2018

The Predictive Utility and Longitudinal Student Growth of NWEA MAP Interim Assessments in Two Pennsylvania Schools

David Christopher Finnerty
Lehigh University

THE PREDICTIVE UTILITY AND LONGITUDINAL STUDENT GROWTH OF NWEA MAP INTERIM ASSESSMENTS IN TWO PENNSYLVANIA MIDDLE SCHOOLS

by

David C. Finnerty

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Doctor of Education

in

Educational Leadership

Lehigh University

April 25, 2018

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Education.

_____
Date

_____
Craig Hochbein,
Dissertation Director
Assistant Professor of Education
Lehigh University

_____
Accepted Date

Committee Members:

_____
George P. White
Professor of Education
Lehigh University

_____
Louise Donohue
Professor of Practice
Lehigh University

_____
Bridget O'Connell
Superintendent
Palisades Area School District

ACKNOWLEDGEMENTS

I could not have completed this dissertation without the love and support of my family and friends and the guidance from the wonderful educators at Lehigh's College of Education.

First, I would like to thank my advisor, Craig Hochbein. From our first interactions, during which your passion for improving student outcomes was on full display, through the whole of this dissertation process, you made this work better. Thank you.

To my committee members, Louise Donahue, Bridgett O'Connell, and George White. Each of you inspired me with your personal story and professionalism. Picking my committee was the easiest part of this journey. Thank you.

To Stephen Kimball, an outstanding educator, colleague, and, for my part, friend. Thank you for serving as my editor-in-chief! Your thoughtful feedback challenged my thinking and improved my writing. I hope to repay the favor when you someday pursue the suffix.

To my friend and reader, Rachel Holler. You got me started on this journey! Thank you for being an outstanding professor, wonderful educator, and dear friend.

To my family, especially my parents, Jack and Mary Beth. We didn't have much but we always shared a love for each other and for learning and that somehow has always been enough. Also my siblings, especially my brothers John, Steve and Matt, and my sisters Nancy and Beth. Pursuing a doctorate is difficult enough without all the "life changes" I added to the mix. Thank you for your unwavering support.

To my children Owen and Olivia – the Wonderorfs! You amaze me in so many ways great and small. I am so proud of you! Thank you for understanding when I had to "dissertate."

Lastly, to my future wife, Gina DeBona. This doesn't happen without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Accountability pressures in NCLB and continued in ESSA combined with a perceived void in actionable data have led districts to implement NWEA MAP interim assessments. NWEA MAP interim assessments purport to predict performance on accountability exams and to inform instruction in advance of these exams with the ultimate goal of improving student achievement. Interim assessments such as the NWEA MAP interim assessments are typically administered multiple times per year and therefore consume a significant amount of instructional time.

This study analyzed the longitudinal student data of 405 student from two Pennsylvania middle schools, grades 6-8, that had implemented NWEA MAP interim assessments. Using a purely quantitative design, this study investigated whether NWEA MAP scores grew significantly and to what extent NWEA MAP interim assessments contributed to the predictive utility of existing student achievement data.

Using RM-ANOVA and descriptive statistics, this study found statistically significant growth of NWEA group means but overall mixed evidence of sustained growth. Using block-wise multiple regression, this study found that while each administration of the NWEA MAP made a statistically significant contribution to the overall predictive utility of the model, the contribution was of limited practical value. Furthermore, this study found that additional administrations of the NWEA MAP eliminated the significance of earlier administrations. Existing student achievement data, course grades and especially prior year PSSA 6 scores, persistently and powerfully predicted performance on PSSA 7.

CHAPTER I

Statement of the Problem

Since the No Child Left Behind Act (NCLB) introduced assessment-based accountability into public schools on a national scale, educational leaders in public schools have significantly increased the number of standardized tests students are required to take (Topol, Olson, Roeber, & Hennon, 2012). The increase in assessments has been so dramatic that stakeholders both inside and outside of education have voiced concerns that American public schools have become too focused on standardized tests (Bidwell, 2015; Layton, 2015; Lazarin, 2014). Critics also argued that these tests may have had a negative impact on student outcomes by reducing student engagement (Layton, 2015; Werner, 2011), overinvesting instructional time in test activities (Kerr & Lederman, 2015; Nelson, 2013; U.S. Department of Education [ED], 2015; White House, 2015; Zernike, 2015), narrowing the curriculum (Bidwell, 2015), and creating unnecessary stress (Harris, 2015; Lazarin, 2014; ED, 2015).

The standardized assessments mandated by NCLB, however, accounted for a small part of the overall assessment calendar. As shown in Table 1, federal mandates in NCLB and its reauthorization, the Every Student Succeeds Act ([ESSA], 2016), directed states to administer a total of 17 assessments: once annually in reading and mathematics in grades 3–8, and then once in high school, as well as once in science in grades 3–5,6–9, and 10–12.

Table 1

Federal Assessments Mandated in NCLB

	Grade Level									
Assessment	3	4	5	6	7	8	9	10	11	12
Mathematics	X	X	X	X	X	X	---------X---------			
Reading	X	X	X	X	X	X	---------X---------			
Science	-------X-------			----------X----------				-------X-------		

ED (2002).

These 17 assessments accounted for less than 0.4% of instructional time, below the 2% cap suggested by ESSA (2016). Furthermore, the number of state-mandated standardized assessments has remained constant since the implementation of NCLB in 2001. Similarly, nothing in NCLB suggested any changes in substance or frequency of administration of traditional, teacher-administered classroom assessments. The increase in testing has been in large part due to the wide-scale adoption of a relatively new class of assessments known as interim assessments. Interim assessments purport to predict performance on state-mandated standardized tests and inform instruction in advance of these tests, ultimately to improve student achievement (Goertz, Olah, & Riggan, 2009).

This study examined how the accountability measures in NCLB influenced public schools to incorporate data use to drive decision-making and how a perceived void in actionable data led public schools to implement interim assessment programs. Using student achievement data from two Pennsylvania middle schools, this study analyzed the utility of interim assessments to predict performance on state-mandated standardized assessments, and investigated the degree to which students demonstrated academic growth.

Accountability

NCLB's statement of purpose was "to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging state academic achievement standards and state academic assessments" (ED, 2002, p. 1440). NCLB enumerated 12 action items through which this statement of purpose could be accomplished. These 12 action items might be distilled into a single theme: holding schools accountable for improving the academic achievement of all students as measured by state assessments tied to rigorous academic standards.

With the passage of NCLB, federal mandates adapted accountability measures developed in the manufacturing industry for educational institutions, which were largely unaccustomed to accountability (Davis, 2007; Hamilton & Stecher, 2009). Because legislators had adapted NCLB from the business/industry sector, the accountability outcomes in NCLB were market-based; externally imposed sanctions or rewards were built in to provide incentives for schools to improve student achievement (Mintrop & Sunderman, 2009). NCLB mandated that schools and districts achieve 100% proficiency by 2014. Failure to achieve Adequate Yearly Progress (AYP) toward this 100% goal exposed a school or district to a series of increasingly invasive interventions.

As shown in Table 2, NCLB compelled states to label schools or districts failing to achieve AYP as "Needs Improvement" and facilitated student transfers out of these schools (ED, 2003). A school's failure to meet AYP criteria for two consecutive years granted students the right to transfer schools and be provided transportation at the expense of their school district. A third consecutive annual failure required schools to augment educational services with supplemental tutoring or other programs designed to improve student achievement. Continued failure of a school to achieve AYP led to more drastic consequences for administrators and educators, such as replacement of staff, curricular overhaul, or complete restructuring (ED, 2003). While the early targets for AYP were within reach for most districts, as late as 2006, more than 90% of districts were meeting AYP, pressure mounted as the AYP thresholds increased toward 100% by 2014 (Pennsylvania Department of Education [PDE], 2007). By 2012, a majority of Pennsylvania schools had failed to achieve AYP and nearly 10% were in the lowest category, "Corrective Action" (PDE, 2012).

Table 2

Adequate Yearly Progress (AYP) and Needs Improvement Status Levels

Category	Consequence
Achieved AYP	None
Warning	None
Making Progress	Offer school choice
School Improvement 1	Improvement plan; technical assistance
School Improvement 2	Supplementary educational services
Corrective Action 1	Changes in curriculum, leadership, professional development
Corrective Action 2	Reconstitution, chartering, privatization

Accountability Reporting

At roughly the same time NCLB imposed federal mandates onto states, PDE implemented new reporting requirements for public schools that further raised the stakes not only for administrators, but also teachers and students. Beginning with the 2000 Pennsylvania System of School Assessment (PSSA), PDE reported students' progress toward the Pennsylvania Academic Standards (PAS) by using performance level descriptors (Advanced, Proficient, Basic, and Below Basic) to categorize student performance. Revisions in Chapter 4 of the Pennsylvania Code also directed that school-level results, "be broadly disseminated to an array of audiences including students, parents, educators, citizens, and state policymakers, including the State Senate, the General Assembly, and the State Board" (Data Recognition Corporation [DRC], 2015, p. 1). To meet this Chapter 4 requirement, PDE published an annual School Report Card containing aggregated school performance data and disaggregated performance data for identified subgroups by ethnicity, economic disadvantage, and special education status (DRC, 2015).

