teachers and administrators must act on the data, suggesting that schools set SMART (Specific, Measureable, Attainable, Realistic, Time-sensitive) goals and convene data meetings. Teachers and administrators in the district have access to the data from the benchmark assessments, but they have not developed an adequate process for making the

data actionable. Wayman (2005) observed that this situation was typical for many school districts, and Olson (2005) remarked that the flaw in many districts' programs was with what happens with the data it produced.

In this study, the data training that the teachers and administrators received involved walking them through a process of analysis, but the process was never formalized within the district's program. The expectation that teachers utilize the data analysis process was not mandated by the district or by the principals, leading to the abandonment of the process in subsequent semesters. This is not to say that some of the teachers were not analyzing benchmark assessment data, but the number of teachers completing an analysis and acting on the data was small. In addition, of the teachers who did analyze their data, each teacher approached it in a different way, leading to inconsistency across the schools and district. A protocol for data analysis, as suggested by Blanc et al. (2010), would provide the necessary consistency and expectation that the district needs. The professional development that the teachers received on data analysis would have provided a protocol for them, but the process was not formalized within the district or provided to teachers in a written format.

Instructional practice – Tier 2. Through interviews with teachers and administrators, as well as the researcher's observations, the district's benchmark program is used both formatively and summatively, which is congruent with the idea of multiple purposes discussed by other researchers (Bulkley, Christman, et al., 2010; Olson, 2005). Formative assessment implies that after analyzing the data, teachers implement new strategies and interventions to meet the needs identified through the data analysis. These

strategies and interventions must be different from the initial teaching. During the interview process, few teachers spoke of how their teaching changed as a result of the data from the benchmark assessments. Most spoke of the need to 're-teach' and 'remediate,' but only two gave specific examples of how their teaching had changed. Of those two teachers, one spoke specifically of how she made changes for the next semester, but not how she adjusted her instruction for her current students. According to Popham (2008), true formative assessment must mean that a teacher changes her instructional practices for the students from which the data derived. Armed with that information, teachers can make informed decisions regarding the direction of their instruction for individuals and groups of students with the same needs.

However, the teachers understood the need for differentiation for their students. Although teachers spoke of the need to enrich or accelerate students who were performing well, most of their concern focused on the need for differentiation for students at the lower end of the achievement spectrum. The lack of specificity with instructional practice could be related to a frustration with how to implement a differentiated classroom or to a lack of alternative instructional practices. Additional professional development on the differentiation and instructional strategies could bolster the district's effectiveness in this tier.

The three stages of teacher development created by Brookhart et al. (2008) might be helpful to determine where a faculty is operating in regard to the use of the formative assessment process. In this study, one or two teachers were at the third or "intentional" stage, where they were engaging their students in the formative assessment process. A

few teachers were at the second stage, "skill building," but most would fall into stage one or "consciousness raising." In this first stage, teachers might be participating in some parts of the process, but they had not yet intentionally begun engaging their students in the process.

Program effectiveness – Tier 3. The third or top tier of the model (Figure 4) addresses Program Effectiveness, which incorporates all the themes from the first two tiers. When implementing a comprehensive benchmark assessment program, the quality of the Professional Development provided, the application of Assessment Literacy, Data Literacy, and Instructional Practice knowledge contribute to the effectiveness of the program. A highly effective benchmark assessment program would demonstrate strength in each of these areas. Less effective programs would likely exhibit weakness in one or more areas or perhaps would be lacking one of more of the identified areas.

The district's benchmark assessment program incorporates each of the identified areas in varying degrees. The initial professional development incorporated training on the formative and benchmark tool (FABA) and on analyzing benchmark assessment and state testing data. The professional development did not include assessment literacy or intervention strategies, nor was an ongoing plan created to provide coaching support for teachers or instruct teachers new to the school. In terms of instructional bang for the instructional buck, the district's return on its investment was adequate, but not as aggressive as the district needs it to be in order to meet its Adequate Yearly Progress (AYP) goals. Burch (2010) identified fidelity issues with implementation as a reason some districts and schools did not realize the potential of district initiatives. Fidelity is an

issue confronting the district in the current study and contributes to the district having only a moderately effective benchmark assessment program.

