An Embedded Case Study of the Implementation of A School Division’S Benchmark Assessment System

Chelsea Ireland Kulp
College of William and Mary - School of Education, clirel@gmail.com

Follow this and additional works at: https://scholarworks.wm.edu/etd

Part of the Education Commons

Recommended Citation
http://doi.org/10.21220/W4V36D

This Dissertation is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Dissertations, Theses, and Masters Projects by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.
An Embedded Case Study of the Implementation of a School Division’s Benchmark Assessment System

A Dissertation

Presented to

The Faculty of the School of Education

The College of William and Mary in Virginia

In Partial Fulfillment

Of the Requirements for the Degree

Doctor of Education

by

Chelsea Ireland Kulp

March 2017
AN EMBEDDED CASE STUDY OF
THE IMPLEMENTATION OF A SCHOOL DIVISION’S
BENCHMARK ASSESSMENT SYSTEM

by

Chelsea Ireland Kulp

Approved March 2017 by

Christopher R. Gareis, Ed.D.
Chairperson of Doctoral Committee

Leslie W. Grant, Ph.D.

Charles A. Wagner, Ed.D.
Dedication

I would like to dedicate this dissertation to my family: my husband, my parents, and my sister. Thank you for being my cheerleaders, and therapists, for the last several years! I could not have completed this program without your support and encouragement. I love you!
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgements</td>
<td>vi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td>Abstract</td>
<td>viii</td>
</tr>
<tr>
<td>Chapter One: Introduction</td>
<td>2</td>
</tr>
<tr>
<td>Problem Statement</td>
<td>3</td>
</tr>
<tr>
<td>Conceptual Framework</td>
<td>4</td>
</tr>
<tr>
<td>Research Questions</td>
<td>10</td>
</tr>
<tr>
<td>Significance of the Study</td>
<td>11</td>
</tr>
<tr>
<td>Definition of Terms</td>
<td>13</td>
</tr>
<tr>
<td>Chapter Two: Review of Literature</td>
<td>16</td>
</tr>
<tr>
<td>Assessment in the Classroom</td>
<td>16</td>
</tr>
<tr>
<td>Characteristics of effective assessment</td>
<td>18</td>
</tr>
<tr>
<td>Validity and reliability</td>
<td>19</td>
</tr>
<tr>
<td>Formative assessment</td>
<td>21</td>
</tr>
<tr>
<td>School Improvement</td>
<td>22</td>
</tr>
<tr>
<td>Accountability Era</td>
<td>24</td>
</tr>
<tr>
<td>Intended consequences</td>
<td>26</td>
</tr>
<tr>
<td>Unintended consequences</td>
<td>27</td>
</tr>
<tr>
<td>Every Student Succeeds Act</td>
<td>29</td>
</tr>
<tr>
<td>Benchmark Assessments</td>
<td>30</td>
</tr>
<tr>
<td>Teacher Competencies</td>
<td>34</td>
</tr>
<tr>
<td>Chapter Three: Methodology</td>
<td>37</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Intended Consequences</td>
<td>80</td>
</tr>
<tr>
<td>Unintended Consequences</td>
<td>82</td>
</tr>
<tr>
<td>Instructional Time</td>
<td>87</td>
</tr>
<tr>
<td>Research Question 4</td>
<td>88</td>
</tr>
<tr>
<td>Research Question 5</td>
<td>93</td>
</tr>
<tr>
<td>Research Question 6</td>
<td>95</td>
</tr>
<tr>
<td>School A</td>
<td>95</td>
</tr>
<tr>
<td>School B</td>
<td>97</td>
</tr>
<tr>
<td>School C</td>
<td>98</td>
</tr>
<tr>
<td>Comparison of Three Elementary Schools</td>
<td>98</td>
</tr>
<tr>
<td>Chapter Five: Discussion, Implications, and Conclusions</td>
<td>100</td>
</tr>
<tr>
<td>Validity</td>
<td>100</td>
</tr>
<tr>
<td>Predictive Validity</td>
<td>101</td>
</tr>
<tr>
<td>Consequential Validity</td>
<td>102</td>
</tr>
<tr>
<td>Context, Systems, and Coherence</td>
<td>102</td>
</tr>
<tr>
<td>Conceptual Frameworks</td>
<td>105</td>
</tr>
<tr>
<td>Recommendations</td>
<td>107</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

I would like to thank my dissertation committee members, without whom I could not have completed this project. Thank you to Dr. Chris Gareis for being a role model and for being my dissertation chair. Thank you for encouraging me to produce the best possible work and for pushing me forward. Thank you to Dr. Leslie Grant and Dr. Chuck Wagner for being invaluable resources throughout my dissertation process. I can’t thank you enough for your encouragement and for your time. I admire you both so much and was overjoyed that you would share your time with me.

A huge thank you to Dr. Shelly Cihak! I would have been completely lost without your guidance and help as I moved through the dissertation process. Thank you for being a sound board regardless of the time of day or what we had going on at school. I cannot thank you enough! Thank you to Dr. Antonia Fox for your help and for your support. I really appreciate your time, particularly with regard to reviewing my interview questions and protocols. You are an invaluable mentor to me, and I appreciate you so much. Thank you, also, to Dr. Amy Colley! I cannot express how much I appreciate your willingness to spend 45 minutes on the phone with me at the onset of my research, discussing my vision and bouncing off ideas. Thank you for everything you have done for me!

Thanks to my Poquoson High School colleagues! I hope that I have not been too much of a grouch this year as I worked to complete my dissertation. Special thanks to Brittany Baker and to Molly Holcombe for helping me with my dissertation and for listening to me whine! I adore you.
LIST OF FIGURES

Figure 1 The Alignment Triangle .......................................................... 4
Figure 2 Implementation Framework .................................................. 5
Figure 3 Embedded Case Study Framework ........................................ 38
Figure 4 Embedded Case Study: A School Division ............................ 39
Figure 5 Embedded Case Study: Implementation Framework ............... 40
AN EMBEDDED CASE STUDY OF
THE IMPLEMENTATION OF A SCHOOL DIVISION’S
BENCHMARK ASSESSMENT SYSTEM

ABSTRACT
This embedded case study explored the alignments between division benchmark intents and practices with regard to benchmark assessments and elementary school teachers’ intents and practices. Utilizing the NIRN implementation framework and the embedded case study framework, this study sought to determine what variables are necessary to implement and maintain an effective benchmark assessment system. This study addressed the following research questions: (1) How do teachers describe their intent in making use of the benchmark assessment system? (2) How do teachers describe their practices of making formative use of the benchmark assessment system? (3) What do teachers describe as the perceived outcomes of the benchmark assessment system? (4) What are the similarities and differences between division intent and design of the benchmark assessment system and teacher intent and practices? (5) How competent do teachers feel themselves to be to make use of the benchmark assessment system to progress student learning? (6) How are teachers’ responses similar and/or different relative to the state accreditation status of their respective schools?

CHELSEA I. KULP
EDUCATIONAL POLICY, PLANNING AND LEADERSHIP
THE COLLEGE OF WILLIAM AND MARY IN VIRGINIA
AN EMBEDDED CASE STUDY OF THE IMPLEMENTATION OF A SCHOOL
DIVISION’S BENCHMARK ASSESSMENT SYSTEM
CHAPTER ONE

In the era of accountability, schools are under constant pressure to increase student achievement. As the Every Student Succeeds Act (ESSA) replaces No Child Left Behind (NCLB), it is evident that accountability measures are here to stay. Current research underscores the importance of the school principal with regard to measures of accountability. Horng, Kalogrides, and Loeb (2009) state that,

effective principals influence a variety of school outcomes, including student achievement, through their recruitment and motivation of quality teachers, their ability to identify and articulate school vision and goals, their effective allocation of resources, and their development of organizational structures to support instruction and learning. (p. 1)

Assessment practices within a school influence these key factors of school improvement. Principals and teachers are on the frontlines of school improvement, and, therefore, it is imperative that they are literate in areas of assessment.

While many studies focus on the importance of curriculum and instruction, assessment is the missing piece that completes the effective instructional cycle (Gareis & Grant, 2015). High stakes assessments, in particular, have become the criteria by which schools are judged and evaluated. Strauss (2016) notes that ESSA does “absolutely nothing to limit standardized testing” and that “punishing schools [that aren’t performing] doesn’t help kids learn” (p. 1). In Virginia, the Standards of Learning (SOL) assessments have been the basis for school accreditation. While assessment plays a pivotal role in the perceived success of our schools, research demonstrates that many teachers and administrators lack assessment literacy, which undermines their ability to successfully
utilize assessments for school improvement purposes (Abrams, McMillan, & Wetzel, 2015).

**Problem Statement**

Benchmark testing has grown increasingly prevalent in today’s schools as a measure of predicting student performance on end-of-year high-stakes tests; however, many schools are failing to successfully implement these programs due to a variety of variables. Symonds (2004) notes that variables necessary for successful benchmark implementation include frequent and reliable data, a plan to support teachers in data use, the acknowledgment of the importance of race with regard to assessment, and school-wide focus on the assessments themselves. Wang, Walters, and Thum (2012) also found necessary factors that must be in place in order for benchmark assessments to fulfill their role in instruction. These researchers studied 204 urban schools over the course of three years. They surveyed 6,684 teachers and 149,665 students in grades 1 through 10. With regard to school improvement, these researchers found that high growth schools exhibited strong evidence-based decision-making practices where teachers used the district’s benchmark assessment to reflect on instructional practice, used the core curriculum to guide instruction, and received frequent and high quality professional development on reading and math instruction. (Wang, Walters, & Thum, 2012, p. 517)

Based on their findings, the researchers argued, “Districts will benefit from integrating measures of growth and using school data management systems that integrate benchmark assessment capabilities and provide teachers with the training and tools needed to use the information on their daily practice” (p. 539). Many studies have found, however, that
these conditions for successful benchmark implementation are not found in all schools (Abrams & McMillan, 2013). While schools are being evaluated based on their assessment results, school leaders need to work to ensure that conditions exist for successful benchmark systems in order to increase student achievement.

**Conceptual Framework**

This study is based on knowledge of best practices in assessment and a common language regarding benchmark assessments and their role in schools. One component of the conceptual framework is the alignment among curriculum, instruction, and assessment as represented in Figure 1.

Figure 1

![The Alignment Triangle](attachment:image.png)

**The Alignment Triangle**

Bunch (2012) states,

The goal of alignment is to make curriculum, instruction, and assessment work toward the same ends. Generally, we start with curriculum, lay out goals for instruction, instruct to achieve those goals, and assess to determine how successful we’ve been in achieving the goals set forth in the curriculum. (p. 1)
Figure 1 illustrates the reciprocal relationship between these variables (Gareis & Grant, 2015). Assessments should be designed in such a way that the results help teachers ensure that their instruction is aligned with the intended learning goals, as established by the curriculum. Curriculum and instruction are pivotal components of effective teaching; however, teachers who focus solely on these two areas are missing an important piece of the effective teaching puzzle. Curriculum and instruction focus on the “what” and “how” of teaching, whereas assessment completes the model of teaching and learning by focusing on the “degree to which” students have learned (Gareis & Grant, 2015, p. 3). Benchmark assessments, in particular, are intended to serve a formative role in aligning curriculum and instruction within the classroom setting. The implementation of a benchmark system can help provide formative feedback to classroom teachers in order for them to make sound instructional decisions. There are many components that make up a successful implementation framework, as noted in Figure 2.

Figure 2

**Integrated & Compensatory**

**Leadership Drivers**
- Technical
  - Time
  - Funding
- Adaptive
  - Motivation
  - Clinical inertia

**Organizational Drivers**
- Performance Assessment
- Coaching
- Systems Intervention
- Managerial Support
- Facilitative Administration
- Immediate Appointment
- Nurse/Time
- Caseload

**Compensatory Drivers**
- Decision Support
- Data System
- Change Agents

**Implementation Framework, National Implementation Research Network**

The National Implementation Research Network (NIRN) has created an implementation framework that will be utilized throughout this study. This study,
particular, focuses on competency drivers and organization drivers. Competency drivers are the individuals on the frontlines of benchmark assessment. NIRN argues that there are certain supports necessary for competency drivers to be successful components of implementation. Practitioners (e.g., teachers, division staff, and implementation team members) need to be trained and coached when a “new way of work” is introduced (Implementation Drivers, 2016, p. 1). In schools, for example, this could include new curricular objectives, new instructional methods, or new forms of assessment.

Figure 2 illustrates that conditions that exist for successful competency drivers include training, coaching, and selection. Coaching includes support after training, teacher comfort with the topic (such as new objectives or new forms of assessment), and rapport with client (such as teachers’ sense of clarity about the intents of their school division leadership on a given initiative). The four main roles of a coach include supervision, teaching while engaged in practice activities, assessment and feedback, and provision of emotional support (Implementation Drivers, 2016). When an initiative is implemented, the school or school division is essentially seeking to establish a new behavior among teachers. There are many challenges that can arise from attempting to change behavior. Particularly with regard to establishing a benchmark assessment system, it is imperative that the school division establishes strong coaching practices and the appropriate training necessary for teachers to feel competent in what they are being asked to do, such as administer benchmark assessments and analyze their results.

Training includes intervention training, knowledge, and belief in usefulness. Training is also necessary for teachers to feel competent with regard to the initiative (e.g., the administration and use of benchmark assessments). Approaches to training include
providing information about “the history of an initiative theory, philosophy, and rationales for program components and practices conveyed in lecture and discussion formats” (Implementation Drivers, 2016, p. 1). In the case of benchmark assessment implementation, leaders would need to share research findings that support the use of benchmark assessments with teachers. It could also be beneficial to train teachers in the practices of effective benchmark assessment use.

Lastly, teachers need to be selected based on their attitudes, receptivity to training, and whether or not the implementation relates to their assigned roles. Selection includes determining which individuals will serve in the capacity of practitioner, of organization staff, or as a part of an implementation team (Implementation Drivers, 2016). It is imperative to have individuals properly placed according to their skills. For example, “people who are methodical and comfortable making judgments based on specified criteria may make better evaluators” (Implementation Drivers, 2016, p. 1). With regard to benchmark assessments, it may be worthwhile for early adaptors to serve in the capacity as teacher leaders on an implementation team in order to help fellow teachers better implement the initiative. Selection of roles is a critical task at the organizational level of implementation.

Organization drivers play a critical role in the implementation framework. Organization drivers are facilitative administrators (e.g., superintendents, principals, and other non-teaching staff) who are key facilitators of implementation and of school change. Their role within the implementation framework, in the case of benchmark assessments, for instance, is to establish a data system to guide the process of innovation, create a hospitable environment for change, assess immediate outcomes of the change,
and create support for implementers. In this study, the benchmark assessment system is the change being evaluated, and the school- and division-level administrators are the key organization drivers (Implementation Drivers, 2016). Conditions needed for successful organization drivers include systems intervention, facilitative administration, and a decision support data system.

Systems intervention includes managerial support: “Implementation takes place in a shifting ecology of agency, community, state and federal social, economic, cultural, political, and policy environments” (Implementation Drivers, 2016, p. 1). It is imperative that school divisions understand their context before they implement a new initiative. The success and sustainability of a program depend on the “degree to which agency, community, state and federal systems are supportively aligned and enabling with respect to implementation” (Implementation Drivers, 2016, p. 1). If the current district environment is not conducive to a particular initiative, then the implementation may be a failure before it even begins. Organizational leaders need to be aware of their context in order to determine whether or not an initiative is a good fit for the district, or if prior training and competency building needs to be first conducted. For example, prior to implementing a benchmark assessment system, a division may find that their teachers need training in either creating effective assessments or in the formative uses of assessments. Teachers may need training in both areas in order for a benchmark assessment implementation to be successful. NIRN underscores the importance of this alignment between division conditions and the implementation of a new program:

Systems intervention requires attending to multi-level alignment, maintain leadership and focus, creating and staying connected to champions, intervening to
change policies and funding contingencies, and remaining vigilant at local, state, and federal levels for both windows of opportunity and threats to fidelity and sustainability. (Implementation Drivers, 2016, p. 1)

Facilitative administration is also critical to organization drivers and includes immediate appointment, time, and caseload. NIRN notes that facilitative administrative support is “proactive, vigorous and enthusiastic attention by the administration to reduce implementation barriers and create an administratively hospitable environment for practitioners” (Implementation Drivers, 2016, p. 1). A survey conducted by NIRN found that administrative support was a critical component to success. The survey analyzed two different groups: a successful group (with regard to implementation) and an unsuccessful group. The successful group felt that their administrators had worked to eliminate barriers for teachers, such as reducing the amount of paperwork teachers needed to complete and increasing the time available to complete the task. The unsuccessful group, however, felt overburdened by administrators (Implementation Drivers, 2016).

Decision support data systems are defined by NIRN as “sources of information used to help make good decisions internal to an organization” and are used to “assess key aspects of the overall performance of the organization, provide data to support decision making, and assure continuing implementation of the evidence-based intervention and benefits to consumers over time” (Implementation Drivers, 2016, p. 1). One such system that exists within school divisions is a financial data collection and reporting system. In the case of implementing a benchmark assessment system, a school division may need to evaluate whether or not it is financially able to support such an initiative. Another important component of decision support data systems is feedback. The feedback loop is
critical to keep “an evidence-based program ‘on track’ in the midst of a sea of change” (Implementation Drivers, 2016, p. 1).

The final driver of the implementation framework is leadership; leadership drivers include both technical leadership and adaptive leadership. This driver is a critical component of successful implementation, because “to exercise leadership toward the full implementation of effective innovations means moving a complex and entrenched system through meaningful change—and leading through the resistance that can arise in the process” (Implementation Drivers, 2016, p. 1). While leadership drivers are a critical component of effective implementation, this study will primarily focus on both competency drivers and organization drivers due to the assigned roles of the selected participants. These implementation drivers are critical components of systemic initiative implementation, such as the benchmark system implemented in this school division. This study analyzed the organizational and competency drivers of the school division’s benchmark system in order to determine alignment between teacher and division goals, intent, and practices for benchmark assessments.

**Research Questions**

This study is designed to analyze the manner in which teachers implement benchmark assessments in the classroom setting in order to progress study learning. There are six research questions in this study:

1. How do teachers describe their intent in making use of the benchmark assessment system?
2. How do teachers describe their practices of making formative use of the benchmark assessment system?
3. What do teachers describe as the perceived outcomes of the benchmark assessment system?