As part of a broader reimagining of school accountability that included a waiver from NCLB, in 2013, Pennsylvania replaced AYP as the primary measure of a school's success with a School Performance Profile (SPP). Though the SPP metric broadened the AYP measures of

schools by adding growth metrics, the SPP metric still held schools accountable for proficiency on annual state standardized assessments. As shown in Table 3, the components of Pennsylvania's SPP were almost entirely based on student performance on state-mandated standardized assessments, with only 10% (Other Academic Indicators) of the SPP coming from other data sources.

Table 3

Components of Pennsylvania Middle School SPP

Source Data	Percentage
Indicators of Academic Achievement Percent Proficient or Advanced on PSSA Mathematics, ELA, and Science	40
Indicators of Closing the Achievement Gap—All Students Percent of Required Gap Closure Met	5
Indicators of Closing the Achievement Gap—Historically Underperforming Students Percent of Required Gap Closure Met	5
Indicators of Academic Growth/Pennsylvania Value-Added Assessment System Meeting Annual Academic Growth Expectations for Mathematics, ELA, and Science	40
Other Academic Indicators Promotion Rate and Attendance Rate	10
Extra Credit for Advanced Achievement Percent Advanced on PSSA Mathematics, ELA, and Science	Up to 7 points

PDE (2017).

In addition to the direct consequences for schools and individual educators, the accountability reporting required in NCLB exerted pressure on Pennsylvania secondary schools from stakeholders in the educational community (Schoen & Fusarelli, 2008). Beginning with the introduction of the School Report Card and continuing with the SPP, stakeholders in the educational community gained easy access to standardized assessment student achievement data. The School Report Card made available to the public PSSA data aggregated to the building and district level in each tested subject, and categorized each school and district as having achieved AYP or not based on the percentage of students with a proficient or advanced status. When the

SPP replaced the School Report Card, PDE replaced the AYP categorical designation with a single Building Academic Level Score based on the percentage of possible points achieved. The Building Level Academic Score on the SPP facilitated easy comparisons between schools.

The widespread availability of mandated annual assessment data focused more public scrutiny on student achievement generally, and on students in state-defined subgroups (those categorized as economically disadvantaged, as belonging to certain ethnic groups, or as having a learning disability, etc.) more specifically (Schoen & Fusarelli, 2008). Public scrutiny exerted pressure especially on lower-performing schools. Moore and Waltman (2006) noted negative publicity and decreased teacher morale as a result of pressure to increase test scores. Many teachers in schools identified as low-performing indicated that they planned on leaving their position within five years (Sunderman, Tracey, Kim, & Orfield, 2004). At least one state published a list categorizing teachers based partly on their students' scores on high-stakes tests (Hu, 2012). In addition, some schools implemented merit pay bonuses based on the results of these state summative assessments (Schoen & Fusarelli, 2008). These indirect pressures on teachers notwithstanding, NCLB largely aggregated accountability measures to school and district levels, Pennsylvania increased and focused the accountability measures on classroom teachers and principals with Act 82 of 2012 (PDE, 2013b).

Act 82 Educator Effectiveness

Compelled by Pennsylvania's application for Race to the Top (RTTT) federal funding, Act 82 augmented the traditional observation and practice evaluation model of educators to include student achievement data, as shown in Table 4 (Public School Code, 2012). In addition to representing Pennsylvania's primary criteria for a school's success, the SPP also represents the Building Level Data, or 15% of a classroom teacher's evaluation (PDE, 2013b). Furthermore,

for teachers in subjects for which accountability was assessed, growth, as measured by the Pennsylvania Value-Added Assessment System (PVAAS), represented an additional 15% of a teacher evaluation (PDE, 2013b). Consequently, the roughly eight hours that comprise a student's annual standardized testing represented 90% of a school's SPP and up to 30% of an individual teacher's evaluation (PDE, 2013b). Similarly, Act 82 redesigned the structure of principal evaluations in parallel to that of the teacher evaluations with the Teacher Specific Data being replaced by Correlation Data (Public School Code, 2012). Though districts had some flexibility in identifying Elective Data, half of building principals' evaluation was based upon some form of student achievement data.

Table 4

Teacher Effectiveness System in Act 82 of 2012

Category	Data Source	Percentage
Observation and Practice	Danielson framework	50
Building Level Data	SPP	15
Teacher Specific Data	PVAAS growth, three-year rolling average	15
Elective Data	Student learning objective (SLO)	20

In sum, the assessment-based accountability established in NCLB created significant pressure for administrators and teachers indirectly through reporting requirements and directly through Act 82. Similarly, they have created pressure for students by characterizing performance using categorical performance level descriptors. Twelve states further increased direct pressure on students by requiring proficiency on state exams before graduation (Gewertz, 2017). Like the manufacturing industry, in which NCLB accountability has its roots, educational organizations facing pressure have responded by employing data to predict outcomes and inform practice (Davis, 2007).

Data-Driven Decision Making in Education

Modeled after the quality improvement frameworks in the manufacturing sector, such as Total Quality Management, data-driven decision making (DDDM) refers to the systematic collection and analysis of data from a variety of sources to inform decisions (Marsh, Pane, & Hamilton, 2006). The drive toward accountability has led educators to become more interested consumers of data to predict performance on standardized assessment, inform pedagogical decisions, and ultimately, improve student outcomes (Love, 2004; Mandinach, Honey, & Light, 2006).

Data are widely available in education, yet the mere presence of data is meaningless without deliberate action through which it can be transformed to provide actionable value. Several researchers have offered theoretical frameworks for DDDM (Bernhardt, 2004; Mandinach et al., 2004; Love, 2004; McLeod, 2005; Marsh et al., 2006; Means, Padilla, & Gallagher, 2010), with each acknowledging the enormous volume of data available to educators and the necessity for identifying practical utility to transform data into meaningful improvement. Data must be actionable, that is, timely, varied in source and type, and possess valid, student-level detail to inform practice.

Educators have access to two general categories of student achievement data: standardized assessment data and classroom assessment data. Viewed through DDDM frameworks, data from state-mandated standardized assessments such as the PSSA have some utility to predict performance on future PSSA assessments, but little utility to inform instruction. PDE reports PSSA data to districts broadly using categorical performance level descriptors without the necessary student-level detail to inform instructional practice. Additionally, PSSA data have virtually no value for informing instructional practice during the same year, as they are

not available until after the school year has ended and students have progressed to the next grade level.

Classroom assessment data has limited utility in predicting performance on standardized assessments (Helwig, Anderson, & Tindal, 2002; Noble & Sawyer, 2004; Sawyer, 2007; Willingham, Pollack, & Lewis, 2002). Teacher classroom assessment practice varies from classroom to classroom and may not be aligned to standards (Parke & Lane, 2008). Several studies have shown moderate to high correlation between predictions based on classroom assessment and actual performance on standardized assessments (Hoge & Coladarci, 1989), especially those students performing at the higher (Demaray & Elliot, 1998) and lower performance levels (Gaines & Davis, 1990). However, for students performing near the threshold of proficiency, arguably the most important student group in categorical accountability measures, predictions based on classroom assessment were not as strongly correlated. Furthermore, Bowers (2010) noted that though grades were strong predictors of student outcomes such as graduation, they were less predictive of student mastery of standards.

Optimally, assessment data should be aligned to standards to predict student performance on standardized assessments, be available in time to inform instruction, and include sufficient student-level detail to inform instructional decisions. Standardized assessment data, though aligned to state standards, are not timely and do not provide the student-level detail. Classroom assessment varies from classroom to classroom, may not be well aligned to standards, and provides low to moderate predictive utility for threshold students. Educators trying to predict student performance on these standardized exams and inform instructional practice before these exams must, therefore, find other data.

Interim Assessment

Interim assessments that purport to be aligned to standards and provide timely, actionable data to predict and inform instruction have been developed to meet this need. Perie, Marion, and Gong (2009) defined interim assessments as:

> Assessments administered during instruction to evaluate students' knowledge and skills relative to a specific set of academic goals in order to inform policymaker or educator decisions at the classroom, school, or district level…the results of which must be reported in a manner allowing aggregation across students, occasion, or concepts. (p. 6)

Interim assessments vary in form, but are typically designed to be shorter than the state-mandated standardized assessments to which they claim to be aligned, and are more frequently administered, generally from three to five times per subject per school year (Perie, Marion, & Gong, 2009; Success For All, 2007).

The theory of action for interim assessments is that if educators have timely access to assessment data aligned with state-mandated standardized assessments to predict and inform instructional practice, these schools can use that data to improve student learning outcomes on the state-mandated standardized assessment. Districts that support interim assessment see these assessments as filling a void in actionable data. Interim assessments provide the student-level detail and timeliness missing with standardized assessment data as well as the alignment and validity often missing with classroom assessment. However, Goertz et al. (2009) noted that, "much of the rhetoric on interim assessments paints a rosy picture" (p. 1) and further suggested that the connection between interim assessments and improved student achievement warranted additional study.