The findings suggested that the administrators saw the benchmark assessments as having a formative purpose, and the assessments were only one piece in the overall formative assessment process. The formative assessment process includes students participating in activities such as goal-setting, tracking of their data, and decision-making regarding specific strategies. Strengthening this component of the formative assessment process would provide more ownership and relevance for students. In turn, student involvement in the process would require more involvement on the part of teachers. If teachers were coaching students through the process of making decisions about their learning strategies, based on the data from the benchmark assessments, then teachers would be more likely to make changes in their teaching strategies, based on the data.

Several researchers (Crane, 2010; Marshall, 2006; Perie et al., 2009) advocate for district's to create a basic plan or Theory of Action prior to implementing a benchmark assessment program. The findings indicate that the district from this study may have enjoyed a greater degree of effectiveness if a Theory of Action had been articulated and formalized at the outset of the benchmark assessment program. Such a plan would have explicitly stated the type of professional development the district would provide to each of its teachers, how often the professional development would occur, and when coaching and follow-up support would be available. In addition, a Theory of Action would have provided a roadmap for teachers, postulating the district's vision regarding the purpose of the benchmark assessment program, where it fits with the district's formative assessment

process or Response to Instruction (RTI) process, when assessments would be administered, timeframes and protocols for data analysis, expectations for adjusting instruction, and the level of student involvement in the process. With a detailed, written Theory of Action document, implementation across the district would be more consistent, and fidelity to the benchmark assessment program would be greater.

Prediction. The quantitative research question involved how well the benchmark assessment system predicts later performance on the state's End-of-Grade (EOG) and End-of-Course (EOC) assessments. The district is required to use the EOG and EOC assessments for federal accountability, and the ability to predict which students may not be on track for proficiency on these assessments could be an important tool for the district. The FABA tool does not provide cut scores for the district. Thus, teachers have no way of determining from the students' scores if students are at the basic, proficient, or advanced achievement level in the content area. One goal for the district is to move forward with the data to create cut scores that students, teachers, and principals can utilize.

The benchmark assessments correlated significantly with the subsequent state assessments in all cases except one (i.e., benchmark 2 for grade 5 math). With the exceptions of grade 5 math and grade 8 math, each of the assessment areas demonstrated a moderate to strong (.5 or greater) correlation between at least one of the benchmark assessments and the subsequent state assessment. Thus, the multiple regression model worked well for most of the areas, accounting for 50% or more of the variance. In most cases, each predictor variable contributed to the model, although in grades 3, 5, and 8

math, the pattern did not hold. In addition, in grade 7 reading and in English I, the third predictor variable (the final benchmark assessment) did not contribute significantly to the model.

Generally, the benchmark assessment system that the district utilized predicted well for the state assessments. The final benchmark assessment that the district gives its students most resembles the subsequent state assessment in that it contains test items from all standards. The final benchmark assessment, in most cases, was a strong predictor of a student's End-of-Grade (EOG) or End-of-Course (EOC) assessment score. The district could analyze the few areas where the model did not fit well to determine the reasons for the poor fit. The problem may be in the assessment itself (e.g., weighting of standards on the assessment) or the conditions for administration of the assessment. Other factors, such as how important the students (or teachers) perceive the assessment or the alignment between the taught curriculum and the assessment, could affect how well the benchmark assessment predicts later performance. If the benchmark assessments did not align to the district's pacing guide, then they may be unreliable measures and would account for the inconsistent statistical test results. Additionally, the few unanticipated results could be a problem with the benchmark tests themselves. In the cases where benchmark assessment 1 was the best overall predictor, the assessment could be the best predictor because the information covered early in the year is weighted more heavily on the subsequent End-of-Grade assessment. If benchmark assessments given earlier in the school year are the best predictors, then that information could be useful to educators for identifying early in the school year the students who might need additional support.

The results of the prediction model extend the qualitative findings of the study. By and large, the district's benchmark assessment program is moderately effective from a qualitative lens. The addition of the quantitative prediction model offers refinement of the qualitative findings. The qualitative portion of the study indicated that the benchmark assessment program provided benefits for the district, although some areas of the model (Figure 4) require additional attention before the district can realize the full potential of its benchmark assessment program. Similarly, except for the areas noted earlier, the benchmark assessments that the district administers provide strong predictors for student achievement on their state assessments.