4. What are the similarities and differences between division intent and design of the benchmark assessment system and teacher intent and practices?

5. How competent do teachers feel themselves to be to make use of the benchmark assessment system to progress student learning? (a) Why do teachers feel more or less competent? (b) What do teachers feel would contribute to their competency?

6. How are teachers’ responses similar and/or different relative to the state accreditation status of their respective schools?

An embedded case study of participants in three elementary schools within a school division was conducted in order to look for themes among teacher use of benchmark assessments.

Significance of the Study

This study is designed to produce clear guidelines for successful benchmark programs in order to help educators increase student achievement. A study conducted by Slavin, Cheung, Holmes, Madden, and Chamberlain (2013) found intriguing results with regard to the success of benchmark assessments in terms of school improvement. These researchers administered a division-level reform model created by the Center for Data-Driven Reform in Education (CDDRE). Fifty-nine districts in seven states were randomly assigned either CDDRE or control conditions. The researchers found that “helping school leaders understand student data is helpful but in itself does not produce educationally important gains in achievement. Schools must actually take action to change teaching and learning” (Slavin et al., 2013, p. 390). The purpose of benchmark
assessments, they argue, is to “find out early where problems exist so that changes can be made before it is too late” (p. 374). While this study analyzed student achievement in light of application of a reform model, the study failed to analyze whether or not teachers had the ability to take action needed based on the results. This study of a school division’s benchmark system is designed to examine whether teachers are able to successfully utilize benchmark results.

Benchmark assessments can be used to inform policy, instructional planning, and decision-making at the classroom, school, and district levels (Herman, Osmundson, & Dietel, 2010). Many obstacles can inhibit schools from creating and implementing successful benchmark systems, including time, expertise, support, and understanding of the potential value of benchmarks themselves (Murnane, Sharkey, & Boudett, 2005). Researchers argue that there have been limited investigations of the impact of benchmark testing and of their formative uses (Abrams et al., 2015). This study of a division’s benchmark system is designed to narrow the gap in literature by examining the variables necessary for successful benchmark implementation. Furthermore, school systems would benefit from clear procedural recommendations, because the effective use of benchmark assessments has been found to increase student achievement (Wang et al., 2012). Researchers have found that

high growth schools exhibited strong evidence-based decision-making practice where teachers used the district’s benchmark assessment to reflect on instructional practice, used the core curriculum to guide instruction, and received frequent and high quality professional development on reading and math instruction. (Wang et al., 2012, p. 517)
These researchers argue that districts will benefit from creating school data management systems that integrate benchmark assessments and provide teachers with the capacity to effectively implement these assessments. However, many researchers have found that schools are not adequately building this capacity in educators through professional development and support (Abrams & McMillan, 2013). Abrams and McMillan (2013) found that teachers who use assessment results are not necessarily making strong associations between students’ conceptual misunderstandings and an appropriate instructional response, at a pedagogical level. This study of a division’s benchmark system is designed to produce clear guidelines for successful benchmark programs in order to help educators increase student achievement.

**Definition of Terms**

This investigation of benchmark assessments and their current uses in schools uses the following definitions:

- **Assessment:** An assessment should be designed to provide data about student learning, instruction, and curricula (Niemi, Wang, Wang, Vallone, & Griffin, 2007).

- **Benchmark assessment:** Operationally defined by Herman et al. (2010) as assessments that are administered periodically throughout the school year, at specified times during a curriculum sequence, to evaluate students’ knowledge and skills relative to a prescribed set of longer-term learning goals. The design and choice of benchmark assessments is dependent upon its intended uses.

- **Competency drivers:** Practitioners (e.g., teachers, district staff, and implementation team members) need to be trained and coached in order for
successful implementation of an initiative to take place. The competency drivers in this study will be the elementary school teachers in a school division who are responsible for implementation of the benchmark assessment system.

- **Curriculum**: “The written set of educational outcomes and associated content that students are to learn. This will include the knowledge, skills, and abilities we expect students to acquire or master after a period of appropriate instruction” (Bunch, 2012, p. 2).

- **Embedded case study**: A case that is embedded within a larger case study (Scholz & Tietje, 2003; Yin, 2003).

- **Formative assessment**: “A process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students’ achievement of intended instructional outcomes” (Council of Chief State School Officers [CCSSO], 2008, p. 3)

- **Instruction**: The delivery method of the intended learning outcomes.

- **Organization drivers**: Facilitative administrators (e.g., superintendents, principals, and other non-teaching staff) are key components of implementation and school change. They must establish a data system to guide the process of establishing innovation, create a hospitable environment for change, assess immediate outcomes, and create support for implementers. Data collected through assessment should be used to identify what types of and the amount of coaching needed (Implementation Drivers). The organization drivers in this study will be the deputy superintendent of curriculum, instruction, and assessment, in addition
to the organizational structures put in place for the benchmark assessment system (such as any capacity building done by the division, etc.).

- **Reliability**: Refers to the extent to which assessments are consistent in their measurement of a student’s ability or performance (Niemi et al., 2007).

- **Summative assessment**: “Any method providing information to aid in making judgements about the success of instruction or learning” (Schafer, 2013, p. 136).

- **Validity**: Refers to the accuracy of an assessment; does the assessment measure what it is intended to measure? There are four components of validity: content, criterion, construct, and concurrent (Grant & Gareis, 2015).
CHAPTER TWO

Review of Literature

Educators share the common goal of student achievement. Their involvement in this process depends on their role as a stakeholder: as parent, as student, as teacher, as building-level administrator, or as a division-level administrator. After the passing of *No Child Left Behind* in 2001, educators have been under intense scrutiny with regard to student performance on high-stakes assessments. As a result, educators have sought strategies for improving student performance on these assessments. The use of benchmark assessments has become a prevalent way for educators to measure student progress towards successful performance on these end-of-course assessments. However, research demonstrates that not all educators understand how to effectively use benchmark assessments and their resulting data in formative ways (Abrams & McMillan, 2013). It seems evident that several conditions are necessary in order for teachers to effectively use benchmark data to increase their students’ achievement (Abrams & McMillan, 2013; Goertz, Oláh, & Riggan, 2009; Symonds, 2004; Wayman & Cho, 2009). This literature review creates a common language with regard to assessments, and benchmark assessments in particular, and the ways in which they can be used to improve schools and to increase student achievement.

**Assessment in the Classroom**

A critical component of student achievement is assessment, and the classroom teacher is on the frontline of assessment. Curriculum and instruction are pivotal components of effective teaching; however, teachers who focus solely on these two areas are missing an important piece of the effective teaching puzzle. Curriculum and
instruction focus on the “what” and “how” of teaching, whereas assessment completes the model of teaching and learning by focusing on the “degree to which” students have learned (Gareis & 2015, 2015, p. 3). O’Malley et al. (2013) argue that assessment should be “consistent with curriculum sequencing” and “should be useful, providing actionable information for improving instruction” (p. 156).

There are three distinct roles of assessment in the classroom: pre-assessment, formative assessment, and summative assessment (Gareis & Grant, 2015). A pre-assessment is designed to measure student learning prior to instruction. A formative assessment is “the assessment of student learning integrated into the act of teaching” (Gareis & Grant, 2015, p. 5). The 2008 CCSSO defines formative assessment as “a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students’ achievement of intended instructional outcomes” (CCSSO, 2008, p. 3). Riggan and Oláh (2011) consider formative assessments to be short cycle assessments that have the potential to be one of the most powerful means to improve the quality of teaching and raise student performance.

Summative assessment intends to capture a snapshot of student learning at the end of an instructional period. Summative assessment can be defined as “any method providing information to aid in making judgements about the success of instruction or learning” (Schafer, 2013, p. 136). State-based high-stakes testing, such as the Virginia Standards of Learning tests, are examples of summative assessment. It is important to note, however, that summative assessments can be used for formative purposes, and formative assessments may also be used for summative purposes (Schafer, 2013). For example, a released Virginia Standards of Learning assessment can be used during the
course of the school year to inform instruction and to predict student performance on the 
SOL test at the end of the year. In order for assessments to be effective tools for 
improving student achievement, there are certain characteristics that an assessment must 
have.

Characteristics of effective assessment. In 2003, the Joint Committee on 
Standards for Educational Evaluation deemed four key characteristics of effective 
assessment: propriety, utility, feasibility, and accuracy. Propriety standards are focused 
on preventing students from “undue harm” during assessment (Gareis & Grant, 2015, p. 27). This includes keeping student data confidential and only sharing information with 
individuals who are directly related to the student’s learning, such as parents, counselors, etc. Propriety standards also note that assessments should be clear and fair, avoiding 
cultural biases that could systematically inhibit any groups from success on the 
assessment. Secondly, utility standards “remind us as teachers that our classroom-based 
assessments must be purposeful and practical” (Gareis & Grant, 2015, p. 27). Assessments results should be able to directly impact teaching and learning. Educators 
should be “timely and appropriate” with regard to communication (Yarbrough, Shulha, 
Hopson, & Caruthers, 2011). Furthermore, “evaluations should be conducted by 
qualified people who establish and maintain credibility in the evaluation context” (p. 15). The relevant data derived from assessments, and their subsequent evaluations, needs to 
made clear to all involved stakeholders.

The third characteristic of assessment is feasibility. The feasibility standards 
encourage teachers to think about whether or not our assessments are appropriate with 
regard to time. Is the assessment both effective and efficient? Is it appropriate for the
context in which it will be used? Educators must have the goal of helping “stakeholders understand the feasibility and value of addressing specific evaluation purposes at specific times in the program life cycle” (Gareis & Grant, 2015, pp. 31-32). Finally, accuracy standards “remind us as teachers that we must assess students to adequately and dependably represent student learning, so we can make decisions that ultimately support further learning” (p. 28). Educators must work to ensure that the assessment has met the needs of the particular goal that was set. “Treating all information as equally useful, evaluators must work with stakeholders to weight the relevance, scope, and accuracy of information” (Yarbrough et al., 2011, pp. 46-47). An assessment cannot meet the needs of educational stakeholders if it is not both valid and reliable. Validity and reliability are critical components of the accuracy stands; how can we ensure that our teacher-made assessment fulfill these requirements?

**Validity and reliability.** Teachers must ensure that their assessments are both valid and reliable. Assessment validity is concerned with “the truthfulness or appropriateness of decisions resulting from assessments” (Gareis & Grant, 2015, p. 33). It is the “extent to which inferences drawn from assessment results are appropriate” (p. 34). There are four attributes of validity: construct, content, criterion, and consequential. Construct validity refers to “the extent to which a test measures the construct it is supposed to measure” (Niemi et al., 2007, p. 12). Content validity “examines how well the items from a test represent the entire content domain to be measured” (Niemi et al., 2007, p. 11).

Brown and Coughlin (2007) define criterion validity as “The ability of a measure to predict performance on a second measure of the same construct, computed as a
correlation” (p. 2). A specific form of criterion validity commonly used in the world of assessment is predictive validity, which indicates how well a student will perform on a later assessment. Benchmark assessments are administered with the goal of predicting later performance on a summative assessment, such as a state standardized test at the end of the school year. Lastly, consequential validity has to do with “The appropriateness of the intended and unintended outcome that ensue from an assessment” (Gareis & Grant, 2015, p. 35). For instance, in the case of benchmark assessments, the actual use of results by teachers and administrators to make informed decisions and undertake effective instructional actions would be indicative of consequential validity.

An assessment needs to be both valid and reliable. Niemi et al. (2007) state that “Reliability addresses the extent to which a test consistently measures what it is supposed to measure and informs on how well the estimated test score reflects a student’s ‘true score’ on a test” (p. 10). If a test is found to be unreliable, it may be suffering from either systematic or random error. Systematic error decreases the validity of test scores and is typically a function of a formatting error. For example, “if there is a typo in the test which distorted the whole meaning of an item, all student performance will be affected in the same way due to that” (Niemi et al., 2007, p. 11). Random error is unsystematic and can differ between test-takers. There are various techniques one can utilize to evaluate the reliability of an assessment: test-retest reliability, parallel-form reliability, Cronbach’s alpha method, split-half method, and the Ge-Richardson method (Niemi et al., 2007). Many school systems fail to evaluate both the validity and the reliability of assessment, particularly if they are commercially made or when students perform well on that respective assessment (Brown & Coughlin, 2007; Oláh, Lawrence, & Riggan, 2010).
Furthermore, Brown and Coughlin (2007) found that evidence supporting the validity of state assessments is generally lacking, and many of our instructional decisions are made as a result of the data from these assessments. In short, the validity and reliability of assessments—especially consequential assessments—is paramount; however, there is evidence that validity and reliability are not always adequately ensured.

**Formative assessment.** Wayman and Cho (2009) found that teachers can improve student achievement through the use of data. The researchers analyzed how data systems should be implemented in schools and determined that, for successful implementation, “the uses must fit directly into the fabric of educator work” (p. 94). Wayman and Cho found that “data systems can help make administrative work more efficient while improving individual student outcomes” (p. 95). Some ways that data can be utilized to this end include: Prioritizing instructional time, targeting additional individual instruction for students who are struggling with particular topics, more easily identifying individual students’ strengths and instructional interventions that can help students continue to progress, gauging the instructional effectiveness of classroom lessons, refining instructional methods, and examining schoolwide data to consider whether and how to adapt the curriculum based on information about students’ strengths and weaknesses (Wayman & Cho, 2009). Professional development is critical to achieving these instructional goals and should be offered “prior to data system implementation” (p. 96). The goal of professional development should be to “help educators become proficient in using data systems in everyday practice” (p. 95). Through the implementation of an effective data system at the school level, teachers can work toward increasing student achievement.
Formative assessment is a practice that is “embedded within classroom instruction” (Goertz et al., 2009, p. 1). The CCSSO (2008) define this as “a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students’ achievement of intended instructional outcomes” (p. 3). Furthermore, the U.S. Department of Education has used its Race to the Top program to encourage school districts to develop formative assessments as part of comprehensive state assessment systems (Goertz et al., 2009).

Formative assessment is intended to provide teachers with meaningful data on individual students in order to make instructional decisions. The data produced by formative assessments initiates a feedback system in which interventions are informed by assessments followed by actuation processes, or as Halverson (2010) states, “structured occasions to turn assessment information into actionable knowledge” (p. 133). He calls for the setting up of actuation spaces (such as grade-level or team meetings) so that practitioners can reflect on the data provided. Halverson argues that, without actuation spaces, practitioners will have difficulty determining how any of the data they encounter could lead to improvements in teaching and learning. Supovitz and Klein (2003) insist that formative assessments provide a forum for teachers to discuss and test their ideas about what instructional strategies produce evidence of student learning. However, it is critical that these assessments are valid and reliable in the first place in order for the data to be both meaningful and useful to the classroom teacher.

**School improvement**

While teachers serve on the frontlines of assessment, principals also share the task of improving student achievement: “Expectations of educational accountability put
building principals at the center of public and political discussions related to the improvement of education” (Daresh, 2002, p. 153). DuFour and Marzano (2011) note, “Schools are to bring every student to dramatically higher standards of academic achievement. No generation of educators in the history of the United States has ever been asked to do so much for so many” (p. 1). One common effort to address this charge is for school principals to use data in order to make informed decisions regarding student learning and thus work to meet the needs of all students.

Principals play a critical role in supporting teachers through the analysis of student data. Kerr et al. (2006) found that data-driven decision-making can lead to improved school culture and teacher practice. After reviewing the literature available, these researchers summarized common findings to include “teacher reports of greater differentiation of instruction, greater collaboration among school faculties, and improved identification of students’ learning needs as a result of data use” (p. 501). Symonds (2004) also found evidence that underscores the importance of teacher use of data in the classroom. In that study, the Bay Area School Reform Collaborative surveyed 32 K-8 schools in the San Francisco Bay Area in order to uncover the characteristics of schools that are successfully closing the achievement gap. Symonds (2004) found that teachers at gap-closing schools are more likely to:

Use data, administer frequent assessment of students, receive professional development on analyzing low-performance student data, receive PD on linking low-performing student data to instructional strategies, have leaders that encourage or lead systemic inquiry into the gaps, discuss low-performing student
achievement data with colleagues, and visit each others’ classrooms to observe instructional strategies more frequently. (Symonds, 2004, p. 1)

The study’s overall recommendations include that schools need frequent and reliable data, and teachers need support in order to effectively use data. Marzano et al. (2005) found that principal leadership has a significant and positive relationship with student achievement. Principals serve an invaluable role as instructional leader, and one way to increase student achievement is to improve classroom assessments and their alignment with both curriculum and instruction. Based on the research of both Symonds (2004) and Marzano et al. (2005), principals use assessment results every day to make critical decisions at the school-wide level.

**Accountability Era**

In the era of accountability, teachers and principals are not the only educators who bear the responsibility of improved student achievement. Division administrators also play a pivotal role in the school improvement process. The implementation of *No Child Left Behind* in 2001 increased the pressure on school divisions to raise test scores, close achievement gaps, and turn around under-performing schools (Shepard, Davidson, & Bowman, 2011). Recent federal policies, such as NCLB, require the testing of curriculum standards at each grade level, with serious consequences for schools that fail to make ‘adequate yearly progress’ on state tests over a series of years, including conversion to a charter school or dismissal of the administration. (Ylimaki, 2014, p. 4)

*Education Week* compiled a variety of data from 2003 to 2015 in order to evaluate the success of NCLB. Some modest success includes the improvement of 4th and 8th
graders in reading and math; proficiency in these areas increased from 29.6 to 34.8 out of 100 (Solis, 2015). The National Assessment of Educational Progress (NAEP) combined proficiency rates also increased, but subgroups scores remained notably lower than Asian and White students. Unfortunately, the poverty gap has grown wider in this time span. “The combined NAEP proficiency rate for students in poverty increased from 14.2 in 2003 to 20.9 in 2015” (Solis, 2015, p. 2). Most states saw improvement, with Washington, DC, increasing its NAEP proficiency rate by 15 points (Solis, 2015). Interestingly, demographics shifted a great deal from 2003 to 2015. The percentage of non-White students enrolled in public education increased from 40.8 in 2002-03 to 49.8 in 2013-14 (Solis, 2015). “In the 2002-03 school year, Latino students made up 17.7 percent of enrollment in pre-K through grade 12. By 2013-14, one-quarter of all students were Latino” (Solis, 2015, p. 3).