Several studies have explored the link between interim assessment and student achievement with mixed results (Henderson, Petrosino, Guckenburg, & Hamilton, 2007, 2008; Slavin, Cheung, Holmes, Madden, & Chamberlain, 2013). Missing from these studies is an exploration of the dramatic increase in the number of interim assessments students are being required to take and whether these additional assessments improve student outcomes. Whereas state-mandated standardized assessments occur once annually, schools that use interim assessments typically assess in multiple subjects, multiple times per school year (Goertz et al., 2009). Thus, interim assessments represent a much larger percentage of a district's investment and impose a much higher cost in lost instructional time. For schools to make an informed decision on implementing interim assessments, it is important to investigate the utility of interim assessments to predict performance and to inform instruction.

Summary of Background

In sum, the accountability pressure exerted by NCLB and continued with ESSA, drove districts to seek data to predict performance on accountability assessments and inform instruction in advance of these assessments. Schools possess large amounts of student achievement data that fit generally in two categories: accountability assessment data and classroom assessment data. However, many districts perceived that neither of these data sources presented actionable data needed to effectively predict student achievement and inform instruction. Based upon research that suggests formative assessment practice positively affected student outcomes, districts and third-party companies designed interim assessments to function as shorter versions of accountability assessments. Interim assessments are administered far more frequently than accountability assessments and therefore consume significantly more instructional time. The increase in assessment time has raised concerns about lost instructional time and whether interim

assessments effectively predict performance and inform instruction. Furthermore, if interim assessments do predict performance and inform instruction, districts must assess whether repeated administrations of interim assessments significantly contributed to prediction of performance and informing instruction.

Purpose of the Study

This study will focus on two Pennsylvania middle schools, grades 6–8, as these grades are among the most frequently tested (Hart et al., 2015). Students at these schools took the Northwest Evaluation Association's (NWEA) Measures of Academic Performance (MAP) interim assessments and the PSSA in mathematics and reading/English Language Arts (ELA). During the course of their middle school career, students took 27 or more interim assessments in addition to the seven PSSA assessments. Interim assessments serve to predict performance on standardized tests and to inform instruction in advance of these tests. Several studies have explored the utility of interim assessments to predict performance on standardized assessments. However, research regarding the value of interim assessments to inform instruction and ultimately demonstrate growth in student outcomes, especially with regard to repeated administrations of an interim assessment are virtually non-existent.

The purpose of this study is twofold, to investigate the utility of NWEA MAP interim assessments (1) to predict performance on the PSSA and (2) to improve student outcomes through informed instruction in advance of the next PSSA. To investigate predictive value, this study will employ a multiple regression designed to test to what extent each administration of NWEA MAP contributed to the utility to predict actual performance on the corresponding year-end PSSA. If NWEA interim assessments provide predictive value, then this study would expect to find strong correlation between NWEA proficiency projections and actual performance on the

corresponding PSSA. To investigate the utility of interim assessments to inform instruction this study will analyze the variation in student performance on each successive NWEA MAP over time. Additionally, this study will analyze the actual growth by performance level descriptor. If interim assessments do provide formative value for improving student learning outcomes, then this study would expect to find improved student outcomes longitudinally as measured by the scaled scores and the percentage of students scoring proficient or better.

Research Questions

The following research questions guide this study:

1a: Do NWEA MAP mathematics interim assessment scores differ significantly over time?

1b: Do NWEA MAP reading interim assessment scores differ significantly over time?

2a: To what extent do repeated administrations of NWEA MAP mathematics assessments contribute to the overall utility to predict performance on the mathematics PSSA?

2b: To what extent do repeated administrations of NWEA MAP reading assessments contribute to the overall utility to predict performance on the Reading/ELA PSSA?

3: Do the changes in NWEA MAP scores over time and the predictive utility of NWEA MAP scores vary by subject?

Significance of the Study

MAP interim assessments purport to predict student proficiency on the PSSA assessments and inform instructional decisions ultimately resulting in improved student outcomes. NWEA and other proponents of interim assessment products base their support of interim assessments on the rich, though complicated, body of research on formative assessment. Based on these assertions and the perceived lack of actionable data from other sources, many

school districts have invested significant financial resources and time, both instructional and otherwise, to implement interim assessment programs. Many studies have documented the positive effects of formative assessment on student outcomes (Black & Wiliam, 1998a). However, the formative assessment activities studied by Black and others bear little resemblance to interim assessment practice. Because both financial resources and time are scarce, it is critical for districts to ensure that these investments are providing a significant return.

Additionally, stakeholder criticism of assessment practice in public schools has the potential for profound policy implications. The implication of charges of over-testing is that public schools need not administer as many tests to accomplish assessment goals. Moreover, if public schools are indeed over-testing, then this incurs critical opportunity costs in lost instructional time and misplaced resources, which results in negative effects on student learning. Stakeholder criticisms primarily target state-mandated annual standardized assessments such as the PSSA and Keystone Exams, which are used to satisfy the standards-based accountability metrics in NCLB. These assessments, however, account for only a small part of the overall investment in assessment (Lazarin, 2014), and interim assessments make up a much larger percentage of the overall assessment calendar.

While many of the shareholder claims of over-testing are anecdotal in nature and lack a clear basis in research, public pressure can have important policy implications. Former President Barack Obama, federal lawmakers on both sides of the political aisle, and former Secretary of Education Arne Duncan have been critical of the amount of instructional time lost to assessment, suggesting that a cap of 2% of instructional time be dedicated to assessment (Kerr & Lederman, 2015; Nelson, 2013). Mr. Obama further warned that over-testing leads to disengagement and reduced student achievement (Werner, 2011; Zernike, 2015). In a December 2015 press release,

Mr. Obama trumpeted a revision of education policy that "rejects the overuse of standardized tests" and provides states with increased flexibility "to audit and streamline their current assessment systems" (ED, 2015). Also in December 2015, Congress acted upon concerns about a perceived over-reliance on standardized testing and the amount of instructional time spent on tests when it reauthorized the NCLB, as ESSA (ED, 2015).

Paradoxically, policymakers, educators, and parents have both criticized and affirmed the practice of standardized testing in public schools. Although they agree that public schools over-invest instructional time in assessing students, thus negatively impacting student learning, these same stakeholders agree that not all tests are bad. Mr. Obama affirmed the importance of statewide annual assessments in grades 3 through 8 and again in high school (White House, 2015), echoing the assessment requirements of NCLB (ED, 2003). Similarly, former education secretaries from both major political parties support annual statewide assessment (Hefling, 2015; ED, 2003).

The criticism and support among shareholders suggests the need for a solution vaguely defined by Mr. Obama (2015) in an open letter to parents and teachers: "Let's make our testing smarter." (p. 1). It is critical that these solutions be informed by research not only on state-mandated standardized assessments, but also on the significantly more frequently administered interim assessments.

Delimitations

Though interim assessments purport to inform instruction, this study will not directly investigate informed instruction, that is, how classroom teachers use interim assessment data. Rather this study will investigate student growth which would be the desired outcome of informed instruction. The focus on student growth rather than informed instruction is both a

deliberate delimitation and a limitation. A focus on student growth allows for a quantitative research design using longitudinal student achievement data whereas an investigation of informed instruction would necessitate a qualitative element that would be difficult to employ in a longitudinal design. The value of interview or survey data regarding how a classroom teacher used a specific set of interim data three years ago would likely be of little value. However, by excluding an investigation of informed instruction, this study will not be able to inform educational leaders on potential best practices in how teachers employed interim assessment data to improve student outcomes.

Definition of Terms

Accountability Assessments – summative assessments designed to meet the federal mandates in NCLB and ESSA.

Adequate Yearly Progress (AYP) - categorical determination by the state of a public middle schools progress as measured by a school's proportion of students achieving proficient level in reading and mathematics as well as meeting criteria in attendance and sub-group student achievement.

DDDM – Data-driven Decision Making

DRC – Data Recognition Corporation, third-party vendor contracted PDE to create and score assessments

ED – United States Department of Education

ELA – English Language Arts

ESEA – *Elementary and Secondary Education Act of 1965*

ESSA – *Every Student Succeeds Act of 2015* reauthorization of ESEA

Formative Assessment - Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited. (Black & Wiliam, 2009, p. 9)

Interim Assessments – "Assessments administered during instruction to evaluate students' knowledge and skills relative to a specific set of academic goals in order to inform policymaker or educator decisions at the classroom, school, or district level…the results of which must be reported in a manner allowing aggregation across students, occasion, or concepts." (Perie, Marion, & Gong, 2009, p. 6).

MAP – Measures of academic progress, an interim assessment from NWEA

Middle School - Schools that contain grades 6 and 7 with no grade lower than grade 5 or higher than grade 9.

NCLB – No Child Left Behind Act of 2001 reauthorization of ESEA

NWEA – Northwest Evaluation Association

PAS – Pennsylvania Academic Standards for Reading, Writing, Speaking and Listening, and Mathematics (1999-2014)

PCS – Pennsylvania Core Standards (2015 – present)

PDE – Pennsylvania Department of Education

PSSA – Pennsylvania System of School Assessment

PVAAS – Pennsylvania Value-Added Assessment System

RTTT – Race to the Top

SPP – School Performance Profile

CHAPTER II

Review of Literature

This literature review begins with the evolution of the assessment-based accountability systems that drive interim assessments, specifically those that affect the study group, Pennsylvania middle schools. The review of literature continues with an analysis of existing data sources, accountability data and classroom assessments, through a DDDM framework. Lastly, this review critically analyzes the research on interim assessments.