Significance

This study sought to explore one district's benchmark assessment program through a mixed methods approach. It is probable that the findings from the study are accurate because the qualitative methodology included participants from all the schools in the district and those participants represented all subject areas included in the benchmark assessment system. In addition, participants from all grade levels, with the exception of grade 5 and grade 7 were represented in the study. Administrators from each of the three schools were included as well.

The quantitative portion of the study was also comprehensive. The data files used for the quantitative analysis included all grades and subject areas in the district's benchmark assessment program. The data was cleaned so it contained only complete records.

If these findings are true, then school districts who have not implemented a benchmark assessment program could use the model (Figure 4) as a framework from which to plan their benchmark initiatives. Having an articulated plan or Theory of Action (Crane, 2010; Marshall, 2006; Perie et al., 2009) that included each of the themes identified in the model would assist a district in beginning the initiative with a clear vision that all stakeholders could understand. The plan would allow districts to plan the appropriate professional development prior to the actual administration of the benchmark assessments, providing their teachers and administrators with a foundation in assessment literacy, data literacy, and instructional practice.

In addition, districts that have a benchmark assessment system currently in place could use the model as a means to critique the components of their current program. The model (Figure 4) could assist districts in identifying the components in their programs, allowing them to determine whether any of the major components are missing or lacking in development. If so, the district could then proceed to add or strengthen the component. While the model in its present form cannot provide a comprehensive evaluation of a benchmark assessment program, it can provide districts with an initial review of their programs.

Regardless of whether a district is embarking on a new benchmark assessment program or has one currently in place, this study indicates how important it is for teachers to have a firm grasp of formative assessment and to have implemented a formative assessment process in their classrooms. In addition, teachers need continued support

through professional development and coaching for incorporating new instructional strategies into their teaching repertoire.

Finally, this study is significant because it involves American Indian students in a tribally controlled school district, an often underrepresented population in educational research. The research could be helpful to teachers and administrators who work in tribally or Bureau of Indian Education (BIE) operated school districts or to educators who work in districts with large populations of American Indian students. Developing or strengthening a benchmark assessment program could possibly allow schools with American Indian populations to improve student achievement among that population.

Limitations. The accuracy of the findings of this study is limited by several factors. The school district involved in the study is an extremely small district, serving approximately 1,100 students. The student population is predominantly American Indian, an often overlooked population in the literature, but the findings may not generalize to other populations. Another limitation related to the small size of the district involves the small sample size after cleaning the data files of student test records. Because of the small sample sizes, the adjusted R^2 statistic was used.

Phase I of the study was qualitative and involved interviews with 10 teachers and four administrators from the district. All participants freely agreed to the interviews and were candid in their answers to the interview questions. However, the study did not include classroom observations of the teachers utilizing formative assessment, administering the benchmark assessments, or analyzing the data. Such observations

could have validated the teacher and administrator responses, thus strengthening the accuracy of the findings.

Another limitation of the study involves exceptional populations. Although one of the teachers interviewed was a teacher in the special education program, students with disabilities was not a focus of the study. In addition, records from students assessed with an alternate assessment were not included in the statistical analysis. Records from students with disabilities who take the same assessments as their non-disabled peers were included, but these records were not analyzed separately. The focus of the study was not on students with disabilities, nor would an adequate sample size have been available if this was the focus. The same situation holds true for the other end of the spectrum. Records of students who are academically and intellectually gifted were not analyzed separately in the study. The findings of the study might not hold true for either of these populations.

Student responses to the assessments may also affect the accuracy of the study's findings. While students may randomly mark responses on any test, whether high-stakes (state) or low-stakes (benchmarks), the probability is greater with district benchmark assessments, especially with older students when they know they will not receive a grade for their performance. The findings of this study indicate that some students did not always take the benchmark assessments seriously and would sometimes rush through completing them. This situation might account for the areas where the benchmark assessment did not predict well.