With regard to the data generated by NCLB, Halverson (2010) argues that teachers are now under pressure to “turn assessment information into actionable knowledge” (p. 133). Abrams and McMillan (2013) note that teachers use assessment results in a variety of ways:

- identifying and addressing areas of student weakness, providing remediation for gaps in student learning, setting instructional priorities and increasing efficiency,
- determining instructional approaches such as whole class instruction, and differentiating instruction for small groups or customizing learning activities for individual students. (p. 110)

Furthermore, Daresh (2002) states, “The outcome of these measures is often a public statement that teachers or principals are not doing their jobs because students are not
performing well on tests…. this trend is likely to continue well into the future” (p. 161).

In 2011, Peter Meyer wrote,

Nine years after the enactment of No Child Left Behind, the public’s appetite for standardized tests appears undiminished. More than two in three Americans believe that the federal government should “continue to require that all students be tested in math and reading each year in grades 3–8 and once in high school,” whereas less than 10 percent actually oppose this requirement. (Meyer, 2011, p. 1)

While many Americans believe that we should be holding schools accountable for the education of our students, there is a debate regarding whether or not high-stakes accountability measures are the way to evaluate our teachers. Daresh (2002) notes, “An increasing number of states are seeking to determine accountability and effectiveness of schools through the use of testing programs that supposedly verify whether students are actually learning in their classrooms” (p. 161). However, this method evaluates educators based on relatively narrow assessment results without a holistic measure of student improvement and growth.

**Intended consequences.** *No Child Left Behind* was intended to increase transparency among educators and educational stakeholders, such as parents and the community (Camera, 2016). According to the NCLB Parent’s Guide (2003), the law was designed to “improve student achievement and change the culture of America’s schools” (p. 1). That report also provides an in-depth list of what NCLB does for both parents and children: supports learning in the early years (to prevent learning challenges later on), provides more information for parents about their child’s progress, alerts parents to important information on the performance of their child’s school, gives children and
parents a lifeline, improves teaching and learning by providing better information to teachers and principals, ensures that teacher quality is high priority, gives more resources to schools, allows more flexibility, and focuses on what works (NCLB, 2003). Supporters of NCLB argued that the law would “do for the quality of education what Brown v. the Board of Education did for the equality of America’s schools” (DuFour, DuFour, & Eaker, 2008, p. 45). The NAEP results, previously mentioned, are the main evidence that is drawn upon in order to argue that these goals have been achieved. However, DuFour et al. (2008) note, “The promise of booming student achievement as a result of increased accountability, sanctions, and parental choice has, to date, failed to come to fruition” (p. 45). While the intended outcomes of No Child Left Behind are valuable and important, there have been many unintended consequence that have arisen following its implementation.

Unintended consequences. Many states have revolted against the strict accountability system of No Child Left Behind. In 2011, for example, Montana’s education secretary explicitly stated that the Big Sky state would not be following NCLB. Denise Juneau, Montana’s education secretary, stated, “We won’t raise our annual [NCLB-mandated] objectives this year… and we’re not asking for permission” (Meyer, 2011, p. 1). Many states have since joined the fight, but the era of accountability shows no signs of ending any time soon. While the goals of NCLB and other related accountability measures may have been noble at the onset, there are a variety of unintended outcomes that have set many school systems into a tailspin. Larry Shumway, superintendent of schools in Utah, claims, “Pretty soon all the schools will be failing in America, and at that point the law becomes meaningless…. States are going to sit and
watch federal accountability implode” (Meyer, 2011, p. 1). In addition to the issues of struggling schools, other movements have arisen as a result of NCLB.

One of the unintended consequences of high-stakes testing has been the “opt-out movement,” a movement that almost every state has experienced to some degree (Harris, 2015, p. 1). “According to the state education department [of New York], last year about 49,000 (4%) didn't have a known valid reason for not taking the English test and 67,000 (6%) didn't take the math exam” (Wallace, 2015, p. 1). In 2015, the state superintendent of Indiana recommended that parents homeschool their students during the week of standardized testing in order to opt out (Wallace, 2015). FairTest, the National Center for Fair and Open Testing, encourages students and parents to “just say no to the test” (FairTest.org). This organization argues,

Testing overuse and misuse is damaging public education by eating up classroom time, narrowing curriculum and driving many students out of school. It is perpetuating a false narrative of failure and putting schools in low-income communities at risk of closure or privatization. (FairTest.org)

Harris notes,

There are generally few repercussions for students who do not take the tests, but if more than 5 percent of the student body at a given school or district opts out, that school may risk certain consequences, like greater monitoring or the loss of money for needy students. (Harris, 2015, p. 2)

Administrators hold varying opinions about the opt-out movement. Some administrators, such as the Indiana state superintendent mentioned above, support parents’ and students’ right to opt out. Salvatore Goncalves, superintendent of a school system in New York,
states, “Our board has taken a very strong stance against standardized testing” (Harris, 2015, p. 3). Goncalves argues that children are being tested too much. However, other educational leaders blatantly disagree with opt-out movements. James Crisfield, a former superintendent in New Jersey, argues, “I just worry about opting out as a conceit, that if it extends beyond PARCC, it will start eating away at the strength of public education” (Harris, 2015, p. 3). Regardless, there appears to be a widespread public and professional distrust of external standardized assessments, based on the pervasiveness of opt-out movements across the country.

In addition to opt-out movements, NCLB and other accountability measures have increased pressure for teachers and students alike. DuFour and Mattos (2013) argue:

> Principals are being asked to improve student learning by implementing mandated reforms that have consistently proven to be ineffective in raising student achievement. The current emphasis on using more intensive supervision and evaluation of teachers to improve school performance illustrates this irony.

According to Race to the Top guidelines, this more rigorous supervision process should influence a teacher’s professional development, compensation, promotion, retention, tenure, and certification. (p. 34)

Efforts have been made to revise NCLB, and the *Every Student Succeeds Act* of 2015 was designed to replace and improve upon the goals of *No Child Left Behind*.

**Every Student Succeeds Act.** On December 10, 2015, President Obama signed the *Every Student Succeeds Act* into law. ESSA was designed to “shed No Child Left Behind’s one-dimensional accountability system in favor of one that results in
measureable improvements” (Camera, 2016, p. 1). Chris Minnich, executive director of the Council of Chief State School Officers, noted,

As state agencies, we had a lot of information, but we didn’t do a good job putting it out in a way parents or teachers could engage with… I think the promise of ESSA is to go beyond transparency, to go into the idea that it’s not enough just to tell a school they’re not getting it done for kids, but we have to actually help that school get better. (p. 1)

However, it appears that ESSA is still missing the mark when it comes to implementation at the school level. Many states pleaded that the Department of Education need not be so heavy handed with regulations and specifics in defining various aspects of the accountability measures (Camera, 2016). Director of the National Education Association’s (NEA) education policy and practice, Donna Harris-Aikens, wrote,

We believe that the regulations miss the mark in terms of fidelity to the spirit and letter of ESSA, and instead revive elements of NCLB’s test, label, and punish system by adding the agency’s own restrictions on goals, indicators, weights, labels, interventions, and state plans. (Camera, 2016, p. 1)

As the federal accountability standards debate continues, teachers are still left with the task of helping students achieve on state standardized assessments. In an effort to increase student achievement, one strategy commonly undertaken by schools is the use of benchmark assessments. Despite their prevalence, educators often lack a common language for and understanding of how benchmark tests should be created and how we should use their results to increase student learning.

**Benchmark Assessments**
Benchmark assessments are intended to serve as an umbrella structure to support curricular planning, assessment, and feedback among them…. Benchmarks have a hierarchical structure that allows educators to think about benchmarks as embedded within educational goals, spanning different age-ranges, and involving activity across multiple domains of child development. (Feldman, 2010, p. 234) Benchmarks evaluate student knowledge and skills, typically within a limited time frame, and the results of these assessments can be aggregated and analyzed across classrooms, schools, or even districts (Perie, Marion, & Gong, 2009). By providing benchmark assessment results and a forum for discussing them, school leaders can provide teachers with guidance on how to make mid-year adjustments (Supovitz & Klein, 2003). Shared benchmark assessments can be powerful, because they provide teachers and administrators with common student performance data, presumably based on similar curricular coverage, which provide comparative feedback across classrooms (Supovitz & Klein, 2003). The information can be used formatively to increase student achievement and to inform classroom instructional practices.

There are three core purposes for benchmark assessments: instructional, evaluative, and predictive (Bulkley, Christman, Goertz, & Lawrence, 2010). Instructional purposes “provide results that enable educators to adapt instruction and curriculum to better meet student needs” (Perie, Marion, Gong, and Wurtzel, 2007, p. 4). An evaluative purpose means that the assessment is being used to “enforce some minimal quality through standardization of curriculum and pacing guides” (Perie, Marion, Gong, & Wurtzel, 2007, p. 5). Lastly, predictive assessments are “designed to determine each
student’s likelihood of meeting some criterion score on the end-of-year tests” (p. 5). Furthermore, there are several key criteria necessary for appropriate and effective benchmark assessments: validity, alignment (to both the curriculum and to the intended purposes of the assessment), reliability, fairness, high utility, and balance (Herman et al., 2010). These criteria were also referenced in the previous section regarding characteristics of effective benchmark assessments. In addition to validity and reliability, an assessment must be fair and unbiased: “A fair test is accessible and enables all students to show what they know; it does not advantage some students over others. Bias emerges when features of the assessment itself impede students’ ability to demonstrate their knowledge or skill” (Herman et al., 2010, p. 6). Utility is another critical component of effective benchmark assessments. To determine the utility of an assessment, one must ask the following question: “How useful will this assessment be in helping us to accomplish our intended purposes?” (p. 7). It is critical for test creators to understand the purpose of the benchmark assessment and to ensure that that purpose is also transmitted to the teachers who are administering it.

In current practice, benchmark assessments are designed either to predict future performance, such as on a state test, or used to assess student mastery at that particular point in the pacing guide for the respective subject and grade level (Olson, 2005). When planning a benchmark assessment, Perie et al. (2009) argue that there are five questions educators must ask:

What do we want to learn from this assessment? Who will use the information gathered from this assessment? What action steps will be taken as a result of this assessment? What professional development or support structures should be in
place to ensure the action steps are taken and are successful? How will student learning improve as a result of using this interim assessment system and will it improve more than if the assessment system was not used? (p. 7)

Oftentimes, school systems rely on previously used or commercially created benchmark assessments, trusting that they have been previously evaluated utilizing these criteria. Perie et al. (2009) claim that benchmarks should be created and implemented differently depending on their intended purpose. For example, if a benchmark assessment is being used for instructional purposes, it “should provide results that enable educators to adapt instruction and curriculum to better meet student needs” (Perie et al., 2009, p. 8). If a benchmark assessment is designed for evaluative purposes, then its primary goal is to provide information to help the teacher, school administrator, curriculum supervisor, or district policymaker learn about curricular or instructional choices and take specific action to improve the program, affecting subsequent teaching and thereby, presumably, improving the learning. (Perie et al., 2009, p. 9)

Lastly, a predictive benchmark assessment is “designed to determine each student’s likelihood of meeting some criterion score on the end-of-year tests” (p. 10). Predictive benchmarks are most commonly used in school districts that are trying to gauge where students will score on end-of-course standardized tests and how to respond to those predictions accordingly. When selecting the appropriate benchmark, there are many pitfalls into which educators can stumble.

Researchers have found that there are common variables in districts with ineffective benchmark assessment practices. For example, Davidson and Frohbieter
(2011) found that administrators (both at the school and district level) often hold different perspectives on the purposes, uses, and quality of an assessment system. Davidson and Frohbieter analyzed interview responses from 24 district administrators and 14 principals in seven districts across two states. As a result of the study, the researchers discovered that professional development is often lacking in the area of assessment, and the actual benchmark assessment selected may not necessarily reflect the district’s intent for its implementation (Davidson & Frohbieter, 2011).

Other common issues with benchmark assessments have to do with the test themselves. Oftentimes, the validity of locally developed assessments is not adequately analyzed (Brown & Coughlin, 2007). While commercially developed benchmarks should have been checked for validity, school districts often blindly trust in this process without further analysis (Brown & Coughlin, 2007). Furthermore, when utilizing commercially developed benchmark assessments, educators often fail to judge them against their intended uses (Brown & Coughlin, 2007).

**Teacher Competencies**

In order for benchmarks to be effective, the teachers who administer them must be competent in areas of assessment. Abrams and McMillan (2013) found several factors that contributed to teachers’ successful implementations of benchmark assessments. Some of these variables include clear and consistent district- and building- level expectations for teachers’ role in the use of benchmark testing and clear expectations for administration and subsequent remediation derived from benchmark data (Abrams & McMillan, 2013). Abrams and McMillan (2013) found several additional variables that ensured effective and efficient use of benchmark testing. The researchers found that
teachers needed immediate access to the results of the assessment; needed regularly scheduled meetings to collaborate with other teachers regarding benchmark data; needed sufficient time to be able to review and reteach benchmark items; and needed to possess the ability to identify weaknesses in student learning and to remediate with regard to those weaknesses (Abrams et al., 2015). Furthermore, the strength of the benchmark itself was important with regard to its alignment with content of instruction, use of high-quality test items, and the accuracy of scoring (Abrams et al., 2015). While teachers use benchmark assessment results, they are not necessarily making strong connections between students’ conceptual misunderstandings and the appropriate instructional response (Abrams & McMillan, 2013). Based on their research, Abrams and McMillan (2013) determined that it is unclear what pedagogical connections teachers are making between remediation and the nature of the students’ misunderstanding in the first place.

Other researchers have found that teacher interest or “buy-in” to the benchmark process is a critical component of successful implementation. Bancroft’s (2010) study revealed that, for her particular subjects, teachers found benchmark testing to be an interruption in valuable class time. Bancroft (2010) notes, “If teachers have little buy-in for a reform, their resistance can overturn the reform’s best intentions” (p. 55). In this particular study, benchmark testing was found to be ineffective. A contributing factor to this failure was the attitude of the educators involved: “School participants—teachers and administrators—are ambivalent at best regarding the benchmark testing practice and see little evidence that the teaching/reflection/re-teaching process has any efficacy in terms of boosting either test scores or more authentic student learning” (Bancroft, 2010, p. 18). Davidson and Frohbieter (2011) also found that it was critical for both administrators and
teachers to be on the same page with regard to the objectives of benchmark assessments. Building capacity for assessment literacy is also lacking in many schools and is a contributing factor to ineffective benchmark systems (Davidson & Frohbieter, 2011).

When administered correctly, benchmarks can be an invaluable way to increase student achievement and, thus, improve schools. In fact, many teachers have reported that benchmark test results helped them monitor student progress and identify skill gaps for their students and led them to modify curriculum and instruction (Goertz et al., 2009). It is imperative that educators develop a shared understanding of effective benchmark systems: how to design benchmarks, how to implement them, and how to utilize the data that they generate in an effort to improve our schools, particularly at the elementary level. Successful implementation at the elementary level is imperative, because the learning that occurs in elementary school serves as the foundation for our students’ educational futures.

There is limited research with regard to the impact of benchmark testing and the formative uses thereof (Abrams et al., 2015). Many studies have failed to examine how individual teachers actually analyzed and used the data to inform their classroom practice, the policy conditions that supported teachers’ ability to use benchmark assessment data to improve instruction, or the interaction of benchmark assessments with other classroom assessment practices (Goertz et al., 2009). This study of a school division’s benchmark assessment seeks to close the gap in literature by analyzing the ways elementary school teachers utilize benchmark results in their respective classrooms.
CHAPTER THREE

This study was designed to analyze the manner in which teachers implement benchmark assessments in the classroom setting in order to progress study learning. There were six research questions in this study:

1. How do teachers describe their intent in making use of the benchmark assessment system?

2. How do teachers describe their practices of making formative use of the benchmark assessment system?

3. What do teachers describe as the perceived outcomes of the benchmark assessment system?

4. What are the similarities and differences between district intent and design of the benchmark assessment system and teacher intent and practices?

5. How competent do teachers feel themselves to be to make use of the benchmark assessment system to progress student learning? (a) Why do teachers feel more or less competent? (b) What do teachers feel would contribute to their competency?

6. How are teachers’ responses similar and/or different relative to the state accreditation status of their respective schools?

An embedded case study of participants in three elementary schools within a school division was conducted in order to look for themes among teacher use of benchmark data.

Method

A case study is the appropriate method to conduct this study because it is an “in-depth study of one or more instances of phenomenon in its real-life context that reflects the perspective of the participants involved in the phenomenon” (Gall, Gall, & Borg,
2007, p. 447). Gall et al. (2007) note that examples of phenomena include “programs, curricula, roles and events” (p. 447). The phenomenon in this study is the use of benchmark assessments and their presumed role in increasing student achievement. The present study is an embedded case study, because three elementary schools were studied within the larger context of a school division, as depicted in Figure 3. A benefit of the embedded case study research methodology is that this type of research is often considered more compelling than single-case studies and is “more likely to lend [itself] to valid generalization” (Fraenkel & Wallen, 2009, p. 431). Given the prevalent use of benchmark assessments among K-12 public schools, increased generalizability is useful. Furthermore, the cases are “chosen in order that theories can be generated about a larger collection of cases. In this way they employ a very different mode of thinking from the single case study” (Wellington, 2015, p. 166). By sampling teachers from three different elementary schools within the same school district, the researcher hopes that the results will be more generalizable than a single case study conducted just at one school. Figure 3 demonstrates the framework of an embedded case study.

Figure 3

*Embedded Case Study Framework*
The structure and characteristics of an embedded case study provide the methodological framework for this study. The term “embedded case study” typically refers to a case that is embedded within a larger case study (Scholz & Tietje, 2003; Yin, 2003). Figure 4 specifically illustrates that the elementary schools selected for this study fit within the larger context of the school division.

Figure 4

![A School Division](image)

**Embedded Case Study: A School Division**

In this study, there are several cases (the three elementary schools) embedded within the larger case (the school division). The benchmark assessment system in the division is a significant component of the larger context of this study. This study evaluates the alignment between division and teacher goals for benchmark assessments, as well as the overall climate for benchmark system implementation established by the division. Figure 5 demonstrates how these elements of an embedded case study are connected to the Implementation Framework, particularly with regard to competency drivers and organization drivers.