Assessment-Based Accountability in Pennsylvania

Evolution of the Pennsylvania System of School Assessment

Pennsylvania has required secondary schools to administer statewide assessments for more than 45 years, although the design and purpose of these assessments has evolved dramatically (DRC, 2015; Pennsylvania Bulletin, 2010). The current form of assessment took shape with the 1992 introduction of the PSSA. Districts were required to administer the PSSA on a three-year cycle, with assessments in mathematics and reading in grades 5, 8, and 11, and an optional assessment in writing in grades 6 and 9 (DRC, 2015). In a 1994 revision to Chapter 5 of the Pennsylvania School Code, the State Board of Education established the PSSA as an annual assessment for all public schools with assessments in reading and mathematics in grades 5, 8, and 11 (DRC, 2015). Additionally, Chapter 5 eliminated the district option for assessing writing, and required districts to administer the PSSA writing assessment on a three-year cycle in grades 6 and 9 (DRC, 2015).

In 1999, the Pennsylvania State Board of Education revised Chapter 4 of the Pennsylvania School Code, repurposing the PSSA as a criterion-referenced, standards-based instrument aligned to the new Pennsylvania Academic Standards (PAS) for Reading, Writing,

Speaking and Listening, and Mathematics (DRC, 2015; Pennsylvania Bulletin, 2010). The 2001 passage of NCLB further compelled Pennsylvania to augment the existing PSSA with annual assessments in reading and mathematics in grades 3, 4, 6, and 7, and science assessments in grades 4, 8, and 11 (PDE, 2007). By 2007, the PSSA consisted of reading and mathematics in grades 3-8 and 11; science in grades 4, 8, and 11; and writing in grades 5, 8, and 11.

In 2013, during the period of this study, PDE completed the transition from the PAS, which had been in place since 1999, to the Pennsylvania Core Standards (PCS). The State Board of Education adopted the Common Core State Standards (CCSS) in 2010 and shortly afterward charged a group of educators with creating the PCS by adapting these CCSS to "reflect the organization and design of the PA Academic Standards" (PDE, 2013a). PDE (2013a) described the transition to PCS as a shift away from high school completion to college and career readiness, which emphasizes higher order thinking and increased academic rigor.

Predictably, the change in standards affected the structure of the standards-based PSSA. To ensure alignment with PCS, PDE replaced the PSSA reading and writing exams with a redesigned English Language Arts (ELA) assessment that incorporated elements of writing into each grade-level assessment (DRC, 2015). After two years of embedded and stand-alone field testing, the ELA assessment went into effect for the 2014-15 school year (DRC, 2015). During the transition to PCS, the mathematics PSSA consisted of content common to both sets of standards.

Keystone Exams

In 2008, Pennsylvania introduced plans to replace the 11th-grade PSSA with the Keystone Exams as a "comprehensive graduation competency program" (DRC, 2015, p. 8). The Keystone Exams were originally designed to include end-of-course (EOC) exams for 10 high-school-level

content areas (Biology, Literature, Algebra 1, Algebra 2, Geometry, English Composition, Civics and Government, Chemistry, U.S. History, and World History) that would comprise at least 33% of the student's grade for the class (DRC, 2015). After field testing in 2010, PDE administered the first wave of Keystone Exams, including Algebra 1, Biology, and Literature, in spring 2011. Following a one-year hiatus in 2012 during which no Keystone Exams were administered, PDE required public school districts to administer Keystone Exams annually in Algebra 1, Biology, and Literature (DRC, 2015). PDE field tested Algebra 2, English Composition, and Geometry in 2011, but as of 2017, these exams have not yet advanced past the initial field test. PDE has not developed the remaining Keystone Exams in Civics and Government, U.S. History, and World History. In sum, the requirement for Keystone Exams to comprise at least 33% of the course grade has not yet been implemented.

At present, NCLB and ESSA require states to administer 17 exams, once annually in reading and mathematics from grades 3–8 and once in high school, as well as one science exam in grades 3–12. In Pennsylvania, students must annually take the PSSA standardized exams in ELA and mathematics in grades 3–8, with an additional assessment in science in grades 4 and 8 (PDE, 2015b). Pennsylvania satisfies the high school accountability testing requirement of ESSA with EOC Keystone Exams in Algebra 1, Biology, and Literature. While Keystone Exams were designed as high school assessments, many middle school students take algebra and thus take the Keystone Exam for algebra while still in middle school.

Beginning with the graduating class of 2020, Pennsylvania will require public school students to demonstrate proficiency in Algebra 1, Biology, and Literature to earn a diploma (PDE, 2015a). This graduation requirement was originally mandated for the class of 2017, but state legislation delayed the implementation by three years. Students may demonstrate

proficiency with a score of proficient or advanced on the EOC Keystone Exams in each of these subject areas. As shown in Table 5, a Pennsylvania secondary school student must take 10 state-mandated standardized exams from grade 6 through high school.

Table 5

Secondary School PSSA and Keystone Exams Testing Time

Grade	Testing Time (min)			Total Testing Time (min)
	ELA	Math	Science	
Sixth Grade PSSA	249	148		397
Seventh Grade PSSA	249	148		397
Eighth Grade PSSA	249	148	112	509
Keystone EOC Exams	146	150	144	440
Total	893	594	256	1,743

PDE (2015ab).

A total of 10 standardized exams over seven years of school, accounting for a little more than 29 hours or approximately 0.4% of instructional time, seems unlikely to have generated a clarion call for less testing. However, state-mandated standardized assessments are not the only assessments that students must take. The high-stakes application of these state-mandated standardized assessments for accountability has driven districts to seek data that can be used to predict student performance and inform instruction.

DDDM and Existing Student Achievement Data

Driven by the accountability movement in education, DDDM in education refers to the often broadly defined practice of systematically collecting and analyzing data to inform instructional outcomes (Marsh, Pane, & Hamilton, 2006). The use of data in education has grown rapidly. Federal policy in NCLB and Race to The Top (RTTT) have been powerful drivers of data use in educational organizations (Coburn & Turner, 2012; Mandinach, et al., 2006; Marsh, et al., 2006; McCaffrey & Hamilton, 2007). Proponents of DDDM contend that student achievement data are critical to improved student outcomes (Faria, Heppen, Li, Stachel, Jones, Sawyer, Thomsen, Kutner, & Miser, 2012). Critics charge that DDDM proponents often

present an overly optimistic connection between data use and improved student outcomes heralding the transformative power of data to positively affect student outcomes despite a weak empirical connection (Coburn & Turner, 2012; Militello & Heffernan, 2009; Slavin, et al., 2013).

The research on DDDM in educational practice can be categorized in three ways: descriptive studies of the contextual supports that promote the systemic use of data, quantitative studies of data use related to student outcomes, and qualitative studies of how teacher use data (Coburn & Turner, 2012; Marsh, Pane, & Hamilton, 2006). Much of the research relating data use to student outcomes studies implementation of interim assessments and will be reviewed later in this paper. This section reviews theoretical frameworks for DDDM and considers the utility of existing data sources, accountability assessment data and classroom assessment data, to predict performance and inform instruction within the context of a DDDM framework.

Theoretical Frameworks for DDDM

Educators have access to an abundance of student achievement data especially since the implementation of assessment-based accountability (Hamilton, Halverson, Jackson, Mandinach, Supovitz, & Wayman, 2009). The mere presence of these data is meaningless. To engage in DDDM, educators must participate in an iterative process of deliberate interaction with these data to improve student outcomes. To facilitate this deliberate interaction with data, several researchers offered theoretical frameworks for DDDM (Ikemoto & Marsh, 2007; Mandinach et al., 2006; McLeod, 2005; Means et al., 2010). McLeod (2005) discussed DDDM in instructional practice as possessing five elements: good baseline data, measurable instructional goals, frequent formative assessment, professional learning communities, and focused instructional interventions. Means et al. (2010) defined a conceptual framework for DDDM as a continual

process with five components (plan, implement, assess, analyze data, and reflect), and identified six conditions and supports for successful DDDM in education:

1) State, district, and school data systems;
2) Leadership for educational improvement and the use of data;
3) Tools for generating actionable data;
4) Social structures and supported time for analyzing and interpreting data;
5) Professional development and technical support for data interpretation; and
6) Tools for acting on data. (p. 3)

Though these and other DDDM frameworks differed in their precise language, they each included collection of actionable data – that is, data that are timely, varied in source and type, and contain valid, student-level detail (Darling-Hammond & Adamson, 2010; Mandinach et al., 2006; McLeod, 2005; Means et al., 2010). Within the context of meeting districts' accountability data needs, that is, to predict performance and inform instruction, further explication of valid, student-level data is needed. To predict performance, valid data would possess power to predict performance on a future assessment. Often predictive power is accomplished through alignment with future assessment (PDE, 2016). To inform instruction, data would include sufficient student-level detail to provide task-oriented feedback to the learner (Black & Wiliam, 2009; Hattie & Temperley, 2007). Student achievement data fit generally in two categories: standardized assessment data from accountability testing such as the PSSA, and classroom assessment data, which include all forms of informal and formal assessment data generated within the normal conduct of instruction.