A final identified limitation that could lead to erroneous findings is with the technical characteristics of the benchmark assessments used. While the FABA test items have been vetted through a rigorous item development process, the actual assessments have not undergone a similar process. Li et al. (2010) believe that test item quality is paramount. The district creates its own assessments by choosing test items from the FABA item bank. According to Li et al., the validity of a benchmark assessment is related to its purpose. If the purpose is low-stakes, then a lower threshold for the technical characteristics exists. Conversely, if the purpose is high-stakes, then a higher threshold must be met. In this study, the benchmark assessments were low-stakes assessments, meaning that the technical characteristics of the assessments were not as important to the district as other characteristics (e.g., instructional). Because the purpose of the benchmark assessment program in this study was low-stakes, districts with high-stakes benchmark assessments should be cautious in generalizing the information to their situations.

Recommendations for Future Research

If little research exists on benchmark assessments, as suggested by Bulkley, Nabors Oláh, et al. (2010), then this study furthers the literature in the field. Although this study involves a population often overlooked in the literature, additional studies involving benchmark assessment programs with American Indian students and other under-represented populations is needed. In addition, according to the research (Black & Wiliam, 1998), the formative assessment process is especially effective with students

who are behind academically, and so future research with students with disabilities is a promising avenue.

A benchmark assessment program could be an element in a district's Response to Instruction (RTI) initiative. RTI includes universal screening (benchmarks) as well as progress monitoring (formative assessment). Several studies (Atkins & Cummings, 2011; Hintze & Silberglitt, 2005; Nese et al., 2011; Pearce & Gayle, 2009; Wood, 2006) indicate that the universal screenings used with RTI often predict how well students perform on later assessments. A study involving how a district uses its benchmark assessments as part of its RTI program could provide valuable research for improving instruction for all students, as well as strengthening a school or district's benchmark assessment or RTI programs.

Another area for further research is developing the Dimensions of Benchmark Assessment Program Effectiveness model (Figure 4) to become a more thorough evaluation instrument for districts seeking a method of evaluating their benchmark assessment programs. One possibility to explore is providing rubrics for each of the themes contained in the model.

Summary

Based on the previous discussion, the mixed methods study successfully answered the four research questions. Results from the study indicated that the district had implemented a benchmark assessment program that was meeting the district's purposes as articulated by the participants. The benchmark assessments consisted of high quality test items with a web-based delivery for primarily American Indian students in a small,

rural school district. The benchmark assessment program consisted of professional learning opportunities in data analysis as well as the assessment system. Teachers and administrators were provided with data that was used formatively by teachers. Teachers used the data to adjust their instruction, though some struggled with differentiating their instruction based on student needs identified in the data. Though moderately successful, the benchmark assessment program has the potential to demonstrate greater benefits for the district. The qualitative portion of the study identified five themes: Professional Development, Assessment Literacy, Data Literacy, Instructional Practice, and Program Effectiveness. The study found district strengths in the Professional Development, Data Literacy, and overall Program Effectiveness themes, and weaknesses were identified in Assessment Literacy and Instructional Practice. The five themes fashioned the Dimensions of Benchmark Assessment Program Effectiveness model (Figure 4). The model lends itself to a district working to implement a benchmark assessment program or to schools or districts wanting to informally evaluate an existing program.

The quantitative portion of the study provided an answer to the fourth research question regarding how well the program could predict End-of-Grade (EOG) and End-of-Course (EOC) scores. The study concluded that the district's benchmark assessment program correlated strongly with EOG and EOC scores in all but two areas. The benchmark assessment scores predicted the subsequent EOG and EOC scores well in most of the grade levels and subject areas. Multiple regression statistics were used to determine how well the benchmark assessment scores predicted the EOG or EOC scores, and simple linear regression statistics were for individual benchmark assessments when

multicollinearity was suspected. Equipped with the knowledge that the benchmark assessments are strong predictors, teachers and administrators can utilize the knowledge to better personalize student learning experiences.

References

Atkins, T. L., & Cummings, K. D. (2011). Utility of oral reading and retell fluency in predicting proficiency on the Montana Comprehensive Assessment System. *Rural Special Education Quarterly, 30*(2), 3-12. Retrieved from EBSCOhost on September 20, 2011.

Barger, J. (2003). *Comparing the DIBELS oral reading fluency indicator and the North Carolina end of grade reading assessment (Technical Report).* Asheville, NC: North Carolina Teacher Academy. Retrieved on October 30, 2011, from http://dibels.uoregon.edu/techreports/NC_Tech_Report.pdf

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan, 86*(1), 9-21. Retrieved from EBSCOhost on October 30, 2011.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148. Retrieved from EBSCOhost on October 30, 2011.

Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. E. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education, 85*, 205-225. Retrieved from EBSCOhost on October 1, 2011.

Brookhart, S., Moss, C., & Long, B. (2008). Formative assessment that empowers. *Educational Leadership, 66*(3), 52-57. Retrieved from EBSCOhost on October 30, 2011.

Brown, R. S., & Coughlin, E. (2007). *The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region (Issues & Answers Report, REL 2007 - No. 017).* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs

Bulkley, K. E., Christman, J. B., Goertz, M. E. & Lawrence, N. R. (2010). Building with benchmarks: The role of the district in Philadelphia's benchmark assessment system. *Peabody Journal of Education, 85,* 186-204. Retrieved from EBSCOhost on September 15, 2011.

Bulkley, K. E., Nabors Oláh, L., & Blanc, S. (2010). Introduction to the special issue on benchmarks for success: Interim assessments as a strategy for educational improvement. *Peabody Journal of Education, 85*, 115-124. Retrieved from EBSCOhost on September 15, 2011.

Burch, P. (2010). The bigger picture: Institutional perspectives on interim assessment technologies. *Peabody Journal of Education, 85*, 147-162. Retrieved from EBSCOhost on October 1, 2011.

Bureau of Indian Education. (2011). *Bureau-wide Annual Report Card, 2009-2010.* Retrieved from http://www.bie.edu/idc/groups/xbie/documents/text/idc012921.pdf

Center for Urban Affairs and Community Services. (2008). *FABA Assessment System: Preliminary Analysis of Effectiveness*. North Carolina State University. Retrieved from http://classcape.mck.ncsu.edu/FABA3/docs/userDocs/Research/docs/FABA_Analysis_Report.pdf

Center for Urban Affairs and Community Services. (2011). *FABA overview*. North Carolina State University. Retrieved from http://classcape.mck.ncsu.edu/FABA3/presentations.jsp

Crane, E. (2010). *Building an interim assessment system: A workbook for school districts*. Washington, DC: Council of Chief State School Officers. Retrieved October 30, 2011, from http://www.ccsso.org/Resources/Publications/Building_an_Interim_Assessment_System_A_Workbook_for_School_Districts.html

Creswell, J. (2005). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (2nd ed.). Upper Saddle River, NJ: Pearson Education.

Creswell, J. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Dedoose, Version 4.5, web application for managing, analyzing, and presenting qualitative and mixed method data. (2013). Los Angeles, CA: SocioCultural Research Consultants, LLC.

Graney, S. B., Missal, K. N., Martinez, R. S., & Bergstrom, M. (2009). A preliminary investigation of within-year growth patterns in reading and mathematics curriculum-based measures. *Journal of School Psychology, 47*, 121-141. Retrieved from EBSCOhost on September 15, 2011.

Gravetter, F. J., & Wallnau, L. B. (2009). *Statistics for the behavioral sciences* (8th ed.). Belmont, CA: Wadsworth.

Grigg, W., Moran, R., & Kuang, M. (2010). *National Indian Education Study - Part I: Performance of American Indian and Alaska Native Students at Grades 4 and 8 on NAEP 2009 Reading and Mathematics Assessments (NCES 2010–462).* Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Halverson, R. (2010). School formative feedback systems. *Peabody Journal of Education, 85*, 130-146. Retrieved from EBSCOhost on September 15, 2011.

Halverson, R., Prichett, R. B., & Watson, J. G. (2007). *Formative feedback systems and the new instructional leadership (WCER Working Paper No. 2007-3).* Madison, WI: University of Wisconsin-Madison, Wisconsin Center for Educational Research. Retrieved from EBSCOhost on October 1, 2011.

Hassel, E. (1999). *Professional development: Learning from the best. A toolkit for schools and districts based on the National Awards Program for Model Professional Development*. Oak Brook, IL: North Central Regional Educational Laboratory. Retrieved April 28, 2013, from http://www.learningpt.org/pdfs/pd/lftb.pdf

Hefflin, P. (2009). Do benchmark assessments increase student achievement on state standardized tests? (Doctoral dissertation, Duquesne University, 2009). *ProQuest Dissertations and Theses*, ID 3371521.