Figure 5
Using the Implementation Framework as a conceptual organizer, the school division’s benchmark assessment system is the overarching context for this study (Implementation Drivers, 2016). The teachers interviewed at the three elementary schools are the competency drivers of the benchmark system, and school and district leaders serve as the organization drivers. This study includes an analysis of organization drivers in the school division, because teachers’ alignment to the division’s goals for benchmark assessments is presumed to be important (Herman et al., 2010; Riggan & Oláh, 2011; Supovitz & Klein, 2003). Furthermore, the deputy superintendent for curriculum, instruction, and assessment, was interviewed. Thus, the researcher analyzed the administrative vision for benchmark testing and the supports set in place to increase teacher competency. For the scope of this study, the researcher was only able to analyze the organizational drivers and competency drivers in place within the school division. The study is designed to analyze the alignment between division intent and teacher intent and practice with regard to benchmark assessments.

The purpose of this embedded case study was to analyze the manner in which teachers at the school level use benchmark assessment results to improve instruction and
to increase student achievement within the context of a school division with a sustained history of and system for benchmark assessments. A case study can have one of three purposes: “to produce detailed descriptions of a phenomenon, to develop possible explanations of it, or to evaluate the phenomenon” (Fraenkel & Wallen, 2009, p. 451). The goal of this case study is to evaluate the phenomenon of benchmark testing in today’s elementary school setting. Upon the conclusion of the study, the researcher made a generalization about the results. That is, the researcher made determinations about benchmark assessments “that apply to more than one individual, group, object, or situation” (p. 432). These generalizations focus on the manner in which teachers at the three elementary schools utilize benchmark assessment results in the classroom setting within a school division.

Participants

Wellington (2015) argues that the design of a case study is like a funnel:

The start of the study is the wide end: the researchers scout for possible places and people that might be the subject or the source of data, find the location they think they want to study, and then cast a net widely trying to judge the feasibility of the site or data source for their purposes. (p. 164)

For the purposes of this study, this school division was selected due to its proximity to the researcher and due to the size of the school system. The school division is the 14th largest school system in Virginia, and the size of the division increases the chances of finding elementary schools that meet the desired profiles with regard to state accreditation requirements.
The school division was selected for the study due to the large size of the school system and the varying levels of both state and federal accreditation therein. The town itself boasts a rich history, as it was established in 1610. The population is estimated at 137,000 with a median household income of $50,705. The community is racially diverse, and the most recent census data reports that 49.6% of the population is African American, 42.7% is white, and the remainder includes bi-racial, Asian, American Indian, and Hispanic (census.gov). The demographics of the public school population differ from that of the community itself. Of the current student population, 59.7% of the students are African American, 25.3% are White, 6.1% are Hispanic, 6.2% are bi- or multi-racial, 2.1% are Asian, and less than 1% are American Indian or Hawaiian/Pacific Islander (census.gov). With regard to demographics, more than 530 students in the division are from 47 different countries. These students speak 46 different languages, adding to the diversity of the school division.

This school division is comprised of one early childhood center, nineteen elementary K-5 schools, one gifted center, two PK-8 schools, five middle schools (including one fundamental and one magnet), and four high schools (including a specialized academic program). The division’s website illustrates both the mission and vision of the school system. The mission statement reads, “In collaboration with our community, [this school division] ensure academic excellence for every child, every day, whatever it takes.” The vision statement reads, “[This school division]: the first choice for success for every student.” The website also underscores the core values:

We believe that the developmental needs of children are central to every aspect of the operation of [this school division] and that all interactions with our
stakeholders must be governed by our core values- integrity, responsibility, innovation, excellence, and professionalism.

This school division is the 14th largest in Virginia and has a student population of 20,358. There are 1,537 teachers, and 773 of those teachers have a Master’s degree or higher. Eighty-two teachers are National Board Certified. There are 20 Advanced Placement courses offered. The division’s budget for 2016-17 is $200,450,417, and the estimated per pupil expenditure is $11,302. In 2016, the division graduated 1,300 students, and more than $30.5 million was awarded to students in the form of grants and scholarships. Among the students who graduated, 738 planned to attend a 4-year college, and 215 planned to attend a 2-year college. Of these students, 91% graduated on time, according to the Virginia Department of Education’s On-Time Rate Schedule. The average pupil/teacher ratio for Grades K-3 is 1:23, and the ratio for Grades 4-5, for middle school, and for high school is 1:25.

The division has implemented a new strategic plan for 2016-2020, which was adopted on August 19, 2015. In order to create the strategic plan, the division gathered input in over 45 sessions with teachers, administrators, parents, and community members. There were over 630 total participants. The strategic plan connects with NCLB and ESSA, because the introduction states, “virtually everyone concluded that setting lower standards would be an unspoken agreement to leave some children behind; and that, we will not do” (https://www.hampton.k12.va.us/about/Vision2020.pdf). The current strategic plan boasts several accomplishments achieved by this school division during the implementation cycle of the 2010-2015 strategic plan. Among these accomplishments are the decrease in the truancy rate by 46%, the honor of receiving the All America City
Award in 2014, the opening of the new and state-of-the-art PK-8 combined schools in 2010, and the increased in graduation rate by 13% from 2008.

However, the school division also realistically notes areas in which growth is needed. For instance, only 40% of the schools in this division are accredited, a decrease from 97% in 2010. Another challenge faced by this division is the increase in the number of students who are eligible for free/reduced-price meals: 58% as opposed to 47% in 2010. Furthermore, 515 students are classified as homeless, an increased from 211 in 2010. The division also acknowledges major achievement gaps, particularly with regard to students with socio-economic and disability status. On a recent school climate survey, only 44% of the students surveyed reported that they felt challenged by their school work, and only 34% claimed that lessons related to life beyond the school walls. In order to address these concerns, the strategic focus model underscores several goals for the upcoming school years: maximizing every child’s learning; creating safe, nurturing environments; attracting, developing, and retaining exceptional staff; enhancing family and community engagement and satisfaction; maintaining effective, efficient, and innovative support systems; and managing fiscal resources effectively and efficiently.

Benchmark assessments connect to several of the goals of the division’s strategic plan. For example, one of the goals is to increase support for students by providing frequent formative assessment feedback in order to more rapidly identify students who need more help and to provide that help “without delay” (p. 9). Furthermore, a component of the strategic plan is to provide professional development differentiated by employee needs. One form of professional development mentioned is to develop a “formative assessment system” (p. 13). There is also a plan in place to ensure that staff and students are
technologically liter in order to “successfully navigate online…assessment systems” (p. 14).

**Elementary schools and teacher participants.** Three elementary schools in the division were selected based on prior state accreditation results, with the intention of representing three characteristic situations relevant to accreditation status and progress. Of the three selected elementary schools, one school has not met state accreditation requirements within the last five years (2015-2016, 2015-2014, 2014-2013, 2013-2012, and 2012-2011); one school has recently met state requirements but had not done so prior to the last few years; and one school has consistently met state accreditation requirements over the last five school years. Three teachers were interviewed at each school. Teacher participants must have taught in the division for a minimum of five years in order to ensure experience with the current benchmark system. It was important to select teachers who have worked under the benchmark system in order to get a more accurate impression of benchmark testing in the division at the elementary school level. The building administrators were the gatekeepers to these participants. The researcher asked the school principals to identify teachers within the school who have worked under the benchmark system for several years and who were known to be teacher leaders in their respective elementary school. Administrators were kept abreast of the study and the interview schedule, but teacher responses were kept confidential in order to protect the participants and to encourage their honest responses. The assistant superintendent in charge of assessment was also interviewed in order to gain a deeper understanding of the benchmark assessments and their role in elementary schools within the larger context of the school division.
School A. The first school selected has struggled to meet state accreditation standards over the last several years. Currently, this school is not accredited. School A is a Title I school, and its staff includes three reading interventionists, a math interventionist, and a family engagement specialist. The school currently serves approximately 400 students. Currently state accreditation results indicate that the accreditation benchmark was not met in either English or in Science, with one-year averages of 65% and 40% respectively. However, School A averaged 70% in math and, thus, met accreditation standards in that subject. Due to the three-year average of 80% in History, School A met the accreditation benchmark in that subject. A third grade, a fourth grade, and a fifth grade teacher were selected by the principal to participate in this research; all three teachers are members of the school leadership team. The 3rd grade participant has taught in the division for 7 years, the 4th grade participant has taught in the division for 3 years, and the 5th grade participant has taught in the division for 11 years.

School B. School B met all state accreditation benchmarks in all four subjects during the 2016-2017 school year. However, scores in all four subjects increased from 2014-2015 to 2015-2016 but decreased, for the most part, in the following year. English increased from 63% to 76% in the first two years, and the average increased to 77% for 2016-2017. However, scores increased in math from 71% to 82% in the first year, but decreased from 82% to 74% in 2016-2017. Similarly, averages in History increased from 73% to 94% and then dropped to 87% last year. Science scores increased from 58% to 78% and dropped to 75% for the 2016-2017 school year. A third grade, a fourth grade, and a fifth grade teacher were selected by the principal to participate in this research.
The 3rd grade participant has taught in the division for 5 years, the 4th grade participant has taught in the division for 5 years, and the 5th grade participant has taught in the division for 9 years.

**School C.** School C’s scores have notably increased from 2014. This year, state accreditation standards have been met in English, in math, in History, and in Science. This fully accredited school has seen an impressive improvement in state testing averages. English scores increased to 85% from 75% in 2015-2016 and 66% in 2014-2015. Math averages have increased to 85% from 71% in 2014-2015. The average in History is 94%, an increase from the 88% average earned in both previous years. Science averages have increased the most: the current average is 87% and the average in 2015-2016 was 62%. Three teachers were selected from School C, by the school’s principal, to participate in this study. The teachers were chosen from third grade, fourth grade, and fifth grade. All three participants have been teaching in the division for at least seven years. The 3rd grade participant has taught in the division for 13 years, the 4th grade participant has taught in the division for 15 years, and the 5th grade participant has taught in the division for 7 years.

**Sampling.** The participants are a homogenous sample, and the commonalities among the participants include the fact that they are all elementary school teachers and they have all worked under the same benchmark system for several years. The principals at each elementary school selected a teacher representative from third, fourth, and fifth grade. This school division refers to their benchmark tests as critical skills assessments. The deputy superintendent for curriculum, instruction, and assessment the division states that curriculum leaders and teacher specialists create the benchmark assessments at the
division level. Feedback on the benchmark drafts is gleaned from select teachers in the division. These assessments are also revised each year, and teachers receive a blueprint prior to administration so that they are aware of how many questions align with each Virginia SOL reporting category. The deputy superintendent reports that the division has utilized this benchmark system for the past 15 years. Currently, benchmarks are administered at the elementary school level grades 3-5. Students in Grade 3 complete quarterly benchmarks just in math and in English, while Grade 4 and Grade 5 students complete quarterly benchmarks in all four core subject areas. In order to gain more information about the benchmark system in the division, the researcher utilized interviews as the primary instrument of data collection.

**Instrumentation**

One method of data collection in this study is interviews. The researcher conducted interviews with the participants in order to gain information about the manner in which teachers utilize benchmark test results in the classroom. Tuckman (1999) states that interviews help researchers to convert into data the information they receive directly from [research subjects]. By providing access to what is ‘inside a person’s head,’ [this approach] allows investigators to measure what someone knows (knowledge or information), what someone likes and dislikes (values and preferences), and what someone thinks (attitudes and beliefs). (p. 237)

Furthermore, Tuckman (1999) argues that there are three questions researchers must consider in order to address issues of validity while constructing interview questions:
To what extent might a question influence respondents to show themselves in a good light? To what extent might a question influence respondents to attempt to anticipate what researchers want to hear or learn? To what extent might a question ask for information about respondents that they may not know about themselves? (p. 237)

These questions are critical to creating and maintain a successful interview protocol that makes the participants feel safe in sharing honest responses. The researcher utilized open question phrases such as “please describe your process” in order to encourage participants not to feel pressured to respond with a “correct” yes or no answer. The participants were also informed that all responses were kept anonymous in order to prevent them from feeling pressured to respond in a certain way in the event that their superiors, such as building administrators and central office staff, would have access to their answers. Furthermore, participants were asked to share an anecdote in which the division benchmark system worked the way it is intended. This question, and the question regarding their competency in administering benchmarks, challenged participants to share information they may not have known about themselves.

Interview protocol. Interviews “consist of oral questions asked by the interviewer and oral responses by the research participants” (Gall et al., 2007, p. 228). While group interviews and focus groups have become increasingly prevalent in contemporary research, individual interviews were conducted for this study. A focus group concept was considered but then rejected because participants may feel more comfortable sharing their true feelings without the presence of colleagues, particularly department chairs or administrators. While there are benefits to selecting the method of
questionnaires over interviews, the researcher selected interviews because “questionnaires cannot probe deeply into respondents’ beliefs, attitudes, and inner experience” (Gall et al., 2007, p. 228).

A protocol was established with participants prior to the interview. The interviewer scheduled an appropriate time and meeting space for each interview. Saldana (2011) notes that the interview space must feel “comfortable and secure” (p. 35). Once an appropriate location was found, the interviewer reserved the space and secured a meeting time with the participant. The participant was contacted at least one day prior, either via email or phone, as a courtesy reminder or to reschedule, if necessary (Saldana, 2011). In the interview protocol and the construction of the interview questions, the researcher made sure to avoid asking multiple questions within a single question and asking either/or questions (Saldana, 2011). All responses were audio recorded, with consent of the participants. The participant consent form is included in Appendix A.

There were two sets of interview questions: one for the deputy superintendent of curriculum, instruction, and assessment and one for the selected teachers at the elementary school level. The interview questions for the deputy superintendent are included in Appendix B. The teacher interview questions are listed in Appendix C. Follow-up questions were asked when the researcher needed to clarify a point made by an interviewee or when more information was needed. For example, after the deputy superintendent responded to the question regarding the helpfulness of benchmark assessments, the researcher asked, “How do you know that this is the case?” Both sets of interview questions were piloted prior to being administered in the division. The researcher piloted the deputy superintendent’s interview questions with his administrative
counterpart in a neighboring school division, and the teacher interview questions were piloted with two teachers at a neighboring school division.

Upon completion of the pilot interviews, the administrative interviewee shared some key insights regarding the interview protocol and questions. The accepted recommendations include allow for follow-up responses in the event that interviewees think of information they wish they had shared after the interview; move a question about the potential unintended consequences of benchmark testing after the question about the benefits of this system; add a question, or follow up question, regarding scoring; ask a question about whether the benchmarks are scored online or if teachers are scoring their own students; and give the participants a copy of the interview questions to help them process as the researcher moves through the interview. The teacher interview questions were piloted with two teachers at a neighboring school division. The changes made as a result of these interviews are as follows: add a question about how the school administers the benchmark assessments and the teacher’s role in that process, provide participants with a copy of the interview, add question about whether there is a mandated benchmark system, and add a question about how benchmark data impacts their individual teacher evaluations.

Validity and reliability. There are often concerns about validity and reliability in a qualitative research study. The fact that this study is an embedded case study improves its reliability, because “data [was] analyzed and reported at the group level,” in addition to the responses from the individual participants themselves (Gall et al., 2007, p. 229). Furthermore, this method of data collection is typically more valid and reliable because the researchers are “typically collecting information that is highly structured and
more likely to be accurate” (p. 229). Yin (2008) argues that one can judge the quality of a case study design by analyzing three types of validity criteria and one reliability criterion:

(1) Construct validity is the extent to which a measure used in a case study correctly operationalizes the concepts being studied. (2) Internal validity is the extent to which the researcher has demonstrated a causal relationship between X and Y by showing that other plausible factors could not have caused Y. The criterion of internal validity is not applicable to descriptive case study research, because it does not seek to identify causal patterns in phenomena. (3) External validity is the extent to which the findings of a case study can be generalized to similar cases. (4) Reliability is the extent to which other researchers would arrive at similar results if they studied the same case using exactly the same procedures as the first researcher. (p. 24)

By being aware of these concerns with regard to case study research, the researcher hoped to address issues of construct validity, internal validity, external validity, and reliability in order to make the study results both replicable and generalizable to a larger population. Construct validity was achieved as the researcher sought to operationalize the variables of interest. Furthermore, the researcher remained aware of her expectations and biases as she entered the research. Internal validity was not as grave of a concern, because “it does not seek to identify causal patterns in phenomena” (Yin, 2008, p. 24). While the researcher sought to identify patterns in the interview responses, this study was not designed to prove causality among any of the variables. Lastly, the researcher sought to improve the reliability of the study by asking open-ended questions. The researcher
avoided “yes” or “no” questions in favor of encouraging open dialogue among participants. Also, the researcher also improved reliability by utilizing coding categories from a replicable study (Abrams & McMillan, 2013).

Furthermore, the validity of this study was strengthened through the triangulation of data. Kemmis (1983) notes that “what makes the case study work ‘scientific’ is the observer’s critical presence in the context of occurrence of phenomena, observation, hypothesis-testing (by confrontation and disconfirmation), triangulation of participants’ perceptions, interpretations, and so on” (p. 103). In this study, the triangulated data included teacher interview responses, sample elementary-level benchmark assessments, benchmark data, and any relevant teacher documents shared during the course of the interview.

The researcher also controlled for issues of reliability and validity by keeping a reflective journal, member checking, peer examination, and collaboration with participants. Merriam (1998) states, “Validity and reliability are concerns that can be approached through careful attention to a study’s conceptualization and the way in which the data were collected, analyzed, and interpreted, and the way in which the findings are presented” (pp. 199-200). In this study, the researcher used “multiple sources of data” in order to construct “plausible explanations about the phenomena being studied” (p. 204). With regard to member checking, the researcher took “data and tentative interpretations back to the people from whom they were derived and ask them if the results are plausible” (p. 204). Furthermore, the participants were involved throughout the research process, and not just to clarify the findings. Prior to the interview, the researcher shared the consent form, located in Appendix A. The consent form listed the research questions
and described the purpose of the researcher’s study. The participants were also encouraged to ask questions of the researcher before, during, and after the interview. In addition to collaborating with participants, the researcher also asked colleagues to comment on the findings as they emerged, in order to add to the validity of the study through the use of peer examination. Throughout the course of this study, it was important for the researcher to clarify “the researcher’s assumptions, worldview, and theoretical orientation at the outset of the study” (Merriam, 1998, p. 205). By addressing these concerns, the internal validity of the study was strengthened.