PSSA Assessment Design

State summative assessments such as the PSSA were "designed to improve instruction" (PDE, 2009, p. 10), in part by aligning curriculum to standards and informing instructional decisions at the district level and in the classroom. PDE constructed the PSSA to measure student achievement relative to specific grade-level standards, the PAS from 1999 through 2014, and the PCS since 2015. Prior to 2015, the PSSA consisted of six test sections, three each in math and reading, ordered in a single test booklet (PDE, 2014). Each of the three mathematics sections consisted of 24 multiple choice questions with one or two open-ended questions per section. The three reading sections each contained 16–24 multiple choice questions with a slight variation in number of questions over the years studied (2012–2014), and five open-ended questions. In 2014, constructed-response questions replaced all five open-ended reading questions and three of the four open-ended math questions.

Table 6

2015 Grades 6–8 PSSA Test Design

PSSA	Item	Number	Question Value	Total Value
ELA	Passage Multiple Choice	23	1	23
	Standalone Multiple Choice	18	1	18
	Evidence-Based Selected Response	3	2	6
	Evidence-Based Selected Response	3	3	9
	Text-Dependent Analysis	1	16	16
	Writing Prompt	1	12	12
	ELA Total	49		84
Math	Multiple Choice	60	1	60
	Open-ended	3	4	12
	Math Total	63		72

PDE (2014).

The structure of the PSSA changed significantly in 2015 when PDE completed the transition to PCS. Since 2015, the PSSA has been administered in seven sections, three each in mathematics and ELA reading with an additional ELA writing section. The structure of the redesigned PSSA is shown in Table 6. Despite the change in structure, PSSA score reporting remained consistent across the change in standards; PDE reported PSSA scores as scaled scores, and categorically using four performance level descriptors (Below Basic, Basic, Proficient, and Advanced). Additionally, PDE reported disaggregated scores by reporting categories—five categories in math and eight in ELA—noting points achieved, points available, and a categorical strength profile (High, Medium, or Low) for each reporting category (Rivera, 2015). Typically, raw scores and scaled scores have not been widely used in favor of the categorical performance level descriptors.

As shown in Table 7, PDE defined three scaled score cuts representing the lowest scaled score necessary for each performance level. To establish these cut scores, the Pennsylvania Board of Education employed a "modified bookmark method" protocol. The Board gathered a panel of educational experts to evaluate each PSSA assessment with items ordered from easiest to most difficult. In an iterative process, panelists placed a bookmark "at the point in the booklet that best represented each level (basic, proficient, and advanced)" (PDE, 2013b, p. 60). The change in standards necessitated a significant recalibration of cut scores, again using the modified bookmark method. Table 7 shows the cut scores for 2012 through 2014 before PDE transitioned to the PCS and the cut scores for 2015 and 2016 after the transition to PCS had been completed.

Table 7

2012–2014 PSSA Scaled Score Cuts

	Grade	Minimum	Below Basic	Basic Proficient	Proficient Advanced	Maximum
Reading	6	700	1121	1278	1456	2391
	7	700	1131	1279	1470	2319
	8	700	1146	1280	1473	2610
Mathematics	6	700	1174	1298	1476	2649
	7	700	1183	1298	1472	2561
	8	700	1171	1284	1446	2337

DRC (2012, 2013, 2014).

2015–16 PSSA Scaled Score Cuts

	Grade	Minimum	Below Basic	Basic Proficient	Proficient Advanced	Maximum
ELA	6	600	875	1000	1115	1699
	7	600	845	1000	1130	1652
	8	600	886	1000	1130	1636
Mathematics	6	600	897	1000	1105	1531
	7	600	904	1000	1109	1536
	8	600	906	1000	1108	1558

DRC (2015).

Clearly, the cut scores were significantly different under PAS as compared with PCS. This recalibration is evident in the distribution of students by performance level shown in Tables 8 and 9. The percentages of students who achieved proficiency, proficient or advanced, on the 2015 ELA assessment declined dramatically compared to the 2015 Reading assessment across grades 6-8. The decline was more pronounced in the advanced category especially in grade 8. Math performance experienced similar declines but with significant increases in the percentages of students performing in the lowest category not observed in the ELA scores. The percentage of students who scored below basic more than doubled from 2014 to 2015 in both grades 7 and 8. The changes in test design and cut scores limited educators' ability to make useful comparisons between the PAS and the PCS aligned PSSA scores.

Table 8

State PSSA Results in Reading and ELA

Grade	Assessment	Percentage of students scoring in each performance level				
		Below Basic	Basic	Proficient	Advanced	Proficient and Advanced
Grade 6	2012 Reading	14	17	31	37	68
	2013 Reading	15	21	28	37	65
	2014 Reading	18	18	27	37	64
	2015 ELA	10.2	29.5	39.2	21.1	60.3
	2016 ELA	8.6	29.8	38.9	22.7	61.7
Grade 7	2012 Reading	11	13	35	41	76
	2013 Reading	13	17	31	39	70
	2014 Reading	12	16	30	41	72
	2015 ELA	6.6	35.1	41.5	16.8	58.3
	2016 ELA	5.0	33.5	43.3	18.2	61.5
Grade 8	2012 Reading	9	11	24	55	79
	2013 Reading	12	11	22	55	77
	2014 Reading	11	9	25	54	79
	2015 ELA	11.1	31.3	43.3	14.3	57.6
	2016 ELA	11.3	30.4	40.9	17.5	58.4

PDE (2017b).

Table 9

State PSSA Results in Mathematics

Grade	Assessment	Percentage of students scoring in each performance level				
		Below Basic	Basic	Proficient	Advanced	Proficient and Advanced
Grade 6	2012 Math	9	14	27	50	77
	2013 Math	14	13	27	46	73
	2014 Math	15	14	23	48	71
	2015 Math	25.6	35.1	28.2	11.2	39.3
	2016 Math	30.1	28.8	24.2	16.9	41.1
Grade 7	2012 Math	10	11	25	55	80
	2013 Math	13	11	25	51	76
	2014 Math	13	12	24	52	75
	2015 Math	34.0	33.3	23.2	9.5	32.7
	2016 Math	34.9	28.1	23.7	13.3	37.0
Grade 8	2012 Math	12	12	25	51	76
	2013 Math	13	13	28	45	74
	2014 Math	16	11	22	51	73
	2015 Math	38.2	32.4	21.5	7.9	29.4
	2016 Math	40.2	28.6	20.8	10.5	31.2

PDE (2017b).

Utility of PSSA Data in DDDM Framework

The DDDM framework requires actionable data to predict performance on the next accountability assessment, that is, the next grade level PSSA, and to inform instructional practice in advance of that assessment. With regard to the utility of prior PSSA data to predict performance on the next PSSA, PDE provides evidence that prior year scores do have value in predicting future scores. PDE asserted that the "PSSA exams are aligned to the appropriate grade level standards that are sufficient for longitudinal modeling and prediction" (PDE, 2016a, p. 6). PDE annually calculates projections of future proficiency based upon past performance in its calculations of Pennsylvania Value-Added Assessment System (PVAAS) Student Projections (PDE, 2016b). PDE annually calculates PVAAS student projections for middle school students using an analysis of covariance which includes all prior PSSA test data (SAS Institute, 2016). The value-added modeling which uses past performance to predict performance on future assessments has been studied for more than 30 years and PVAAS has been validated by independent research (PDE, 2016a; SAS Institute, 2016). While the purpose of the PVAAS projection is to provide a projection for growth, it is reasonable to conclude based upon PVAAS that prior PSSA data do have utility to predict performance.

Regarding the utility of PSSA to provide actionable data to inform instruction, the DDDM framework suggests limited utility. First, state accountability assessments such as the PSSA are administered at the end of the year and data arrive too late to inform within-year instructional decisions (Henderson, et al., 2007, 2008; Herman & Baker, 2005; Marsh et al., 2006; Shanahan, Hyde, Mann, & Manrique, 2005; Wiliam, Kingsbury, & Wise, 2013). Schools

administer PSSA assessments in the spring but do not receive results until summer, when students have moved on to the next grade or school.

District- and building-level administrators found these state standardized summative assessment data useful for organizational decision-making, such as improvement plans, curriculum decisions, and professional development. However, school principals reported difficulties with the timeliness of state summative data, particularly as it pertained to informing educational practice in real-time. More than 95% of Pennsylvania school principals responded that state summative assessment data were available and more than 80% responded that these data were moderately to very useful (Marsh et al., 2006). Several researchers noted that both district-level administrators and building principals valued and used these data in curricular and program evaluation, yet there was little evidence that they had value at the classroom level (Black & Wiliam, 2009; Herman & Baker, 2005; Marsh et al., 2006). RAND's ISBA survey data of mathematics and science teachers supported this conclusion, reporting higher utility for state-mandated standardized assessment data in aligning curriculum than for individual instructional needs of students (Hamilton, Berends, & Stecher, 2005). Not only does the lack of immediacy in data availability eliminate the possibility of informing within year instructional outcomes, Wiliam (2013, p. 6) and colleagues noted that the delay raises a concern about the "shelf-life" of the data. Since students have moved on to the next grade level with a new set of standards, the lack of immediacy of the data may have decreased the value of student-level inferences educators can make from the prior-year data.