Heppner, P. P., & Heppner, M. J. (2004). *Writing and publishing your thesis, dissertation & research: A guide for students in the helping professions*. Belmont, CA: Brooks/Cole CENGAGE Learning.

Herman, J. L., & Baker, E. L. (2005, November). Making benchmark testing work. *Educational Leadership*, *63*(3), 48-54. Retrieved from EBSCOhost on September 15, 2011.

Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of –CBM and high-stakes testing. *School Psychology Review*, *34*(3), 372-386. Retrieved from EBSCOhost on September 20, 2011.

Hunt, W. L. (2008). Effect of grading a benchmark assessment on student performance. (Doctoral dissertation, Widener University, 2008). *ProQuest Dissertations and Theses*, ID 304800538.

Individuals with Disabilities Education Improvement (IDEA) Act of 2004, Pub. L. 108-466.

Learning Forward. (2012). *Standards for professional learning*. Retrieved on April 28, 2013, from http://learningforward.org/standards/standards-list#.UX1PDlcQLNo

Li, Y., Marion, S., Perie, M., & Gong, B. (2010). An approach for evaluating the technical quality of interim assessments. *Peabody Journal of Education, 85*, 163-185. Retrieved from EBSCOhost on October 1, 2011.

Marshall, K. (2006). *Interim assessments: Keys to successful implementation*. Interim Assessment Project, New Leaders for New Schools. Retrieved from EBSCOhost on October 1, 2011.

Marshal, K. (2008). Interim assessments: A user's guide. *Phi Delta Kappan, 90*(1), 65-68. Retrieved from EBSCOhost on October 1, 2011.

Marzano, R. J., Brandt, R. S., Hughes, C. S., Jones, B. F., Presseisen, B. Z., Stuart, C., & Suhor, C. (1988). *Dimensions of thinking*. Alexandria, VA: Association for Supervision and Curriculum Development.

McManus, S. (2008). *Attributes of effective formative assessment*. FAST SCASS, Council of Chief State School Officers. Retrieved October 30, 2011, from http://www.ccsso.org/Documents/2008/Attributes_of_Effective_2008.pdf

McTighe, J., & O'Connor, K. (2005, November). Seven practices for effective learning. *Educational Leadership, 63*(3), 10-17. Retrieved from EBSCOhost on September 15, 2011.

Nabors Oláh, L., Lawrence, N. R., & Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody Journal of Education, 85*, 226-245. Retrieved from EBSCOhost on September 15, 2011.

Nese, J. F. T., Park, B. J., Alonzo, J., & Tindal, G. (2011). Applied curriculum-based measurement as a predictor of high-stakes assessments: Implications for researchers and teachers. *The Elementary School Journal, 111*(4), 608-624. Retrieved from EBSCOhost on September 15, 2011.

Newsom, J. T. (1999-2007). *Lecture 20: More on multiple regression*. Retrieved on February 22, 2013, from www.upa.pdx.edu/IOA/newsom/pa551/lectur20.htm

No Child Left Behind Act of 2002, Pub. L. 107-110.

North Carolina Department of Public Instruction. (n.d.). *A vision for 21st century assessment. Accountability Services Division.* Retrieved September 20, 2011 from https://www.ncpublischools.org/accountability/educators/vision/

North Carolina Department of Public Instruction. (2007, October). *North Carolina end-of-grade test of reading comprehension – Grade 3. Test information sheet.* Retrieved January 1, 2012, from http://www.ncpublicschools.org/docs/accountability/testing/eog/20071001eogreadingtestinformationsheetgrade3.pdf

North Carolina Department of Public Instruction. (2008a, March). *The North Carolina English language arts test: Edition 3, English I. Technical Report.* Retrieved January 2, 2012 from http://www.ncpublicschools.org/docs/accountability/testing/reports/english1techmanualdraft.pdf

North Carolina Department of Public Instruction. (2008b, June). *The North Carolina mathematics tests, Edition 3. Technical Report.* Retrieved January 2, 2012, from http://www.ncpublicschools.org/docs/accountability/reports/mathtechmanualdrafted2.pdf