Reliability can often be problematic in qualitative research, because “human behavior is never static” (Merriam, 1998, p. 205). However, qualitative research is intended to describe and explain the phenomena as the participants experience it, and “the reliability of documents and personal accounts can be assessed through various techniques of analysis and triangulation” (p. 206). In order to assess the reliability of a study, the researcher determined whether the results were consistent with the data collected. Merriam (1998) notes, “Just as an auditor authenticates the accounts of a business, independent judges can authenticate the findings of a study by following the trail of the researcher” (p. 207). By providing enough detail, the researcher ensured that the study in replicable, thus increasing the reliability of the research.

**Data Collection**

Interviewing was a critical component of this case study research. The researcher also created and maintained case records. The case record included “lightly edited, ordered, indexed and public version of the case data,” while the case data included all material collected (Wellington, 2015, p. 171). The interview responses were transcribed
and coded for commonalities. For critical background knowledge of the benchmark assessments, the researcher worked with the assessment coordinator for the division to identify key stakeholders in benchmark assessment creation at the elementary school level. The researcher gained information about the process, the stakeholders, and the evaluation of the assessments with regard to validity and reliability.

**Data Analysis**

All interviews were recorded, and the responses were transcribed. Wellington (2015) notes four valuable principles for high-quality analysis in a case study:

1. It considers or attends to all the evidence.
2. The analysis must consider and weigh up possible and plausible “rival interpretations” and explanations, i.e. alternative ways of viewing the data.
3. The analysis must address the “most significant aspect” of the case study, e.g. by focusing on the key issue or issues.
4. Researchers should use their own “prior, expert knowledge” in analyzing their data. (p. 173)

The responses were coded in order to find frequency of views or ideas based on nine key categories illustrated in Appendix C. All coding categories were replicated from Abrams et al. (2015). These codes were developed a priori from a similar study conducted by Abrams et al. (2015). The coding categories were deemed to be appropriate for this study, because the aligned with the research questions. The coding categories that aligned with Research Question 1 were the “value of benchmark testing” and the “usefulness of benchmark testing.” The coding category that was used to analyze research question 2 was “benchmark testing policy.” The coding categories that aligned with research question 3 were “receipt of test results” and “instructional uses of results.”
For research question 4, “expectations for teacher use” was the coding category utilized. “Obstacles/barriers to using test results” was the category used to analyze research question 5, and “recommendations to improve the practice” was used for research question 6. Furthermore, Appendix D includes a table that outlines each research question and its corresponding Implementation Driver(s) and the related interview question(s).

Gall and colleagues (2007) note that developing the coding categories is one of the most important steps in qualitative research. The researchers note the importance of creating categories that directly relate to the data itself. After developing the category system, the researcher uses it to code each segment. A segment is defined as “a section of the text that contains one item of information and that is comprehensible even if read outside the context in which it is embedded” (p. 466). It is critical to examine each segment in order to determine under which category it falls within the category system.

With regard to data analysis, the researcher needed to analyze the applicability, or generalizability, of the findings (Gall et al., 2007, p. 477). This was accomplished by using the constant comparison method of data analysis. Constant comparison refers to the “continual process of comparing segments within and across categories” (p. 469). By using the multiple-case design, the researcher analyzed relational or causal patterns, and the generalizability of constructs and themes was assessed. Patterns became evident, because the results were compared between several participants in three different locations. Structural analysis was also utilized to explore the patterns evident in the data; in this process, case study data was examined “for the purpose of identifying patterns inherent in discourse, text, events, or other phenomena” (Gall et al., 2007, p. 471).
example, similar patterns emerged at School A with regard to the intent of benchmark assessments. Participants in this school shared similar responses that were not replicated at either School B or School C. Furthermore, the absence of patterns was also critical to the findings of this study. For example, two teacher participants at School C shared vastly different approaches to sharing student results.

The data were triangulated through the comparison of teacher and deputy superintendent interview responses and any benchmark documentation that participants chose to share. The participants were also asked, during the course of the interview, to share any relevant documents, in addition to the sample benchmarks and results. With regard to relevant documents, teachers were asked to share any classroom assessments or data that can shed light on benchmarks in the individual teacher’s classroom. These documents could include teacher-made classroom assessments, benchmark data analysis and reports, remediation plans, and anything else that the teacher believes to be relevant. Interestingly, no participant chose to share any data. This may be due to the fact that teachers just receive data during team meetings, and they do not have access to specific student data after the benchmarks.

**Considerations**

Both the names of the participants and the names of the schools were changed in the course of this study in order to protect the identity of the educators and to encourage openness among teachers and administrators at the school division level. According to Gall et al. (2007) there are four types of ethics through which researchers should view their cases. The first type of ethics is utilitarian ethics, which requires researchers to be aware of the morality of their decisions and actions by considering the consequences. It
is important for researchers to remember that the ultimate goal is to “produce the greatest good for the greatest number of people” (p. 459). To this end, the researcher sought to be mindful of the professional relationships of the participants. For instance, one of the principals stated that she selected teachers who were strong leaders, as was requested, but that she also selected teachers because she was “curious to hear what the participant had to say.” The researcher reminded her that all responses would be kept confidential in order to protect the participants.

Another type of ethics with which researchers need to be concerned is deontological ethics. In deontological ethics, “researchers judge the morality of their decisions and actions by referring to absolute values, such as honesty, justice, fairness, and respect for others” (Author(s), YEAR, p. 459). The researcher made a point to show respect for the participants and tried to be very respectful of the participants’ valuable time. The researcher was honest with the participants about her goals in the study and also their role in the study.

With regard to relational ethics, “researchers judge the morality of their decisions and actions by the standard of whether these decisions and actions reflect a caring attitude toward others” (p. 460). The researcher made sure to be kind and gracious as she entered the classrooms of the research participants. Furthermore, ecological ethics involve the morality of the researcher’s decisions and actions with regard to the larger cultural and social system of the participant. The interviews were conducted in the respective participants’ school settings, and the researcher made sure to be respectful of the environment and of the division’s willingness to take part in this research.

Assumptions
An assumption of this study was that a common language of assessment exists among the participants. For example, benchmark assessments can serve a variety of purposes, and educators need to be aware that this study is investigating benchmark assessments designed specifically for formative purposes. It was assumed that all interviewees would answer honestly and truthfully to the questions asked by the researcher. Another assumption was that the principals would select research participants based on the criteria provided, including that the teachers were perceived to be teacher leaders in their respective school and had worked in the division, under the current benchmark system, for at least five years. Furthermore, the researcher assumed that participants would understand the difference between “competent” and “confident” but found the need to clarify the distinction during the course of several interviews. The researcher also assumed that teachers would have anecdotes about the benchmark system working the way it in intended and documentation readily available, and this was not necessarily the case.

**Delimitations**

A delimitation is that the study will be conducted at the elementary school level. This is to control for any differences present among the practices of elementary and secondary programs. A constraint of this study is that the research was exclusively focused on the organization drivers and competency drivers of NIRN’s implementation framework. A future study could be conducted to evaluate the leadership drivers in place for a benchmark assessment system. Furthermore, by establishing the criteria for teacher participants who have worked in the school system for several years, the study is systematically eliminating new teachers. A future study could compare the competencies
and practices of veteran teachers and of new teachers within the same context, either school or division. Furthermore, it would be interesting to research the leadership drivers more closely. Interviewing principals with regard to benchmark implementation could produce interesting findings.

**Limitations**

Participants were limited by experience and the variables of their particular school, such as daily schedule. Another limitation of the study is self-reporting. Tuckman (1999) acknowledges that the self-report approach presents certain problems:

1. Respondents must cooperate to complete a questionnaire or interview.
2. They must tell what *is* rather than what they think ought to be or what they think the researcher would like to hear.
3. They must know what they feel and think in order to report it. In practice, these techniques measure not what people believe but what they *say* they believe, not what they like but what they *say* they like. (p. 237)

The researcher entrusted the participant to respond truthfully, and the researcher also assumed that the participant’s perception of his skills is accurate.

**Summary**

This study is an embedded case study that investigates the implementation of the division benchmark assessment system at three elementary schools. The elementary schools were selected based on their performance with regard to state accreditation over the last five years. This division was selected to participate due to the large size of the division, the varying degrees of success among its elementary schools, and its proximity to the researcher. The embedded case study design provides the methodological
framework for this study, because the three elementary schools are embedded within the division, and teacher competencies and practices are embedded within the larger context of the organization, including its division-level and school-level administrators. The benchmark system itself is the overarching implementation that provides the context for the study. The implementation framework, as developed by NIRN, is the key conceptual framework for this study; the researcher analyzed the data based on both competency drivers and on organization drivers. While the leadership component of the framework was touched on through the interview with the deputy superintendent of curriculum, instruction, and assessment, the researcher did not delve deeply into this component of the implementation framework. The data collected in this study included interviews and relevant documentation shared by teachers. Ultimately, this study was designed to evaluate the alignment between division-level administration and elementary school teachers’ perceptions of and practices with the benchmark assessment system order to aid other school systems in creating successful and effective benchmark systems.
CHAPTER FOUR

[Teachers] don’t always know what we’re supposed to do with the benchmark data. We might look at it closely and break it down for one child, but then if there’s only one question on one skill, do we really know if it was the skill or the question with which [the student struggled]?

The 4th grade teacher from School C highlights some aspects of the confusion surrounding benchmark assessments within this division. While many teachers in the division acknowledged benefits from administering benchmark tests, there were many recommendations made for how to improve the system. Teacher recommendations include the following: allow for teacher input in the creation of the benchmarks, allow teachers to view released benchmark items, and make an attempt to reduce instructional time lost. Based on the interview responses, the researcher also recommends procedural clarification with regard to intervention blocks, remediation expectations, and sharing student benchmark results.

**Findings**

The deputy superintendent of assessment describes the benchmark assessment system in the division in the following terms:

Our benchmark system we call our *critical skills assessment*, so CSA, critical skills assessments. Like many divisions we have a written curriculum, we have a nine week pacing guide, and curriculum leaders really are responsible working along with teacher specialists to draft those quarterly benchmarks or critical skills assessments. Those are secure documents in [our division]…. Really, for us, it was a way to deploy human resources in a strategic manner.
Currently, benchmarks are administered at the elementary school level grades 3-5. Students in Grade 3 complete quarterly benchmarks just in math and in English, while Grade 4 and Grade 5 students complete quarterly benchmarks in all four core subject areas. There have been some changes made to the benchmark system in the division within the last few years. For example, benchmarks are no longer administered at the 2nd grade level. The deputy superintendent of instruction shared the reasons for this change:

We used to give them in grade two; it was a financial reason why we stopped doing that last year, but was also an instructional reason because we saw a lot of our second grade teachers wasting time prepping for a quarterly skills assessment as opposed to using that time for the three R's [reading, writing, and arithmetic]. Just get them to read fluently and comprehend and by the time they leave second grade, the testing will take care of itself.

The deputy superintendent shared that 2nd grade teachers have actually been “kicking and screaming” about having their benchmark tests taken away. He notes that many teachers “relied on benchmarks as a quick way to assess student mastery.” The division has also changed their practice of sharing benchmark data with school level administrators, due to teacher perception of benchmark results as punitive:

As the standards changed at the state department, our central office realized that benchmarks results were viewed by teachers as punitive. They knew we were discussing scores by teacher name. Really, for us, it was a way to deploy human resources in a strategic manner…. Last year, to move away from the punitive mindset, we said, ‘Okay, we’re not going to publish the results. Principals aren’t
going to get the pass rates by teacher. We’re going to publish a student detail by question report so everyone in the division will see a spreadsheet for third grade math and each skill within that subject.’ You don’t see percent passed by teacher….What we realized was that, from an accountability standpoint, we had to go back to publishing this information. We went back to going through data by teacher at each school. In some cases, if you don’t do that, the performance can be masked by one teacher.

In this quotation, the deputy superintendent outlines major changes to data distribution that have occurred within the last several years. Initially, teacher data was shared with principals, then data was just shared by student and grade level/subject area, and now the practice has returned to the sharing of teacher data with principals. It would be interesting to uncover if a communication plan was in place during these major changes to the benchmark system. How aware are administrators and teachers of these drastic changes?

Based on the responses shared during teacher interviews, there appears to be alignment between teacher intent and division intent for benchmark assessments in the division; however, some of this alignment is due to the fact that teachers perceive the benchmarks as mandated by the division, so they are able to share the division’s goals clearly. While a cursory glance through the interview responses seems to suggest that teachers have a clear understanding of benchmark expectations, a deeper analysis reveals that there is a great deal of confusion among teachers with regard to a variety of components of the division’s benchmark assessment system. Teacher interviews
underscored the need for division clarification in several areas, including remediation expectations and the sharing of benchmark results.

**Research Question 1: Teacher Intent for Benchmark Assessments**

All nine teacher participants reported similar intentions with regard to administering benchmark assessments. All teachers shared that they give benchmarks to gauge where their students are with regard to retention and mastery of material covered thus far in the curriculum pacing guide. The 3rd grade teacher at School A shared that her intent in administering benchmarks is to utilize the results in order to see who has mastered specific skills, what needs to be retaught, and what needs to be reviewed during intervention. I have found that the [Technology Enhanced Items] are the ones that students miss the most, so we create TEI questions in our classroom assessments.

Similarly, the 3rd grade teacher at School C stated that her intent is to show growth…. It gives us a baseline of what we need to work on…. It’s a light bulb moment [that makes me think], “Oh, man, I thought that I had gone over that and [these results] are telling me that maybe I didn’t or maybe it was the way that the question was presented… and I need to retweak the way that I present it to [the students].”

The 5th grade teacher at School C shared similar insight with regard to his intent of benchmark testing: “Really to get results to identify problem areas…. It’s to get a feel for how well they master material, but I know there’s always something that’s going to give you a false reading on that.” The 5th grade teacher interviewed at School B shared that her intentions in administering benchmark assessments is to
find out exactly where my kids are. It helps me to differentiate my instruction for the students that have mastered it, then I can either push them a little further or we can move on to something else. The students that haven’t mastered the skill, we can go back and I know exactly what they need help on.

While many teachers shared similar intentions for administering benchmarks, such as mastery of the curriculum, several teachers felt that they administered benchmarks simply because the division mandates this practice.

When asked about their intent for benchmarks, three teachers shared that they give benchmarks because the division requires them to. For example, the 4th grade teacher at School B shared that the intent is “mainly because we are required to do them. The intent is also to see how much of what we’ve taught in the past nine weeks, and for the whole year, they have retained.” Similarly, the 4th grade teacher at School C shared slight reservations about benchmark assessments working as intended:

Well, we are told to use them as a guideline for how our students are performing. I don’t always agree that that’s the best use, because often the kids don’t feel like there’s any value for them, because they don’t get a grade. They don’t really know how they did…. We’re just told to use it as a gauge of how they’re doing in our class.

Furthermore, the 3rd grade teacher at School B explicitly responded, “We don’t have a choice” when asked about her intent in administering benchmark assessments. She elaborated,

It's a requirement. It is what it is. It's third through fifth grade and we have to give them. In third grade we only take Language Arts and math. And that just
started this year…. It’s kind of a lock down, shut down type of feeling, because…it prepares them for the SOLs. You’re told what to do and what not to do. I really don’t have any intentions, because I have to do what they tell me to do.

It is noteworthy that three teachers interviewed mentioned the division’s intention when specifically asked about their intentions as teachers. This participant, in particular, sheds light on the fact that many teachers see benchmarks as something they have to do. Clarification on the value of benchmark testing by division personnel could help assuage some of these sentiments. Furthermore, all nine participants shared that they had absolutely no input in the creation of the benchmark assessments. Five teacher participants argued that teacher input could be helpful, or that even releasing prior test items could be beneficial to improving their benchmark testing practices.

The 5th grade teacher from School B shared that, “I think it would be helpful if we had a little bit of input in [benchmark test creation], because they’re treated almost as strictly as the SOLs.” The 3rd grade teacher from School B emphasized that it would be helpful for teachers to at least see the benchmarks after they are administered:

We get to see the questions and the way they are formatted as we walk around the classroom while the students are taking the assessment, but we’re really not getting to view the full test. We have data meetings after each CSA and they give us a paper copy of it, but then we have to give it back. I don’t understand why we can’t have it….You can teach to the curriculum, but that’s not always going to be the way they format the questions. And I think that it’s an unfair disadvantage to teachers to not be able to know or have some blank questions that we can pull
from to know what they’re going to look like…. Give me a test from a few years ago! I’m not asking for the test you just gave [the students]. I want to know how they’re formatted, how they’re asked certain questions. I think it’s unfair to teachers [not to release this information].

Similarly, the 4th grade teacher from School C shared,

If we could look at the questions, we’d have a better idea of what students need.

The outcomes of benchmark testing should be that we use the data to guide our instruction, but it doesn’t always work out that way.

The 3rd grade teacher from School C responded with a similar viewpoint. She stated that it could be beneficial to have teacher involvement in benchmark test creation:

Even if it were a representative or a handful of representatives from each grade level, I think it couldn’t hurt. Then, it wouldn’t be just from our perspective wondering what people downtown are thinking or wondering what they expect. I think it would be a big help.

With regard to the creation of benchmark assessments, this teacher participant also shared that these assessments should be formatted more like the SOL tests. While the SOLs are administered through TestNav, the benchmarks are administered through PowerTeacher. She stated,

I wish they could have benchmarks formatted to look more like the SOLs. We don’t even use the same software. We’ve done the benchmarks on PowerTeacher, and then we are going to have to show then TestNav, which is the state [program]….We have to have a whole lesson on, “Forget that. That’s not what we’re doing for the SOLs.”
On the other hand, some teacher participants shared reasons to keep teachers from being involved in the benchmark creation process. The 4th grade teacher from School A shared, “I think their worry with showing people [the test] beforehand is people could give hints to her students and could over-prepare them for certain questions.”

The deputy superintendent for assessment shared some of his own goals for the future of benchmark test creation:

The future development, short term, I don’t see too many changes. We are going to continue to monitor school performance; we’re going to continue to use it as a vehicle to be strategic about deploying human resources and even financial resources. Then, hopefully, the results will be in time to remediate students at the end of the quarter, as opposed to waiting for summer school. The hope is that we can close gaps at the end of each quarter rather than waiting until the school year is over…. I think a way to continue to use results to help teachers grow is an area for future development.