Secondly, researchers noted that state standardized test data lack the student-level detail to promote gains in student achievement (Guskey, 2007; Stiggins, 2005). Score reports from the PSSA provided primarily categorical performance level descriptors. While the PSSA score

report provides descriptions of the reporting categories, PDE data reports have not provided data regarding the specific questions or the nature of any misunderstanding. Furthermore, Wiliam et al. (2013) noted that the categorical performance level descriptors are too coarsely defined to provide actionable data to advance learning.

Finally, PSSA data are further limited by its singularity of type and source. Triangulation in type and source of data more successfully provides actionable value (Marsh et al., 2006). Guskey (2007), noting the high volatility of standardized assessments observed by Kane and colleagues (2002), argued that singularity in type and source called into question the validity of decisions based upon such data. Additionally, reliance on categorical data potentially calls into question both the reliability and the validity of state assessment data. Porter, Linn, and Trimble (2005) reviewed the NCLB design decisions of all 50 states and noted that incremental differences in categorizations yielded significantly different results. In other words, minor adjustments to placement of cut scores, minimum scores for each category, produced significantly different results. Aside from the technical justifications for limiting data to a single, high-stakes assessment, teachers need ongoing data to inform instruction, as no single assessment can measure the "full range and depth of learning" (Guskey, 2007, p. 24). State standardized assessment data such as the PSSA do not provide the actionable data source necessary to inform instruction.

Utility of Classroom Assessment Data in DDDM

Classroom assessment data are abundant. Classroom assessments include a broad array of assessments such as informal minute-to-minute formative assessments, and more formal summative assessments, such as paper-pencil unit tests, performance tasks, and other measures (Zhang & Burry-Stock, 2003). Due to the differences in these two classes of assessments,

formative and summative, this section will consider them separately with respect to their utility to predict performance and inform instruction.

Formative Assessment. Although interest in formative assessment as an avenue of school improvement has grown (Stiggins, 2005), the concept has not been well defined (Black & Wiliam, 1998a; Dunn & Mulvernon, 2009; Heritage, 2009; Perie, Marion, Gong, & Wurtzel, 2007) because researchers disagree on what characteristics must be present to constitute formative assessment (Black & William, 1998a; Boston, 2002; Goertz, Olah, & Riggan, 2009; Harlen & James, 1997; Nichol & Macfarlane-Dick, 2006; Sadler, 1998; Shepard, 2005; Stiggins, 2005; Volante & Fazio, 2007). Dunn and Mulvernon (2009, p. 2) noted not only a lack of "inter-individual" consensus in a "constitutive and operational" definition of formative assessment, but also a lack of "intra-individual" consensus, with individual researchers using different definitions in different studies. They emphasized how this inconsistency limits the formation of a meaningful body of research.

Scriven (1967) defined "formative evaluation" as the evaluation of ongoing educational programs. When Bloom (1969) and later researchers applied Scriven's concept to student learners rather than programs, the word "assessment" replaced "evaluation" (Dunn & Mulvernon, 2009; Shepard, 2005). This replacement of "evaluation" with "assessment" suggested an activity rather than a process and may have caused confusion for later researchers (Dunn & Mulvernon, 2009). The difference of opinion about assessment as an activity versus evaluation as a process underlies the problematic absence of a universally accepted definition.

In their seminal review of 250 research studies of formative assessment, Black and Wiliam (1998a) broadly defined formative assessment "as encompassing all those activities undertaken by teachers, and/or by their students, which provide information to be used as

feedback to modify the teaching and learning activities in which they are engaged" (p. 8). Their later definition better explained formative assessment as a process rather than an assessment:

> Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited. (Black & Wiliam, 2009, p. 9)

The purpose of formative assessment is to inform instructional practice to propel the learner forward by providing timely, task-oriented feedback that addresses the gap between the learners' observed state of understanding and the desired state (Bell & Cowie, 2001; Black & Wiliam, 2009; Hattie & Timperley, 2007;Heritage, 2007; Ruiz-Primo & Furtak, 2004; Sadler, 1998). This concept of feedback, and the activities conducted to reduce this gap, are central to the definition of formative assessment (Sadler, 1989, 1998; Black & Wiliam 1998a).

Proponents of formative assessment, an important element in DDDM, argue that classroom assessment data are critical to informing instructional practice and improving student outcomes (McLeod, 2005; Wiliam, 2007). Many studies have shown significant links between formative assessment and student achievement (Andersson & Palm, 2017; Andrade, Du, & Wang, 2008; Bonner, 2009; Herman, Osmundson, & Dai, 2011). Black and Wiliam's (1998b) meta-analysis of more than 20 studies showed that instructional practices that were changed to include formative assessment resulted in significant educational gains, with effect sizes ranging between 0.4 and 0.7. Although some researchers criticized some of the research methodologies used in the studies upon which Black and Wiliam based their conclusions including the effect sizes, researchers agreed that short-term formative assessment can inform instructional practice

(Bennett, 2011; Dunn & Mulvernon, 2009; Kingston & Nash, 2011) and deserved additional study (Briggs, Ruiz-Primo, Furtak, Shepard, & Yin, 2012; McMillan, Venable, & Varier, 2013).

Viewed through the framework for DDDM, formative assessment has little value as a predictor of student performance on accountability assessments. Formative assessment is a process by which data informs instructional practice to improve student outcomes. Formative assessment practice is meant to change student performance, therefore, it is reasonable to expect improvements in student performance would clearly degrade the utility of formative assessment to predict performance. Additionally, formative assessment data are not "aggregatable" (Brookhart, 2013, p. 175) which disallows the possibility for collective inference. With regard to informing instruction, researchers note broad agreement in the utility of formative assessment practice to inform instruction differing only in degree (Hattie & Timperley, 2007). Though formative assessment has shown positive effects on informing instructional practice and ultimately improving student outcomes, research on other primarily summative assessment practice in the classroom has proven problematic.

Summative Assessment. Classroom assessment practice varies among grade levels, subject areas, and individual teacher classrooms and this variation in practice undermines the value of inferences educators can make (Parke & Lane, 2008; Willingham, Pollack, & Lewis, 2002). Furthermore, variation across grade levels, subject areas, and classrooms compromises meaningful aggregation of classroom assessment data. Correspondingly, much of the research on classroom assessment focused on grades rather than other assessments.

Though classroom assessments and course grades are abundant and highly valued by educators (Supovitz & Klein, 2003), several studies suggested that they have historically been poor predictors of student knowledge relative to standardized test scores (Noble & Sawyer, 2004;

Sawyer, 2007; Willingham, Pollack, & Lewis, 2002). Bowers (2010) noted that only "about 25% of the variance in grades is attributable to assessing academic knowledge but that the other 75% of teacher-assigned grades appear to assess a student's ability to negotiate the social processes of school" (p. 2). In their study of teacher grading practice, McMillan, Myran, and Workman (2002) noted a "hodgepodge of factors" (p. 211) that include academic performance and other nonacademic elements from which grades were derived. Furthermore, McMillan et al. (2002) hypothesized that inconsistency in grading was suggestive of differences among teachers regarding relative importance of academic standards. Black and Wiliam (1998b) were similarly critical of assessment practice though they noted some positive predictive utility, "Teachers are often able to predict pupils' results on external tests because their own tests imitate them, but at the same time teachers know too little about their pupils' learning needs" (p. 142).

Some research has shown moderate utility for teachers' predictions of student performance on accountability assessments (Gaines & Davis, 1990; Hoge & Coldarci, 1989). Gaines and Davis (1990) conducted two studies of teachers' abilities, informed by classroom assessment, to predict students' achievement on standardized assessments. In the first study, 30 4^{th} grade teachers were asked to predict the achievement which students would achieve in the lowest and highest quartiles on the Iowa Test of Basic Skills (ITBS). The study group included 530 students, approximately 80% of whom were white, 16% were economically disadvantaged, and 20% of whom had been retained at least once. The second study included 84 teachers in grades 2, 4, and 6. Teachers were asked to predict performance on the ITBS by percentile range, 1^{st}-15^{th}, 16^{th}-35^{th}, 36^{th}-50^{th}, and above the 50^{th} percentile. Teachers correctly predicted performance about 60% of the time. Teacher predictions were most accurate at the lowest levels. Gaines and Davis suggested that non-academic roles including race and socioeconomic status

played a role. These findings were consistent with other research. Hoge and Coladarci (1989) reviewed 16 research studies on teacher-based judgements of achievement and found at least 70% of the time teachers correctly judged student achievement. The 16 studies reviewed employed a variety of designs yielding "judgement/criterion correlations ranging from 0.28 to 0.92" (p. 303). Demaray and Elliot (1998) noted a similar moderately high correlation (r=.70) with evidence of higher predictive accuracy with higher performing students.

The moderately high accuracy levels of teacher predictions should not be surprising especially at either end of the performance continuum. As Cronin and Kingsbury (2008) noted the majority of predictions are easy, "Teachers frequently comment that they can tell you within a few days of instruction which students in their class will be proficient" (p. 3). Students scoring near the proficiency cut scores, arguably the most important students in a categorical accountability system, were more difficult to predict.