North Carolina Department of Public Instruction. (2009, April). *Grade 3 reading comprehension pretest, end-of-grade reading comprehension tests*. Retrieved January 1, 2012, from http://www.ncpublicschools.org/docs/accountability/testing/reports/eogreadingtechman3.pdf

North Carolina Department of Public Instruction. (2011). *Responsiveness to instruction: The problem-solving model and analyzing the core*. PowerPoint presentation. Retrieved November 14, 2011 from http://www.ncpublicschools.org/docs/curriculum/responsiveness/rtimaterials/analyze-core.pdf

Olson, L. (2005). Benchmark assessments offer regular achievement. *Education Week, 25*(13). 13-14. Retrieved from EBSCOhost on September 15, 2011.

Pearce, L. R., & Gayle, R. (2009). Oral reading fluency as a predictor of reading comprehension with American Indian and White elementary students. *School Psychology Review, 38*(3), 419-427. Retrieved from EBSCOhost on October 30, 2011.

Pearson. (2012, February 27). *Pearson 2011 preliminary results (unaudited).* Press Release. Retrieved March 19, 2012, from http://www.pearson.com/media/files/press-releases/2012/2011_Results_Press_Release_FULL.pdf

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practices, 28*(3), 5-13. Retrieved from EBSCOhost on October 1, 2011.

Petscher, Y., & Young-Suk, K. (2011). The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of School Psychology, 49*, 107-129. Retrieved from EBSCOhost on September 20, 2011.

Popham, W. J. (2008). *Transformative assessment*. Alexandria, VA: ASCD.

Protheroe, N. (2001). Improving teaching and learning with data-based decision: Asking the right questions and acting on the answers. *ERS Spectrum*. Retrieved on October 30, 2011, from https://www.ers.org/spectrum/sum01a.htm

Richards, L., & Morse, J. M. (2007). *Readme first for a user's guide to qualitative methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rudner, L. M. (1994). Questions to ask when evaluating tests. Practical assessment, *Research & Evaluation, 4*(2). Retrieved October 30, 2011, from http://PAREonline.net/getvn.asp?v=4&n=2

Shepard, L. A. (2010). What the marketplace has brought us: Item-by-item teaching with little instructional insight. *Peabody Journal of Education, 85,* 246-257. Retrieved from EBSCOhost on September 15, 2011.

Soper, D. (2006-12). *Statistics calculator. [Computer software]*. Retrieved February 12, 2012, from http://danielsoper.com/statcalc3/calc.aspx?id=1

Stiggins, R. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77*(3), 238. Retrieved from EBSCOhost on October 30, 2011.

Stiggins, R. (2005). From formative assessment to assessment for learning: A path to success in standards-based schools. *Phi Delta Kappan, 87*(4), 324-328. Retrieved from EBSCOhost on October 30, 2011.

Stiggins, R., & Duke, D. (2008). Effective instructional leadership requires assessment leadership. *Phi Delta Kappan, 90*(4), 285-291. Retrieved from EBSCOhost on October 30, 2011.

USED. (2001). *No Child Left Behind (NCLB) Act,* Pub. L. 107-110.

Wayman, J. C. (2005). Involving teachers in data-driven decision making: Using computer data systems to support teacher inquiry and reflection. *Journal of Education for Students Placed At Risk, 10*(3), 295-308. Retrieved from EBSCOhost on October 1, 2011.

Wood, D. E. (2006). Modeling the relationship between oral reading fluency and performance on a statewide reading test. *Educational Assessment, 11*(2), 85-104. Retrieved from EBSCOhost on September 20, 2011.

Wright, J. (2010). *"How RTI Works" series.* Retrieved from www.interventioncentral.org

Appendix A

Interview Protocol

Interview Protocol

Dissertation Study: The Nature and Predicative Validity of a Benchmark Assessment Program in an American Indian School District
Date of Interview: _____ Time **of Interview:** _____

Location: _____
Interviewer: Beverly Payne, Investigator

Interviewee Code: _____

Position: ___ Teacher @ Elem, MS, HS ___ Administration @ Elem, MS, HS, CO

Years in education _____ Years in 2011-2012 position _____

Introduction:

1. Thank you for taking the time to visit with me today.
2. I am conducting dissertation research on the school system's benchmark assessment program. I will be interviewing several staff members from the district for this study.
3. First, I want to assure you that this interview is strictly confidential. Information provided by school and district staff is reported or released in aggregated form only. Districts, schools, and individuals are not identified. Pseudonyms will be used to maintain confidentiality when necessary.
4. I have an Informed Consent form outlining your rights as a research participant. You are free to decide not to participate in this study or to withdraw from the study at any time without adversely affecting your relationship with me, the University of Nebraska-Lincoln, or your school district. Contact persons for the project and the Institutional Review Board are provided on the Informed Consent Form in case you have questions or concerns. I have a copy for you to sign and one for you to keep for your use.
5. It is important that educators participating in this research be willing participants. You are free to decide not to participate or to withdraw from the interview at any time without harming your relationship with your district, this project, or the University of Nebraska-Lincoln. Should you decide not to participate you may return to your normal activities. Are you willing to participate in this interview?
6. I am going to record this interview so that the interview can be transcribed (a typed copy of the interview will be made) and we have an accurate rendering of your responses.
7. It is important that I maintain the integrity of your words and intentions; therefore, I may ask you to review the transcription if I have any difficulties with the interpretation.

8. I am interested in your perceptions and understanding of the development and implementation of your school system's benchmark assessment program and its relationship to the End-of-Grade and End-of-Course scores.
9. Please feel free to discuss your views openly. From time to time, I may have additional questions to further understand a concept that you have shared.
10. Let's begin. Please state your name, school, district, and give verbal permission to record this interview by repeating this statement, "I (your name) at (school/district name) willingly give my permission to record this interview."

Part I.

1. What do you believe is the purpose of the benchmark assessment program?

2. What are the strengths of the FABA benchmark assessments?

3. What are the weaknesses of the FABA benchmark assessments?

4. Have you participated in building any of the district benchmark assessments?

    Probe: How helpful (in what ways) was that experience?

5. How has the benchmark program influenced how you think about assessment in general?

Part II.

6. Describe your FABA training.

    Probe: Did any of the FABA training focus on what to do with the data produced?

7. Describe what kind of training you have had regarding how to use data.

    Probe: Would these techniques or strategies work with benchmark assessment data? Why or why not?

    Probe: How are you provided with ongoing support for data use?

8. What type of professional development would be useful to you for better utilizing the benchmark assessment program?

Part III.

9. How confident are you in your ability to analyze benchmark assessment data?

10. What is your normal procedure for data analysis once a benchmark assessment is completed?

Probe: Do you use a protocol? Please describe.

11. Do you collaborate with other educators in analyzing the benchmark assessment data or in developing strategies or activities to address needs identified through the analysis?

Probe: If yes, how often and with whom?

Probe: If no, why not?

12. How do students participate in analyzing the benchmark assessment data?

Probe: How helpful do your students find the data?

Probe: Do they set goals or track their data?

13. How do you obtain your data?

14. What other types of data would you like to see from the benchmark assessment program?

Probe: How helpful would data from constructed response questions be to you?

Part IV.

15. What type of instructional decisions have you made based on the data?

Probe: What type of activities or strategies have you implemented based on benchmark assessment data?

Probe: How often do you incorporate new activities based on benchmark assessment data?

16. How useful do you believe the benchmark assessment data to be?

17. Do you give students a grade for their performance on the benchmark assessment?

Probe: Why have you chosen to give (or not to give) a grade?

Part V.

18. How many years have you taught?

19. How many years have you taught at this school?

20. How many years have you taught this content (e.g., reading, math, or science)?

Thank you again for participating in this interview. Please remember that your responses will remain anonymous.

Appendix B

Letter from School Board Chair

29 March 2012

Beverly Payne

Dear Mrs. Payne,

The approved your dissertation study in June 2011, and this letter serves as official permission for you to conduct your dissertation study . The board understands that you will:

- Collect benchmark assessment scores.
- Collect End-of-Grade and End-of Course scores.
- Collect program indicators (e.g. regular program, gifted program, special education program).
- Interview pertinent district staff (e.g. administrators, teachers) regarding the assessment program.
- Maintain confidentiality of student information.
- Not identify the school system by name.

We believe this study will contribute to the research base on American Indian education, and we look forward to hearing your findings.

Sincerely,

Lori Blankenship