The deputy superintendent also emphasized that the division would like to move away from the “over reliance on multiple-choice assessment” in order to review “other means for assessing student mastery.” The benchmark assessment system in the division could be strengthened through the inclusion of teacher representatives. In summary, the researcher found that teachers’ intents for using benchmark assessments are (1) to gauge student progress toward performance on the SOL test and (2) to be in compliance with the division’s requirement to use benchmark assessments in the classroom.

**Research Question 2: Teacher Practices for Benchmark Assessments**
All nine teacher participants shared extremely similar responses with regard to their practice of administering the benchmark assessments. The 4th grade teacher interviewed in School B best summarizes the practices shared by all teacher participants:

We get the schedule for the benchmark testing. We’re provided the passwords, because you now have to go into PowerTest. We have to go in the morning of each of the exams and “green light” the test to make sure that we create the test session. We have to create a separate test session if we have students are have read aloud services, except for reading…We have to create the test sessions, make sure [the students] have scrap paper, pencils, the whole nine yards. Everything that they need in order to take the test. Then, of course, they take everything on the computer. We have to make sure that they computers are in working order…Then we have to have everyone start, give directions…. Like on the reading test, we usually have to read directions about the different types of questions that they might see, like the TEI and hot spot items…. We go through that and give them the download password. Then, you have them go in and begin. We have to monitor, walk around, make sure that the students’ desks are far enough apart. You give them the little dividers so that everybody feels like their test and their screen is secure, and it’s just for them. When they are finished, we provide them with the submit password.

All nine teachers interviewed shared similar practices with regard to administering the benchmark assessment. However, differences exist between teachers’ review practices leading up to the benchmark assessment. Only two teacher participants shared detailed review practices. For example, the 5th grade teacher from School C shared,
What I definitely try to do is give [the students] something that outlines all of the skills that they’re responsible for. I can do more of that for social studies and for math. Language Arts is a lot different, a lot more difficult to do that. I try to make sure they have a good awareness of the skills that they’re going to be tested on. Prior to that, I always give them something that forces them to study and use those notes to answer questions. I try to go over that before testing. Even if it’s the morning of testing.

The 5th grade teacher participant from School B also highlighted her review practice leading up to the benchmark assessments:

> When it gets closer to the Critical Skills Assessment, I start making review packets and then I assign them for homework and… when they come in the next day, we’ll go over those questions, but instruction continues as if it were any other day.

Among the teachers interviewed, these two participants were the only ones who shared any sort of review practice leading up to the benchmark assessments. While no categories were established for this particular research question, there were major themes that emerged. Those themes are summarized under the subheadings of “remediation and instructional practices,” “sharing benchmark data,” and “grading.”

**Remediation and Instructional Practices**

While all nine teachers mentioned that they would utilize the benchmark data to gauge student progress, only one of the participants mentioned a specific remediation plan after receiving benchmark data. The 3rd grade teacher at School A stated, “After the test, the students track their scores in their data binders and make a goal for the next
CSA. We discuss what they need to do in order to be successful (show growth).” Out of the nine teachers interviewed, this participant was the only one to mention a remediation plan following the administration of the benchmarks. She was also the only participant to mention the intervention block described by the deputy superintendent of assessment.

While participants failed to mention the intervention block, some participants did share anecdotes about ways benchmarks have successfully impacted student achievement in their respective classrooms. The deputy superintendent of assessment shared that teachers are provided with cut scores for each benchmark test; however, there remains confusion, among some teachers, about what to do with test results. The 4th grade teacher representative from School C shared,

Sometimes we get a score, but we don’t even know what the cut score is. So a 65 might be passing, but we don’t really know. I think, for the kids, it’s a lot of confusion. They don’t really get any feedback, and if I do give them a score oftentimes I don’t even know what to tell them.

Some teachers argued that benchmark assessments should count as a grade for students. The 5th grade teacher from School C shared,

To me, the benchmark assessments should count as a grade, and it should count as a final grade that goes into that subject. The thing I hate is when I see a kid, and I’ve had a kid like this, who got an 86 on his math benchmark. His classroom grade was a C, but he had the third highest grade in the class. To me, that should count for something. Showing that he mastered those skills to the point that he got an 86…. I wish it did count. I think students might take it a little more seriously.
The deputy superintendent for assessment shared insights into the division’s grading practice for benchmark assessments. He stated that there is not a policy, but there is a practice:

At the elementary level, we do not [count benchmark assessments for a grade]. Sometimes what happens as we move into the secondary level is we see a lack of motivation because students know it’s not for a grade. Several years ago, we did begin counting them as grades, so less of a true formative assessment. The benchmarks are 5% of a student’s quarterly grade. Middle schools have the opportunity to curve if needed.

Several of the elementary teachers who were interviewed shared that this apathy towards benchmark testing is also occurring at the elementary level.

Sharing Benchmark Data

Abrams and McMillan (2013) note that teachers use assessment results in a variety of ways:

- identifying and addressing areas of student weakness, providing remediation for gaps in student learning, setting instructional priorities and increasing efficiency, determining instructional approaches such as whole class instruction, and differentiating instruction for small groups or customizing learning activities for individual students. (p. 110)

While all nine teacher participants mentioned that they utilize the benchmark data to gauge student progress, only one of the participants described a specific remediation plan. The 3rd grade teacher at School A stated, “After the test, the students track their scores in their data binders and make a goal for the next CSA. We discuss what they need to do
in order to be successful (show growth).” Out of the nine teachers interviewed, this participant was the only one to mention a remediation plan following the administration of the benchmarks, and even still the plan is not particularly detailed. She was also the only participant to mention the intervention block described by the deputy superintendent of assessment.

Research demonstrates that not all educators understand how to effectively use benchmark assessments and their resulting data in formative ways (Abrams & McMillan, 2013). It seems evident that several conditions are necessary in order for teachers to effectively use benchmark data to increase their students’ achievement (Abrams & McMillan, 2013; Goertz et al., 2009; Symonds, 2004; Wayman & Cho, 2009). It is noteworthy that, of nine teachers interviewed, only one discussed a clear plan for remediation when the purpose of formatively using benchmark testing is to “provide feedback to adjust ongoing teaching and learning to improve students’ achievement of intended instructional outcomes” (CCSSO, 2008, p. 3). The findings of this study replicate the findings of Abrams and McMillan (2013). These researchers discovered that schools are not adequately building the capacity in educators to use benchmark assessments in formative ways through professional development and support (Abrams & McMillan, 2013). Abrams and McMillan (2013) found that teachers who use assessment results are not necessarily making strong associations between students’ conceptual misunderstandings and an appropriate instructional response, at a pedagogical level. When the nine teacher participants were asked to describe their practice of administering benchmarks, all nine responded with the procedural practices of the day. For example, the participants shared that the students are not allowed to talk and that the testing
conditions are similar to those of an SOL test. This information does not necessarily imply that these teachers have an understanding of the formative use of benchmarks or the benchmark administration process, but rather that there are clear guidelines with regard to administering the division-created assessments.

Among the teachers interviewed, confusion exists regarding sharing benchmark test data with students and with parents. For example, the 4th grade teacher and 5th grade teacher representative responses from School C were in stark contrast. Both teachers have taught at School C for the last several years. The 5th grade teacher at this school outlined a clear process for sharing results with parents; in fact, he cites this practice as a major reason for his successful use of benchmark assessments. He stated,

I always send a little sheet home with the parents that show them the class average compared to their child’s average, so they know how well their kid did. It gives me a little more buy-in from parents to help prepare. They understand how serious the tests are so the kids are going to prepare for them because the parents are proud of their scores.

However, the 4th grade teacher from School C stated that teachers were explicitly told, several years ago, not to share benchmark results with parents:

A few years ago, we were told not to even tell [the students] their scores…. When we started doing this several years ago, I worked here and we would send home the scores with the students to share with their parents. We would explain to them what the test was about and how they could work with their students, and we’d give them some suggestions. We’re not supposed to do that anymore. Back then, we could put it on the report card. It wasn’t a grade, but at least parents could see
how their student performed. I think it worked better then; now it’s like a secret.

It feels like this secret thing, you’re not supposed to tell anybody their score.

When prompted to share when and how teachers were informed of this change in practice, the participant stated that teachers were told at a faculty meeting to no longer share these results. She stated,

Some teachers will say, ‘Oh, I’ve never heard that before.’ I’m like, ‘Well, we were all sitting in that meeting,’ but I guess it’s become more relaxed over the years. Nobody’s really paying attention to that anymore. I don’t know if we’re still under that rule or not.

This disconnect may have occurred when a change was made at the administrative level. Transparency regarding division level decisions, such as those quoted, could increase teacher understanding of benchmark assessment practices.

It would be beneficial for the division to share these expectations with teachers yearly, particularly as new teachers enter the division. This is evidenced by the fact that there is confusion even amongst teachers who have worked in the same school for several years. The 4th grade teacher from School A shared,

I think benchmark assessments can be very helpful, if you use them the right way.

If you’re putting too much pressure on the kids and it’s just about passing, then they can be really harmful. I think it’s about how you use it.

This division could benefit from clarifying the intention of this formative use of benchmark assessments among the division’s teachers.

Another discrepancy exists between the use of benchmark results between Schools B and C and School A. School A is not fully accredited, and the 3 teacher
participants shared that they use benchmark data for leadership meeting and for meetings with state representatives. The 5th grade teacher at School A shared,

> We are a school on warning, and we have a state person who comes in every week. She does walkthroughs and things like that and at the end of every nine weeks, the school leadership team, which I am a part of, has to do a presentation. The presentation includes the woman who works for the state. We provide all of our benchmark data to them and a big thing this year, because our school has made a lot of improvements, was comparing our [benchmark] scores at this time last year to now.

This emphasis on growth in School A was also evident in the interview responses from the school’s 4th grade teacher representative. She stated,

> I’ve always had an inclusion class, so I have a lot of kids who struggle and I really emphasize growth. Whenever we’re getting into testing, we focus on where were we at last time? And where do we want to be now? Also, I like [benchmark testing] because it breaks it down into specific skills and I can really see what we have and what we need more work on. So to me it’s more informative than you passed or failed. Are you working hard? Are you growing? Are you improving? And where can we improve more? I’m going to be happier if students go from a 20 to a 30 than if they dropped from a 90 to a 70. The student still passed, but he isn’t showing growth.

Teachers from School A seemed to have clearer understanding of the formative uses of benchmark testing. The 4th grade teacher participant from School A shared a moving
story about how benchmark results were successfully utilized to help increase
achievement in the case of one of her students:

I have one student in particular who is one of my Special Education students. He
started here last year from another school in [the division] and was really, really
low in reading. He actually got the Reading SOL test read aloud to him. He has
made so much progress that he no longer qualifies for that [accommodation]. He went from scoring in the 20s with read aloud to scoring a 60 while reading on his own. To me, it’s not necessarily the benchmark system; it’s everything that
we’ve done…. but [the benchmark assessment] helped us to see that growth. It’s
just kind of another measure. Like we do his reading level on SRI [the Scholastic
Reading Inventory] and that keeps going up, but it’s great to see him able to apply
that to the [benchmark] test…. It’s definitely been helpful to track his progress, so
I think it’s very helpful to be able to track the students’ growth and their progress.

This quotation demonstrates that there is not a shared understanding, among the teachers interviewed, with regard to the usefulness of benchmark assessments within this division. Another theme that emerged through the course of the interviews was the grading policy, or lack thereof at the elementary level, for benchmarks.

**Grading**

The deputy superintendent also referenced time when he discussed the grading practice of the division with regard to benchmark assessments. For the last five years, benchmarks have counted as an exam grade for students at the high school level, and they now, as of this school year, count as a grade for middle school students, as well. He
argued that the benchmarks are now a grade for the middle school for the following reasons:

What we did this year, and I mentioned about teachers having to give up the time so we went to teachers and said okay, let's try something different. How about let's take a high school approach and actually make it an exam? No cordially critical skills assessment, you're just going to have a semester exam, it's going be counted for a grade, 10%. Then oh, by the way, you're going take your SOL test in the spring prior to your second semester exam. Like high school, if you pass the SOL test then you don't have to take the exam. We're going see if that will work.

Interestingly, both the 4th grade and 5th grade teachers at School C argued that benchmark assessments should also count as a grade for elementary students for similar reasons. The 5th grade teacher participant shared, “I wish it did count. I think students might take it a little more seriously. Some of them that just don't test very well, it may affect them, but ... I don't know. I think it should count.”

**Research Question 3: Teacher Perceptions of Benchmark Assessments Outcomes**

Among the research participants, there were mixed responses with regard to the outcomes of benchmark assessments. While all nine teachers shared intended outcomes, the participants had more to share with regard to unintended outcomes. The 3rd grade teacher from School A best summarizes the outcomes, both intended and unintended, for benchmark testing:

Students are able to see their growth, and I can see what skills were mastered and which need to be revisited. With regard to unintended outcomes, the TEI
questions seem to stump the students. With an inclusion class…I do not feel that the division takes into account that special education students try but do not always test well.

Several participants shared similar intended outcomes to benchmark testing. The two categories that were established a priori were intended consequences and unintended consequences. These categories were included in the interview questions that aligned with research question three.

**Intended Consequences**

Several themes emerged within the category of intended consequences, and participants highlighted many positive intended outcomes. For example, the 4th grade teacher at School B stated,

I think [the benchmarks] provide us with some data about how much students are actually retaining… It does help us with even the ones that are getting good grades, so it says, “How much is getting into long term memory?”

The 5th grade teacher from School C also shared some significant outcomes from benchmark testing:

it allows us to clump together skills that we’re having problems with. It’s good when you can see the [the division] data so you can see if other schools, especially a similar school with the same kind of population, have the same problems that you’re having. Then, if you do higher than average, you feel a lot better like, “Hey, I actually did a good job with that particular skill.”

The 3rd grade teacher at School B informed the researcher that positive intended outcomes also include that the data
shows me as a teacher whether or not they comprehended what I was doing and what I was teaching them. To me, it’s a feeling of “oh yeah, I did that right’ or ‘no, I need to do something different with that.”

Teacher participants at School A also found positive intended outcomes associated with benchmark assessments in the division. The 5th grade representative shared,

The positives of [the CSAs] are that we get to see where we need to improve and then it also tells us what we’ve done well on. I do like that they give us comparatives across the division as well, not just what you did. It’s what did the rest of your grade level do and how do you compare to other schools within the division? Then, within that, you can look at schools whose population is similar to this school. How to we compare to them? I like that, as well.

The 4th grade teacher participant from School A also agreed “it is nice to see where we’re at and to see where we’re growing. Are we improving? What do we need to work on?”

The deputy superintendent of assessment shared several intended outcomes of benchmark assessments:

I know one helpful piece has been that we’re not moving on without taking into account where areas for growth exist for students. The other thing that’s been helpful is a change in teacher practice and growth, from a professional development standpoint. Teachers are having reach conversations at their benchmark meetings. There’s a lot of trust now in many of our schools because of these professional learning communities….Teachers are quick to share with one another where they failed and where they excelled and why. I think it’s
forced some dialogue that wouldn’t have happened if we had never gone down this road.

However, only two teachers mentioned the data meetings held after benchmark results were released, and none of the participants discussed collaborating with their colleagues with regard to remediation, instructional planning, or the use of benchmark assessments. While teachers shared many intended outcomes of the benchmark system in the division, there were also many unintended outcomes that concerned the research participants.

**Unintended Consequences**

Several themes emerged within the category of unintended consequences. When asked about the unintended and intended consequences of benchmark testing, the 3rd grade teacher at School B had interesting insights:

Negative consequences are the first thing that I think of, because everything shuts down, and [the students] are so young. These kids are eight and nine and to have them take such a high stakes assessment… They don’t understand that. They don’t understand it and it’s just too much.

The 4th grade teacher in School B also shared that, in addition to the stress for students, lost instructional time is another unintended consequence: “it makes us lose a lot of instructional time, because you have a whole week and half that is testing, testing, testing. Reviewing of what’s going to be on the test instead of having time to continue with instruction.” The 3rd grade teacher participant from the same school also shared that it causes the teachers, in addition to the students, more stress:

What I’m going through, the unintended outcome from my perspective is more
stress on me to try and figure out, “Okay, well that didn’t work. I tried small
group over here and pulling little Suzy Q over because she doesn’t know how to
use the dictionary and it may or may not be on the test…” but I still have to make
sure that everyone does well…. The unintended thing might even be that the kid
knows he didn’t do well. He knows he’s not going to go anywhere, so here he
goes again: “I’m going to have to go to some other kind of tutoring and be pulled
out of class to go try and learn and nobody’s helping me at home.” They have so
much stuff that they’re carrying around with them. There’s no preparation. They
can’t go home and study for the benchmarks. I don’t know what’s on there and
when I open it up, I go, “Oh. I didn’t talk to them much about that.” I think that
the unintended consequence is a whole lot more stress on [teachers].

Furthermore, the 5th grade teacher participant at School B stated,

Sometimes I don’t feel that it’s a true reflection of the kids I have sitting in my
room. I know that it’s to get them prepared for the SOL, but I don’t know. I wish
there were another way that they could prove what they know, other than just
answering the questions. Sometimes, I feel like the questions are designed to
trick them, so it’s like can we beat the test and show what we know? I think
that’s the unintended [outcome] that we get, but it is good data to get back and see
where they are.