In sum, the utility of the primary existing data sources available to educators, accountability assessment data and classroom assessment data, does not satisfy the DDDM framework to predict performance on accountability tests and inform instruction in advance of these test. Accountability assessment data is well aligned with state standards but does not provide actionable data to inform instruction. Classroom assessment data provides formative value to inform instruction but may not be well aligned with standards and does not provide the precision to predict performance on accountability assessments especially for students near the cut scores.

Interim Assessments

In response to assessment-based accountability, educators have become increasingly interested consumers of data, especially data that can be used to predict student performance and

inform instruction to improve student achievement in advance of state-mandated annual standardized assessments. To meet this demand for actionable data, private organizations have developed and marketed various interim assessment instruments that purport to "measure growth (and) project proficiency on high-stakes tests" (NWEA, 2015, p. 1). Some districts have developed locally-made interim assessments that often included released content from previous state standardized assessments. Interim assessments are aligned to state standards and designed to mirror the summative tests such as the PSSA, predict students' success on the PSSA, and provide diagnostic information to inform instruction (Success for All, 2010). The "exams are designed to be shorter, formative assessments that will predict success on the longer, summative assessments used by the state" (Success for All, 2007, p. 18). Interim assessments are medium-cycle in scope, duration, and frequency of administration, falling in between short-cycle formative assessments and longer term summative assessments.

Perie et al. (2009) made explicit the importance of purpose in interim assessment, further categorizing interim assessments as either primarily instructional, evaluative, or predictive. Interim assessments that serve primarily an instructional purpose provide educators data to inform instruction. They can also share commonalities with formative assessment, as data could be readily available to provide feedback to the learner. However, instructional interim assessment differs from formative assessment in that it is usually longer in cycle and certainly allows for aggregation. Interim assessments serving primarily an evaluative purpose can be used to inform curricular decisions and assess the effectiveness of a given program. Evaluative interim assessments are not generally used for interventions and are more aligned with longer term DDDM. Interim assessments designed for a predictive purpose are used to predict individual and collective performance on summative assessments, such as the PSSA. It is also

important to note that interim assessments may be designed for multiple purposes. Perie, Marion, Gong, and Wurtzel (2007) discouraged multiple-purpose interim assessment but leave open the possibility of successful implementation under the right conditions.

Research on Interim Assessment

Interim assessment products have been growing in popularity and, despite budgetary pressures, have been among the most active segments of test publishing (Olson, 2005; Marsh et al., 2006; Sawchuck, 2009). Stecher et al. (2008) found in their longitudinal study of California, Georgia, and Pennsylvania teachers, that districts were requiring administration of interim assessments at higher levels. In Pennsylvania middle schools, the number of teachers reporting a district-required interim assessment more than doubled over the study years, from 28% in 2004 to 60% in 2006 (Stecher et al., 2008). Interim assessments are typically given three or four times a year in reading, language usage, mathematics, and science, increasing the number of standardized tests by 30 or more (NWEA, 2015). The Council of Great City Schools studied the assessment frequency for more than 7,000,000 students across 54 districts and found that students took an average of 112 mandatory standardized exams from kindergarten through 12th grade, with the highest concentration of tests found in secondary school, especially in grade 8 (Hart et al., 2015).

Studies of the impact of interim assessments on within-year growth in student achievement have been inconclusive with some studies showing significant, positive gains (Slavin et al., 2013; Konstantopolis, Miller, & van der Ploeg, 2013) while others have shown no statistical difference in student achievement (Cordray, Pion, Brandt, & Molefe, 2013; Henderson, et al. 2007, 2008). Henderson et al. (2007, 2008) conducted a covariate-paired, quasi-experimental study to investigate the effects of interim assessment implementation. The study

identified 22 Massachusetts middle schools that employed internally-created interim assessments that aligned to state standards and provided quick access to student-level data relative to students' performance on these standards. Study schools were matched to a group of 44 schools that did not use interim assessments. Researchers used prior-year performance on state standardized tests to match schools with those in the treatment group. They found no significant statistical difference in student achievement in these schools as compared with the 44 covariate paired schools that did not employ interim assessment. Similarly, Cordray et al. in their experimental study of 32 schools found no significant overall growth in the reading achievement of grades 4 or 5 students as measured by both NCLB accountability exam scores and NWEA MAP scores

Slavin et al. (2013) studied the implementation of 4Sight interim assessments in 608 schools across 59 districts, spread over seven states and including Pennsylvania. 4Sight interim assessments were implemented quarterly across grades 3–8. First-year results showed small but significant gains in math, but not in reading. Effect sizes increased in years three and four, although changes to the study cohort—from 59 districts in year one to 20 districts by year four—limits the application of these data. Similarly, Konstantopoulos, et al. (2013) conducted a large-scale, experimental design using a stratified sample of 57 schools randomly selected from a population of 116 eligible volunteer schools in Indiana. Thirty-five schools received the treatment with a student sample of 19,167 students in mathematics and 19,173 in reading. The researchers found statistically significant positive effects for the treatment group in grades 5 and 6 in math and for grades 3 through 5 in reading as measured by the Indiana Statewide Testing for Educational Progress-Plus (ISTEP+). In other grades, Konstantopoulos et al. (2013) found consistently positive, though not significant, increases for the treatment group.

Utility of Interim Assessments in DDDM

Interim assessments claim to predict performance on associated accountability assessments with a higher accuracy than other available assessment data. Some research exists that suggested that interim assessments predict proficiency 80-90% accurately (Cronin & Kinsbury, 2008) though other researchers noted a lack of empirical evidence to support this claim (Babo, Tienken, & Gencarelli, 2014; Brown & Coughlin, 2007; Goertz, Olah, & Riggan, 2009).

With regard to the utility of interim assessment to inform instruction, researchers differed. Broad definitions of formative assessment suggested the possibility of formative utility for interim assessments (Artner, 2010; Black, Harrison, Lee, Marshall, & Wiliam, 2004; Chappuis, 2005; Dunn & Mulvernon, 2009), while others argued that interim assessments have little formative utility to classroom teachers (Perie et al., 2009; Shepard, 2005). Furthermore, many studies have documented the positive effects of formative assessment on student achievement (Black & Wiliam, 1998a), yet the activities described as formative in these studies differed significantly from the interim assessment. Nonetheless, manufacturers of interim assessments market their products not only as providing predictive value, but also as building off of these documented positive effects of formative assessment. This contrasts with the research suggesting that the value of these interim assessments as formative is largely non-existent (Geortz, Olah, & Riggan, 2009; Shepard, 2005).

Black and Wiliam (2009) defined a theoretical foundation for formative assessment "as consisting of five key strategies:

1) Clarifying and sharing learning intentions and criteria for success;

2) Engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding;
3) Providing feedback that moves learners forward;
4) Activating students as instructional resources for one another; and
5) Activating students as the owners of their own learning" (p. 8).

They argue that assessments that serve primarily a summative function may also provide formative utility. Within the context of this framework, interim assessments that provide timely, student-level data aligned to state standards would seem to have formative utility.

CHAPTER 3

Methods

Perie et al. (2009) characterized interim assessments using three primary purposes: instructional, those interim assessments concerned with within-year outcomes; evaluative, interim assessments concerned with across-year growth; and predictive interim assessments used to project student outcomes on NCLB accountability assessments. Using Perie et al.'s classification, this study employed a quantitative investigation of NWEA MAP interim assessments at two Pennsylvania middle schools. This chapter begins with an overview of the study including a detailed description of the two assessments, NWEA MAP interim assessments and PSSA. This chapter continues with a description of the study sites and data sets. Lastly, this chapter discusses the data analysis procedures used in this study.

Research Questions

The purpose of this study was to investigate the degree to which repeated administrations of NWEA MAP interim assessments, informed instruction as measured by student growth and predicted student performance on PSSA state accountability assessments. Interim assessments have gained wide-scale acceptance based upon their perceived comparative advantage relative to other existing data sources to inform instruction and predict student performance. Given the level of investment of instructional time, districts should evaluate whether interim assessments did in fact accomplish both objectives, to improve student outcomes through informed instruction and to predict performance. Furthermore, districts should evaluate to what extent additional administrations of the interim assessment contributed to the goals of these assessments. The following research questions guided this study:

1a: Do NWEA MAP mathematics interim assessment scores differ significantly over time?

1b: Do NWEA MAP reading interim assessment scores differ significantly over time?

2a: To what extent do repeated administrations of NWEA MAP mathematics assessments contribute to the overall utility to predict performance on the mathematics PSSA?

2b: To what extent do repeated administrations of NWEA MAP reading assessments contribute to the overall utility to predict performance on the reading PSSA?

3: Do the changes in NWEA MAP scores over time and the predictive utility of NWEA MAP scores vary by subject?