While teacher participants at School A found the benchmark test results very useful
overall, they did share some unintended consequences. The main unintended
consequence that these teachers shared was lost instructional time; however, the 5th
grade teacher from School A shared concerns regarding the impact of benchmark testing on her students:

The negative is really just students that are not on grade level, and I know they are not going to pass the test. Another negative is they’re very time consuming. I mean sometimes students are sitting here for three or four hours taking one test. Ultimately, teacher participants shared several intended outcomes of benchmark testing, such as gauge of student mastery of content. However, there were myriad unintended consequences, including teacher stress, student stress, loss of instructional time, and concerns about whether or not the benchmarks actually demonstrate student learning. During his interview, the deputy superintendent of assessment for the division shared similar concerns to those of his teachers:

Every now and then I'll meet the teachers, and I met with eighth grade teachers last year in a focus group setting, science teachers, and they said, "We're gonna be honest, the pacing's so aggressive, and there's so much pressure being put on teachers from administrators and then more putting pressure on administrators that I know you all said this is supposed to be a formative test, but we know our names are getting put out there, and so we're shutting down instruction a week, a week and a half before to review, so if you think about that over the course of a year, even in aggressive cases where we have our curriculum finished by the end of the third quarter like grade five Social Studies to build in review, that's three weeks of instructions that was devoted to review." While we want the accountability and we want to pinpoint where we need to help, teachers are feeling the stress because principals are feeling the stress, and in some cases
principals are not hesitant to just throw this up in the faculty meeting at the end of the quarter and your data is what it is. Hopefully if you've got 33% of your class, you're in an inclusion setting that everyone knows you're teaching inclusion and that's why your results are lower, but last week we actually sent this file with the division file to all principals. We say be cautious how you share this information…. Where I think it’s been most harmful has been just the pressure, because we are in [a school division] with 55% [of our schools that are accredited] and the pressure has impacted some teaching and learning because we’ve moved, in some cases, to drill and kill for what was supposed to be a formative assessment.

While many participants viewed benchmark assessments as necessary, several participants emphasized the lost instructional time. The 4th grade teacher from School B shared that every quarter, as benchmark testing approaches, “You have to start the process of review. I think it just takes away instructional time. That’s not intended, I’m sure, but it happens.” The 5th grade teacher from School C echoed these concerns:

We always get cut short on our last unit of the quarter [due to benchmarks]. Then, we’re scrambling trying to rush to get the last piece in, just to take a benchmark that doesn’t count for a grade. Often, we don’t have our last test. We often miss our last test or we rush through it just to meet a requirement. This last go around, we did math as our second test after a snow break…. I missed part of fractions. What did they bomb in? They bombed in fractions. Now, they’re looking at data and wondering what to do with it.
The 5th grade teacher from School A shared similar concerns about the amount of time spent on benchmark assessments:

A negative is they’re very time consuming. I mean, sometimes students are sitting here for three or four hours for one subject in one day, especially in math…. The kids who are done early are sitting there for quite a long time.

Benchmark assessments have already been eliminated in 2nd grade in order to address some of these concerns, and the division has cut down on Science and Social Studies benchmarks in 3rd grade.

In addition to the unintended consequence of lost instructional time, the deputy superintendent of assessment shared an anecdote that demonstrates a specific unintended consequence of benchmark testing that they have experienced in the division:

While we want the accountability and we want to pinpoint where we need to help, teachers are feeling the stress because principals are feeling the stress, and in some cases principals are not hesitant to just throw this up in the faculty meeting at the end of the quarter and your data is what it is. Hopefully if you've got 33% of your class, you're in an inclusion setting that everyone knows you're teaching inclusion and that's why your results are lower, but last week we actually sent this file with the division file to all principals. We say be cautious how you share this information. One principal went through… and she left her school in there, she deleted all the other names but left all this here, right? Then she saved the file and then she sent it to her staff, but then she sent the wrong file. She set the original file. That one has since made it around the division, teachers forwarding it to each other, so everyone's score. Unintended consequence.
This anecdote illustrates the high-stakes nature of benchmark testing in the division. Throughout the course of this study, loss of instructional time was another theme that emerged as a major unintended consequence of benchmark testing.

**Instructional time.** In analyzing the transcripts of the responses from all ten participants in this study, the word “time” was mentioned 114 times. This clearly underscores the importance of time to both the teachers interviewed and to the deputy superintendent. Four teacher participants specifically shared concerns about lost instructional time. The 4th grade teacher from School B shared that every quarter, as benchmark testing approaches, “You have to start the process of review. I think it just takes away instructional time. That’s not intended, I’m sure, but it happens.” Similarly, the 5th grade teacher from School A shared similar concerns about the amount of time spent on benchmark assessments:

A negative is they’re very time consuming. I mean, sometimes students are sitting her for three or four hours for one subject in one day, especially in math….

The kids who are done early are sitting there for quite a long time.

The deputy superintendent of assessment acknowledged the concern of lost instructional time and noted ways that the division is attempting to reduce this consequence of benchmark testing. One method is the elimination of benchmarks in 2nd grade, as noted by the deputy superintendent. However, the deputy superintendent maintains that 2nd grade teachers have been very displeased by the change, because they have lost access to valuable student data as they move through the school year. He stated,
We used to give them in grade two, it was a financial reason why we stopped doing that last year, but was also an instructional reason because we saw a lot of our second grade teachers wasting time prepping for a quarterly skills assessment as opposed to using that time for the three R’s. Just get them to read fluently and comprehend by the time they leave second grade, the testing will take care of itself.

Why would this rationale not also apply to other elementary school grade levels? In fact, the four teachers who explicitly discussed the lost instructional time highlighted similar factors to losing time to prepare for the benchmark. In summary, while all nine teachers shared intended outcomes, the participants had more to share with regard to unintended outcomes.

**Research Question 4: Similarities and Differences between Division and Teacher Benchmark Intent**

The deputy superintendent shared the division’s intent for benchmark testing in the following terms:

Really the goal is to teach for mastery, so it's not like we've taught it and now we move on. There is so much content, but we tried to ... The other thing we let them know at the division level for many of our assessments, they can be cumulative in nature, so if we at the division level see significant weaknesses, we'll provide some additional resources to teachers and they'll know to expect some questions on that even though it's covered second quarter. The other thing we do with our curriculum, we either do it at the beginning of a quarter or at the beginning off the year, just like the state has the blueprint, the SOL blueprint, many of our
curriculum leaders give an SOL blueprint. You know on your benchmark assessment you're going to have x number of questions from this reporting category, so it's a way to let teachers know that there is emphasis on these select skills this quarter, because we're given one third of the test represents these two skills, or something like that.

Overall, many teachers felt that their intentions for benchmarks were aligned with these division intents. The 3rd grade teacher participant in School B summarizes this alignment:

Well, I think the division sets the SOLs, pacing, that we have to follow. I would say they would be aligned the same way, because I want my students to pass and be successful. The division wants them to pass and be successful. So I think that they are looking broader-picture than what I am. I'm just looking at these 24 kiddos that I have in here. But I would think they'd be the same.

Furthermore, the 4th grade teacher from School B argues that the intentions do align “in the fact that we want to make sure the kids understand the subjects and that they’re retaining it.” The 5th grade teacher from School B agrees, “Everyone has the same end goal. We want [the students] to be successful and master what they need to, so I think that’s definitely aligned.” Both the 3rd grade teacher participant and the 5th grade teacher participant from School C felt that the intentions were aligned. However, the 4th grade teacher from School C shared, “The tests align with our pacing guide. It’s a good assessment in as far as it aligns with what we should have been teaching… Do the goals align with my goals? I don’t know. They might align.” This quotation illustrates the recurring theme that the teachers in this division view benchmarks assessments as
division business. There is no teacher ownership of the benchmarks, as classroom teachers are not a part of the benchmark creation process in the division.

The deputy superintendent of assessment stated that the goal of benchmark assessments is the division is to “teach for mastery.” Many teacher participants have a similar response when asked about the intent of benchmark assessments. However, the 4th grade teacher interviewed at School C shared that there is some confusion among teachers with regard to how to utilize benchmark assessment data:

I guess as a teacher, and I know a lot of my coworkers feel the same way, is we don’t always quite know what we’re supposed to do with the data. We might look at it closely and break it down for one child, but then if there’s only one question on one particular skill, do we really know that there was a weakness with that skill or was it the question? If we could look at the question, we’d have a better idea.

The intended outcome should be that we use the data to guide our instruction, but it doesn’t always work out that way.

Another disconnect exists with regard to sharing student data with parents. For example, each teacher participant at School C shared different procedures for sharing student benchmark scores. The 4th grade teacher shared that teachers were not allowed to share student scores with parents. She stated,

When we started [giving benchmarks] years ago we sent home scores to the parents…. We would explain to them what the test was about and how they could work with their students, and we’d give them some suggestions. We’re not really supposed to do that anymore.
However, the 5th grade teacher at the same school shared that providing the scores to parents is a huge component of the benchmark process. This teacher stated, “I always send a little sheet home to the parents that shows the class average compared to their child’s average, so then they know how well their kid did.” There is a clear disconnect with regard to expectations for sharing student data with parents.

Furthermore, one participant, the 3rd grade teacher from School A, blatantly stated that her goals for benchmark assessments do not align with the division’s goals:

I do not feel that they align. My philosophy in my class is to show growth. I let my students know the pass score, but I tell them that if they show growth I am happy. They should not be stressing over a score when they are testing. I simply expect them to try their best, and if you try your best you should see growth.

While many participants shared the importance of growth, only one teacher claimed that her classroom goals, in this respect, did not align with the division’s goals.

With regard to remediation as a result of benchmark assessments, the assistant superintendent of assessment describes an intervention block that only one teacher participant mentioned:

One thing that is consistent [among elementary schools] is we have an intervention block built into the elementary schedule. That’s thirty minutes a day, so often from a re-teaching standpoint, once they’ve disaggregated the data, looking at skill deficits, small group instruction re-teaching and that's done during an intervention block. We have a number of after school programs that's done during that as well. At many of our schools what they'll do, particularly in math, is they'll begin to incorporate it into instruction for the next quarter. For example
let's say as a team, based on the math data, there were six skills, and we're going to weave those six skills, not only reteach when we can but weave them into our morning bell-ringers.

While one of the teachers mentioned this intervention block, no teachers discussed specific remediation plans during the course of their interviews. Benchmark assessments are intended to inform instruction moving forward; therefore, it appears that these teachers are missing a critical component of the benchmark assessment system (Abrams & McMillan, 2013).

A noteworthy theme that materialized during the course of the interviews was the “us” versus “them” language that emerged during the course of several interviews. For example, the 3rd grade teacher from School C, in reference to teacher input in the creation of benchmarks, stated, “Even if there were a representative or handful of representatives from each grade level… then it wouldn’t be, from our perspective, wondering what people downtown are thinking or wondering what they expect.” It was noteworthy that this participant referred to division leaders as the “people downtown.” This seems to suggest a disconnect between division-level administrators and classroom-level teachers. Further evidence of this disconnect emerged during the interview of the 3rd grade teacher at School B. When describing her intent in administering benchmark assessments, she stated,

It’s a requirement. It is what it is. It’s third through fifth grade and we have to give them….It’s kind of a lock-down, shut-down type of feeling, because it prepares them for the SOLs. You’re told what to do and what not to do. I really don’t have any intentions, because I have to do what they tell me to do.
Language is critical to the analysis of interview responses, and this response furthers the theme of the “us” versus “them” mentality that emerged in a couple of the interviews. This teacher, in particular, demonstrates a lack of ownership over benchmark assessment administration and, in fact, explicitly states that she lacks intent because she is doing what “they tell me to do.”

**Research Question 5: Teacher Benchmark Competencies**

All nine teacher participants stated that they felt extremely competent with regard to administering the benchmark assessments in their respective schools. When asked about their competency, however, all teachers referred only to the actual administration of the assessment itself rather than the process as a whole. Eight teachers referenced the fact that they had been administering the CSAs for the past several years, and the procedures have not changed. For example, the 4th grade teacher from School A shared,

*I mean, I’ve been doing it since I started here so it’s kind of like oh, another benchmark every quarter. It’s not complicated: They sit down at a computer and take the test and I try to make sure that they’re doing their best.*

Similarly, when asked how competent she felt in administering the benchmark, the 3rd grade teacher from School B stated, “Oh, 100%, because I’ve done it so many times.” Interestingly, only one of the participants discussed a remediation plan when asked to describe their benchmark process from beginning to end (including after the administration of the assessment). The 2008 CCSSO defines formative assessment as “a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students’ achievement of intended instructional outcomes” (p. 3). In this study, the participants appear to utilize benchmark assessments
in summative ways by using them to assess their instruction up to that point, but not necessarily making adjustments for future instruction. A major theme that emerged with regard to this research question is the lack of teacher buy-in and ownership of the benchmark assessment system.

Researchers have found that teacher interest or “buy-in” to the benchmark process is a critical component of successful implementation. Bancroft’s (2010) study revealed that teachers found benchmark testing to be an interruption in valuable class time: “If teachers have little buy-in for a reform, their resistance can overturn the reform’s best intentions” (p. 55). In this division, there is no teacher ownership of the benchmarks, because classroom teachers are not a part of the benchmark creation process. The 3rd grade teacher from School C shared the following insights:

Even if it were a representative or a handful of representatives from 5th grade, 4th grade, 3rd grade….I think it couldn't hurt. Then we, as teachers, wouldn’t be wondering, ‘What are the people downtown thinking? Or what do they expect?’ We wouldn't be able to say that if we had a team of ... Maybe we do, but I've never been aware of that. I think it would be a big help.

Researchers have found that high growth schools exhibit strong evidence-based decision-making practice where teachers used the division’s benchmark assessment to reflect on instructional practice, used the core curriculum to guide instruction, and received frequent and high quality professional development on reading and math instruction. (Wang et al., 2012, p. 517)
According to the participant responses in this study, this division could benefit from reviewing the professional development they have provided teachers with in regard to benchmark testing. Increased transparency of division intents for the benchmark system would also be helpful, because teacher participants lacked a clear understanding of how to effectively utilize benchmark results, with the exception of one individual.

**Research Question 6: State Accreditation and Benchmark Results**

The deputy superintendent for assessment shared that, due to the work of the executive director for research in the division, there is a strong alignment between benchmark results and school SOL performance. He shared,

> We meet every two weeks with the superintendent. Today’s meeting was reviewing quarterly data to see where we need to provide support with limited human resources and financial resources…We’ve seen a strong correlation [between benchmark cut scores and SOL results]. By reviewing performance on our critical skills assessment we can almost predict, particularly in English, school performance on an SOL assessment.

All nine teachers agreed that there was an alignment between the benchmark and SOL test performance, and many stated that the benchmarks are, in fact, more challenging than the SOL tests. Interestingly, major similarities and differences emerged at each elementary school. Each school was selected based on a five-year review of state accreditation measures.

**School A**

School A was selected for this study because the school has struggled to meet state accreditation standards over the last five years. This school is not currently
accredited, but School A did average 70% in math and also met the three-year average in history, with 80%, during the 2015-2016 school year. Interesting themes emerged at School A with regard to the alignment of state accreditation standards with the division benchmark system. The 5th grade teacher in School A shared that “the test itself mirrors the SOLs.” Furthermore, the 4th grade teacher from School A feels strongly that the benchmark system is aligned with the state accreditation system in the division:

I think the whole point of the benchmarks is to get them ready for SOLs, which obviously impacts accreditation. Usually, I feel like our benchmarks are actually a little bit harder than the SOLs. So that by the time we get to that, it's like, oh this is a little easier for them. But it really it depends on the subject. So I think it does help in getting them ready which obviously the more kids are ready for the test, the more kids are going to pass. When I started working here, we were a focus school. We're now are conditionally accredited in math and we're hoping for reading this year. I think one of the things we've tried to do is make sure there is an alignment between what we are teaching in the classroom and then the benchmarks and then the SOLs. I think that definitely has impacted our status as a school and getting accredited.

Interestingly, School A, as a school on accreditation warning, seemed to grasp the importance of benchmark testing more than the teachers from the other two schools. The 3rd grade teacher from School A shared, “I feel as though the accreditation of our school does align to the state status because we have a lot of low readers and reading is our focus for accreditation for this year.” This participant was the only teacher interviewed who
actually linked one of the skills tested in the benchmark assessments to an area of concern with regard to state accreditation.

**School B**

School B was selected for this study because the school has recently met state accreditation standards within the last five years. The teacher participants at this school felt that there was alignment between the benchmarks and state accreditation standards. The 3rd grade teacher participant from School B argued that the benchmark helps prepare students for the SOL test:

> I think it’s necessary to take [the benchmark tests]. I really do. Because you can’t have kids go all year and do no high-stakes testing and then come to the end of the school year and they have to take that ginormous SOL. So [the benchmarks] prepare them for it.

The 4th grade teacher agreed; she stated,

> I feel like it does [align]. We’ve been accredited the past two years running. Before that we were on… I don’t know exactly the level we were. We weren’t on as dire of a warning as some schools have been on, but I think that… I think it should be aligned differently for the accreditation of the schools.

The 5th grade teacher from School B shared,

> I think the benchmark is a little bit harder for [the students], but I think it’s planned out that way. I believe they make it more difficult for them to challenge us, I guess, to raise the bar on our students a little higher so we’re not getting to the SOL and just barely making what we need to make.
Overall, the teacher participants from School B agreed that the benchmark assessment system was aligned with state accreditation standards.

**School C**

School C was selected to participate in this study, because the school’s scores have increased notably from 2014. The school is now fully accredited and has seen major increases in SOL scores over the past several years. The 3rd grade teacher from School C shared,

Oh, year. I’m sure there has to be [alignment]. The benchmark system in my school and the state accreditation—we all have these qualifications that we have to meet. Hence, we have [schools in our division] that didn’t make [state accreditation] and they’re in a watch system. This particular school, I’m very grateful that we’re not there. We’ve been accredited.

The 4th grade teacher argued that it is difficult to be sure of the alignment, because “a quarterly test is only on that quarter” instead of the end of course coverage of the SOL tests. She also stated that the math benchmark assessment, in particular, is more challenging than the SOL test.

The 5th grade teacher in School C agrees that there is alignment between the benchmark tests and state accreditation standards. He stated, “I think they’re right in line, because no matter what you’re studying, you have to have some goal and end point where you want kids to show their mastery. It also forces you, as the teacher, to make sure you have strategies in place that help all kids try to achieve that goal. I think [the benchmarks] do a good job of matching the two of them up.”

**Comparison of Three Elementary Schools**
All nine participants felt that the benchmarks were aligned with state accreditation standards. Many voiced concerns about the format of the SOL tests but argued that the benchmarks definitely help prepare students for high stakes testing at the end of the school year. Notably, participants at Schools B and C were among those interviewed who felt students may take the benchmarks more seriously if they were given for a grade. School A, however, focused that the importance should be growth from benchmark to benchmark.
CHAPTER FIVE

This study of a division’s benchmark system was designed to examine whether or not elementary school teachers in this division are able to successfully utilize benchmark results; it was also designed to narrow the gap in literature by examining the variables necessary for successful benchmark implementation. Slavin, Cheung, Holmes, Madden, and Chamberlain (2013) argue that the purpose of benchmark assessments is to “find out early where problems exist so that changes can be made before it is too late” (p. 374).

Throughout the course of this study, it became evident that increased transparency between division goals could improve teacher understanding of the formative uses of benchmark systems. Participants from School A exhibited a clearer understanding of the formatives uses of benchmark assessments than their School B and School C colleagues. Benchmark practices, particularly with regard to grading and sharing scores, need to be reviewed and explicitly shared with teachers. It was noteworthy that even teachers within the same schools had widely varying practices. Furthermore, teachers recommended increased teacher involvement in the benchmark creation process and the researcher recommends yearly, or even quarterly, clarification of benchmark assessment cut scores and the division vision for the benchmark system.

Validity

Assessment validity is concerned with “the truthfulness or appropriateness of decisions resulting from assessments” (Gareis & Grant, 2015, p. 33). It is the “extent to which inferences drawn from assessment results are appropriate” (p. 34). Many school systems fail to evaluate both the validity and the reliability of assessment, particularly if they are commercially made or when students perform well on that respective assessment
(Brown & Coughlin, 2007; Oláh et al., 2010). This study brings to light implications for both the predictive and the consequential validity of benchmark assessments within this school division.

**Predictive Validity**

A specific facet of validity is predictive validity, which indicates how well a student will perform on a later assessment (Gareis & Grant, 2015). Benchmark assessments are administered with the goal of predicting later performance on a summative assessment, such as a state standardized test at the end of the school year. Predictive assessments are “designed to determine each student’s likelihood of meeting some criterion score on the end-of-year tests” (Perie et al., 2007, p. 5). Predictive benchmarks are most commonly used in school divisions that are trying to gauge where students will score on end-of-course standardized tests and how to respond to those predictions accordingly.

The benchmark assessment system in this division is clearly designed to predict student performance on end of the year SOL tests and to inform efforts to remediate students according to their benchmark performance. Therefore, it is extremely problematic that none of the participants described using the benchmarks in such a formative way. Only one participant mentioned reviewing data with students, and her practice was only to share scores with students and have them set goals for future benchmark performance. There is no indication that the students understood what those scores even meant. Furthermore, there is the issue of the intervention block that only one participant mentioned in passing. According to the deputy superintendent, this seems to be a pivotal component of the benchmark assessment system in this division. Due to the
fact that this is a thirty-minute period of instruction, it is alarming that it would not have come up more during the course of the teacher interviews.

**Consequential Validity**

Consequential validity has to do with “The appropriateness of the intended and unintended outcome that ensue from an assessment” (Gareis & Grant, 2015, p. 35). For instance, in the case of benchmark assessments, the actual use of results by teachers and administrators to make informed decisions and undertake effective instructional actions would be indicative of consequential validity. In evaluating the effectiveness of a benchmark assessment system, it is imperative to take into consideration issues of intended and unintended consequences. As with all initiatives and instructional programs, a division must evaluate whether the program is helping or if there is the potential that the program may actually be doing harm to its intended beneficiaries. Based on this study, there is evidence that benchmarks may, in fact, be causing some harm to students in this division. Teachers note myriad unintended consequences that include student stress and lost instructional time. Unless major reform occurs, there is enough evidence to suggest that, in the current form, benchmarks in this system are doing more harm to students than good.

**Context, Systems, and Coherence**

The structure of this case study played a pivotal role in the analysis of interview responses. If the researcher had just interviewed the deputy superintendent within the larger context of the school division, there would have been vastly different results than to compare that response with those of elementary school teachers at their respective schools. Some of the incoherence among participants with regard to the benchmark
system may be due to the large size of the school system. Regardless of the causes, incoherence exists among the three elementary schools within the context of the larger division, and incoherence is also evident even among teachers at the same elementary school.

Fullan and Quinn (2016) note that coherence, among members of a school division, involves “making sense, sticking together, and connecting” (p. 1). They argue, “Coherence pertains to people individually and especially collectively…. Coherence consists of the shared depth of understanding about the purpose and nature of the work” (p. 1). During the course of this study, the deputy superintendent of the division provided a clear goal and vision for the use of benchmark assessments; however, this goal and vision was not replicated at the classroom level. Based on the interview responses of nine elementary school teachers, it is evident that most teachers just “talk the walk” (Fullan & Quinn, 2016, p. 2). When asked about their intent, as classroom teachers, in administering the benchmark assessment, many teachers had the same canned responses that revolved around growth and/or mastery. However, participants did not delve deeper into their actual formative practices with regard to benchmark assessments. Only one participant mentioned any semblance of a remediation plan, and only one participant mentioned the intervention block. However, School A did seem to have more coherence than either School B or School C.

The three teacher participants from Schools A shared similar goals for benchmark assessments: to demonstrate student growth. Interestingly, this school was the only one selected that was not accredited at the state level. Perhaps this sense of urgency is a contributing factor to the coherence of the benchmark assessment system at this school.
Further evidence of this coherence was that all teachers discussed the use of data in their leadership meetings to demonstrate growth and to plan for improvements. While these teachers still did not present strong coherence or pedagogical decisions made as a result of this data, there were at least conversations about it at a leadership level. However, School C, in particular, demonstrated strong incoherence, particularly with regard to sharing student results.

The 5th grade teacher at School C argued that sharing student benchmark results with the respective parents was the major reason for, what he deemed to be, his success on the benchmark assessments. However, the 4th grade teacher at the same school explicitly stated that she does not share benchmark results with parents or with students. The wide array of data sharing practices within the same school underscore the incoherence of the benchmark at School C, but also the incoherence of the division, as a whole. Fullan and Quinn (2016) note that greater coherence can be achieved through “purposeful action and interaction, working on capacity, clarity, precision of practice, transparency, monitoring of progress, and continuous correction” (p. 2). Based on the interview responses, most of these strategies are lacking in this division with regard to benchmark administration. The only monitoring of progress occurs at the division level, where their deputy superintendent shares and discusses scores with building principals. Teachers did not mention receiving any professional development or training that could have developed their capacity to successfully implement the benchmark assessment system. Furthermore, transparency and precision of practice are major issues that need to be addressed by the leadership of this division. The expectations for benchmark
assessments at the elementary level need to be addressed yearly, or perhaps even quarterly, based on the incoherence of teacher responses.

**Conceptual Frameworks**

Reviewing the interview responses of participants in this study through the lens of the various conceptual frameworks brings to light some alarming inconsistencies in the benchmark assessment system in this school division. The alignment triangle of curriculum, instruction, and assessment, as presented by Gareis and Grant (2015), underscores the major issues in the lack of formative uses of benchmarks. This figure illustrates the importance of the reciprocal relationship among all these components of teaching and learning; however, interview responses from the nine teacher participants in this study depict a more linear alignment: curriculum → instruction → assessment. There was no evidence presented of teacher using assessment results to inform either their instruction or the curriculum.

Furthermore, implementation drivers were clearly at play in this division, both positively and negatively. While the researcher did not include leadership drivers in this study, the organizational drivers and competency drivers in this division played a pivotal role in ineffectiveness of the benchmark assessment system. The results of this study underscore the importance of the competency drivers, in particular. While organization drivers clearly played a pivotal role in the ineffective benchmark system within this school division, more research needs to be done in order to delve deeper into this driver.

Organization drivers are facilitative administrators (e.g., superintendents and principals). Their role within the implementation framework, in the case of benchmark assessments, is to establish a data system to guide the process of innovation, create a
hospitable environment for change, assess immediate outcomes of the change, and to create support for implementers (Implementation Drivers, 2016). The only organization driver who was interviewed in this study was the deputy superintendent of instruction and assessment. While it is evident that there is much work to be done at the organizational level with regard to benchmark assessments, interviews of additional central office staff and of school-level principals would present a richer understanding of both the organization and leadership drivers at play in the failing benchmark assessment system in this division. For the purpose of this study, the ineffectiveness of the competency drivers of the implementation framework was the most evident.

With regard to competency drivers, the NIRN notes the importance of training and of coaching. The NIRN argues that the four main roles of a coach include supervision, teaching while engaged in practice activities, assessment and feedback, and provision of emotional support (Implementation Drivers, 2016). According to the interview responses of the participants in this study, none of these coaching practices have occurred within this division. The teacher participants did not share any feedback, support, or supervision when they described the benchmark assessment system. In order to implement a successful benchmark system, it is imperative that the school division establishes strong coaching practices and the appropriate training necessary for teachers to feel competent in what they are being asked to do, such as administer benchmark assessments and analyze their results. While the vast majority of participants shared that they felt competent in their administration of the benchmarks, digging deeper into their understanding of formative assessments revealed that there is a strong incoherence in the classroom practices and those purported by central office personnel. Another major
component of competency drivers is training. Training includes intervention training, knowledge, and belief in usefulness. It is imperative that leaders share research findings that support the use of benchmark assessments with teachers and to train teachers in the practices of effective benchmark assessment use. Training was not mentioned whatsoever by either the deputy superintendent or the teacher participants in this study. The failure of this division’s benchmark assessment system may be due, in large part, to a failed understanding of the competency drivers necessary for the successful implementation of an instructional initiative.

**Recommendations**

If a school division plans to utilize benchmark assessments, there needs to be a clear vision of purpose. If schools are utilizing these assessments to predict future scores on the Virginia SOLs, for example, then the system put in place should clearly reflect this goal. Based on an interview with the deputy superintendent, the benchmark assessment system in this study is clearly intended to serve a predictive use. However, teacher responses illuminate the fact that this is not a clear goal shared at the classroom level by elementary school teachers. In order for the benchmark system to be effective, there needs to be enough justification for using benchmarks in order to mitigate some of the more serious unintended consequences, such as lost instructional time. Based on the findings of this study, the detrimental impacts of benchmark testing in this division far outweigh the benefits.

More transparency between division level leaders and classroom level teachers could increase teachers’ understanding of the formative uses of these assessments. The deputy superintendent of assessment, during the course of his interview, shared a clear
vision of the system and its uses, and, if clearly articulated to teachers, this could increase teacher buy-in. This division should also review the professional development opportunities they have provided to teachers with regard to successful benchmark implementation. Eight out of nine teachers seemed to lack a clear understanding of the formative uses of these assessments and did not discuss any remediation plan or practice in their classrooms. Furthermore, due to the myriad changes that have occurred in the benchmark system in the division within the last five years, division leaders need to set and review clear expectations on a yearly basis with both teachers and with administrators. The deputy superintendent shared that, within the last few years, major changes have occurred with regard to removing benchmarks from certain grade levels, counting benchmarks for grades at the secondary level, and changes in the manner in which data is shared with principals. Based on these findings, it is clear that confusion remains among veteran teachers at the elementary school level.

Participants from School A could serve as model users of benchmark assessment implementers for their peers interviewed at both School B and School C. These three participants all discussed the importance of growth, and the 3rd grade teacher at this school was the only participant who mentioned and discussed a remediation plan after the administration of these assessments. Formative use of benchmark assessments is not just best practice for schools in jeopardy of, or who have already lost, their state accreditation status. All students need to demonstrate growth in order to achieve at their highest level. Participants at Schools B and C seemed to lack a clear vision of the importance of benchmark assessments and also a clear understanding of what to do with the data they receive from these assessments. If trained properly in the formative uses, many of these
teachers may move away from sentiments like “I give the benchmarks because I have to” and they would perhaps view it as less of a waste of time. Based on the findings of this study, teachers in the division recommend that division leaders allow for teacher input in the creation of the benchmarks, allow teachers to view released benchmark items, and make an attempt to reduce instructional time lost. Based on the interview responses, the researcher also recommends procedural clarification with regard to intervention blocks, remediation expectations, and sharing student benchmark results.
Appendix A

An Embedded Case Study of the Implementation of a School Division’s Benchmark Assessment System

Dear Participant,

The following information is provided for you to decide whether you wish to participate in the present study. Please be aware that you are free to decide not to participate or to withdraw at any time without affecting your relationship with the researcher or with the College of William & Mary.

The purpose of this study is to gain information about the manner in which teachers utilize benchmark test results in the classroom. The research questions are as follows: (1) How do teachers describe their intent in making use of the benchmark assessment system? (2) How do teachers describe their practices of making formative use of the benchmark assessment system? (3) What do teachers describe as the perceived outcomes of the benchmark assessment system? (4) What are the similarities and differences between division intent and design of the benchmark assessment system and teacher intent and practices? (5) How competent do teachers feel themselves to be to make use of the benchmark assessment system to progress student learning? (5a) Why do teachers feel more or less competent? (5b) What do teachers feel would contribute to their competency? (6) How are teachers’ responses similar and/or different relative to the state accreditation status of their respective schools?

Data will be conducted through individual interviews and will be transcribed for purposes of data analysis. Please be assured that all responses will be kept confidential and individual participant names will not be associated with the research findings in any way. Only the researcher will know your identity as a participant. Please do not hesitate to ask any questions about the study either before, during or after participation.

There are no known risks and/or discomforts associated with participation in this study. Participation in this study is expected to yield the benefit of contributing to the body of research on effective benchmark assessments. I would be happy to share the completed findings with you upon completion of this study.

Please provide your consent at the bottom of this form with full knowledge of the nature and purpose of the study and its associated procedures. A copy of this consent form will be provided for your records.

Regards,
Chelsea I. Kulp, Principal Researcher

__________________________
Signature of Participant

__________________________
Date
Appendix B

Table 1

Interview Questions: Deputy Superintendent

Questions

1. Would you please explain the benchmark system in your division?

2. Would you please explain who is responsible for creating benchmark assessments?

3. Would you please explain who is responsible for administering benchmark assessments at the elementary school level?

4. Would you please explain the frequency at which benchmark assessments are administered in your division?

5. Would you please explain who is responsible for scoring the benchmark assessments and how they are scored?

6. Would you please explain how classroom teachers receive the benchmark assessment data?

7. Would you please explain how quickly teachers receive benchmark assessment results?

8. Would you please explain how teachers use benchmark data in their classrooms?

9. Do you feel that the benchmark assessment results are helpful to classroom teachers?

10. How do you know that the benchmark results are either helpful or not helpful?

11. Are there any unintended negative consequences of administering benchmark assessments?

12. To what degree are the division’s benchmark assessments valid and reliable?

13. What is the division’s vision for the future development and use of its BM assessment system?
Appendix C

Table 2

*Interview Questions: Elementary School Teachers*

**Questions**

1. As a teacher, what is your role in the creation of benchmark assessments in the school division?

2. As a teacher, what is your intent in giving/administering benchmarks?

3. What are your practices with regard to benchmark administration? Please describe the process from beginning to end.

4. What are the outcomes of the benchmark assessment system, both intended and unintended?

5. How competent do you feel in administering the benchmark system at your school?

6. To what degree do you feel that the division’s goals for benchmark assessments align with your goals as a classroom teacher?

7. To what degree do your benchmark results impact your yearly teacher evaluation?

8. To what degree is there a relationship between the benchmark system in your respective school and the school’s state accreditation status?

9. Please tell a story, from your own experience, that you believe illustrates the intent of the division benchmark system working the way it is intended.
Appendix D

Table 3

*Interview Coding Categories*

<table>
<thead>
<tr>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Benchmark testing policy</td>
</tr>
<tr>
<td>2. Receipt of test results</td>
</tr>
<tr>
<td>3. Expectations for teacher use (including division expectations)</td>
</tr>
<tr>
<td>4. Instructional uses of results</td>
</tr>
<tr>
<td>5. Supports for using test results</td>
</tr>
<tr>
<td>6. Obstacles/barriers to using test results</td>
</tr>
<tr>
<td>7. Usefulness of benchmark testing</td>
</tr>
<tr>
<td>8. Value of benchmark testing</td>
</tr>
<tr>
<td>9. Recommendations to improve the practice</td>
</tr>
</tbody>
</table>
### Appendix E

**Table 4**

*Research Question, Interview, and Framework Alignment*

<table>
<thead>
<tr>
<th>Research Question</th>
<th>Framework Alignment</th>
<th>Interview Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>How do teachers describe their intent in making use of the benchmark assessment system?</td>
<td>Competency Driver</td>
<td>Teacher Question 1</td>
</tr>
<tr>
<td>How do teachers describe their practices of making formative use of the benchmark assessment system?</td>
<td>Competency Driver</td>
<td>Teacher Question 2</td>
</tr>
<tr>
<td>How do teachers describe the perceived outcomes of the benchmark assessment system?</td>
<td>Competency Driver; Organization driver</td>
<td>Teacher Question 3</td>
</tr>
<tr>
<td>What are the similarities and differences between division intent and design of the benchmark assessment system and teacher intent and practices?</td>
<td>Organization driver</td>
<td>Admin Question 1-12; Teacher Question 5</td>
</tr>
<tr>
<td>How competent do teachers feel themselves to be to make use of the benchmark assessment system?</td>
<td>Competency driver</td>
<td>Teacher Question 4</td>
</tr>
<tr>
<td>What are the similarities and differences in how teachers respond to the above relative to the state accreditation status of their school?</td>
<td>Competency driver; Organization driver</td>
<td>Teacher Question 6</td>
</tr>
</tbody>
</table>
References


Wallace, K. (2015, April 24). Parents all over U.S. are 'opting out' of standardized student testing. *CNN.*


Vita
Chelsea Ireland Kulp

| Birthdate | June 16, 1988 |
| Birthplace | Virginia Beach, VA |
| Education | 2013 – 2017 | The College of William & Mary, Williamsburg, VA, Doctor of Education |
| | 2010 – 2011 | The College of William & Mary, Williamsburg, VA, Master of Arts in Education: Secondary English |
| | 2006 – 2010 | The College of William & Mary, Williamsburg, VA, Bachelor of Arts in English and American Studies |

| Professional Experience | 2016 – Present | Project-Based Learning Project Director, Poquoson City Public Schools, Poquoson, VA |
| | 2015 – Present | High School English Teacher, Poquoson City Public Schools, Poquoson, VA |
| | 2015 – 2013 | Middle School English Teacher, Chesterfield County Public Schools, Chesterfield, VA |
| | 2011 – 2012 | Middle School English Teacher, Williamsburg-James City County Public Schools, Williamsburg, VA |