Assessments

The Northwest Evaluation Association Measures of Academic Progress

This study used two sets of assessments, the NWEA MAP mathematics and reading interim assessments and the PSSA mathematics and reading assessments, administered in grades 6, 7, and 8. The Northwest Evaluation Association (NWEA) is a not-for-profit educational services organization best known for their Measures of Academic Progress (MAP) interim assessment, which was taken by nearly 8,000,000 students annually (NWEA, 2015). NWEA (2010) identified a number of uses for its MAP instrument, including progress monitoring, informing instructional decisions, and "estimating the probability of a student receiving a proficient score on the state assessment" (p. 3). NWEA MAP interim assessments are computer-adaptive instruments that selected a question depending upon the response to the previous question; a correct answer generated a more difficult question, whereas an incorrect answer generated a less difficult question. Compared to conventional testing procedures, computer-

adaptive assessment allows for more accurate assessment of mastery using fewer questions (Weiss & Kingsbury, 1984).

NWEA MAP interim assessments drew from a pool of 34,000 items, which NWEA purported to ensure that "students experience zero item repetition on assessments taken within 14 months" (NWEA, p. 3). NWEA used a cross-grade structure to characterize student performance as on, above, or below grade. The cross-grade structure and computer-adaptive nature of the instrument supported "direct measurement of academic growth and change" (NWEA, 2016a, p. 3). NWEA published MAP assessments in three subject areas: reading, language-usage, and mathematics, as well as a separate MAP assessment for science. The language-usage and science assessments did not factor into projections of proficiency and were therefore not part of this study.

Students in this cohort took MAP interim assessments 3 times in each school year on a computer. NWEA named these assessments based upon the season in which they are taken, Fall, Winter, and Spring. NWEA documented the approximate number of weeks of instruction for each NWEA MAP assessment at 4 weeks, 20 weeks, and 32 weeks respectfully for Fall, Winter, and Spring. MAP tests were not timed; however, NWEA approximated that each assessment should take between 50 and 60 minutes (NWEA, 2013, 2015). When mathematics questions allowed a calculator to be used, a digital calculator appeared on the testing screen (NWEA, 2013). Typically, the mathematics assessment contained 52 questions and the reading assessment contained 42 questions (NWEA, 2016). NWEA scored MAP interim assessments using Rasch Units, which NWEA abbreviated as RIT (NWEA, 2016). NWEA defined RIT scales as an equal-interval scale that allowed for measurement over time regardless of grade level or age of student. NWEA RIT scales ranged from 100 to 300 (NWEA, 2013). Periodically,

NWEA conducted norming studies, most recently Thum and Hauser's 2015 RIT Scale norming study, which evaluated more than 100,000 students and nearly 500,000 test scores (NWEA, 2016a). The 2015 RIT scale norms were developed using nine data sets spanning Fall 2011 through Spring 2014 (NWEA, 2015).

NWEA reported scores for MAP administrations as a total RIT score and disaggregated by content goals, four for math (Operations and Algebraic Thinking, Real and Complex Number Systems, Geometry, and Statistics and Probability) and three for reading (Literature, Informational Text, and Foundational Skills and Vocabulary). Additionally, NWEA reported actual and projected growth measures for both year-over-year growth and growth from the prior NWEA MAP as well as projections for proficiency on the PSSA.

Pennsylvania System of School Assessment

In contrast to the NWEA MAP assessments, the PSSA was not a computer-adaptive assessment, but rather a paper-and-pencil assessment. Additionally, the structure of the PSSA reflected specific grade-level standards rather than the cross-grade structure of the NWEA MAP. Similar to the NWEA MAP, the PSSA reported scores using Rasch ability units, though differently than the NWEA scaling. The PSSA separated defined test sections by subject administered over a number of days. PSSA documentation did not specify time restrictions but estimated test sessions to last between 40 and 80 minutes per section. PSSA directions allowed districts to provide extended time for students who did not finish within the testing period. PSSA assessments included items for psychometric use and field test items that did not factor into the students' scores. The PSSA did not materially differ in construction across grade levels 6–8 in both math and reading/ELA relative to the number and type of questions (DRC, 2013, 2014, 2015).

During the timeframe of this study, PDE completed the transition from one set of standards, PAS, to a new set of standards, PCS. The change in standards coincided with a corresponding change in the PSSA and cut scores for the performance levels. The changes in cut scores reflected in the student performance level distributions presented potential problems for this study. To ameliorate the potential effects of the change in standards, this study analyzed data from the Class of 2018 cohort whose assessment data entirely preexisted the change in standards.

NWEA MAP Concordance with PSSA

Because NWEA advertised MAP interim assessments as predictors of proficiency for state standardized assessments such as the PSSA, it published concordance studies that showed the relationship between the two assessments. NWEA studied MAP and PSSA scores of students from 18 Pennsylvania schools. Table 10 showed a strong correlation between MAP and PSSA. To develop concordance data between MAP and PSSA scores, in both reading and mathematics, NWEA employed an equi-percentile statistical procedure. The equi-percentile equivalent MAP score, $e_y(x)$, was calculated using the percentile score on the PSSA, P(x), using the relation, $e_y(x) = G^{-1}[P(x)]$, where G^{-1} is the inverse of the percentile rank function for PSSA scores (NWEA, 2015).

Table 10

Correlation between MAP and PSSA

Subject	Grade	N	r
PSSA ELA/MAP Reading	6	846	0.78
	7	854	0.72
	8	821	0.75
PSSA Math/MAP Math	6	850	0.86
	7	854	0.87
	8	830	0.85

NWEA, Feb 2016a

To assess the predictive validity of the concordance, NWEA researchers calculated a consistency rate by adding the true negative and true positive values, shown in Table 11. Consistency rates ranged from 0.86–0.87 in reading and 0.84–0.86 in mathematics with Type 1 and Type 2 errors equally likely. Mathematics grade 8 consistency showed the widest discrepancy with false negatives, a Type 2 error, more frequently observed 0.10 compared to 0.06 false positive error rate (NWEA, 2016a).

Table 11

Consistency Rate for PSSA to MAP Concordance

		PSSA Score	
		Below PSSA cut	At or Above PSSA cut
MAP Score	Below MAP cut	True Negative	False Negative
	At or Above MAP cut	False Positive	True Positive

Note. From NWEA (2016a, p. 23).

Study Sites

This study analyzed data from the Class of 2018 cohort from a single district with two middle schools. I selected middle schools because of the high number of assessments middle school students take and the availability of annual accountability assessment data. I selected these middle schools because of their participation in NWEA MAP testing. I assigned fictitious names to the two middle schools and the district. I selected the Class of 2018 cohort because these students had completed middle school before PDE had implemented the change in standards from PAS to PCS. Therefore, the students in this study cohort had taken the standards-based PSSA in grades 6–8 of the same design and aligned to a single set of standards, the PAS.

Wonderorf School District (WSD) served approximately 5,300 students from a rural/residential area of approximately 72 square miles with a population of 30,000. The WSD student population was predominately white, non-Hispanic (85%), with Hispanic (6%)

comprising the next largest group (Table 12). District-wide, slightly less than 29% of the students were economically disadvantaged.

Table 12

Comparison of WSD Middle Schools

	West	East
School Descriptors		
Title 1 School	N	Y
Grades	6, 7, 8	6, 7, 8
Average Years of Educational Experience	20.7	18.7
Enrollment	408	826
Percent Enrollment by Ethnicity		
White (non-Hispanic)	89.0	83.7
Hispanic (any race)	3.2	7.8
Black or African American	3.7	1.5
Asian	2.4	4.5
Multi-Racial (Not Hispanic)	1.7	1.9
Pacific Islander (Not Hispanic)	0	0.6
Percent Enrollment by Student Group		
Economically Disadvantaged	21.1	31.1
English Language Learners	0.5	1.7
Special Education	19.1	15.4
Gifted	4.4	5.3

School Performance Profiles (2015).

Data Sets

The data used in this study consisted of longitudinal assessment data from a single cohort of middle school students, grades 6–8. In parts of the analysis, I focused on data from grade 7. For these areas of the analysis, I needed baseline data, prior year PSSA scores, and consistent course membership, i.e. students having taken the same course. I identified grade 7 data as the best choice for several reasons. First, grade 7 mathematical data demonstrated more course consistency than grade 8 data because typically, the majority of grade 7 students took the identical math course whereas 8th grade course selection diverged more significantly with students distributed among several courses.

Second, I selected grade 7 data rather than data from grade 6 because the prior year accountability assessment data, grade 5 PSSA, were based on instruction in an elementary setting with additional sources of variation. The elementary setting was suboptimal because students took the PSSA 5 at six different schools and in an elementary setting, teachers taught all core subject areas (mathematics, reading, social studies, and science). Furthermore, elementary schools typically operated without a consistently defined bell schedule, the absence of which introduced potential variation in the amount of instructional time dedicated to each subject at each school.

Student demographic information, existing student achievement data (PSSA and end of course grades), and NWEA MAP interim assessment data comprised the data file. Of the available student demographic data, I selected School Membership, IEP status, and Economic Disadvantage for inclusion in the study. I selected school membership to control for school factors such as school data systems, leadership, and other factors identified by Means (2010) and other DDDM researchers. I included IEP status and Economic Disadvantage because these characteristic student groups have been tracked and separately reported by PDE (2017).

Longitudinal PSSA and NWEA MAP interim assessment data from the Class of 2018 cohort formed the data set, as shown in Table 13. In total, three administrations of NWEA MAP interim assessments (Fall, Winter, and Spring) in grades 6–8 in both reading and mathematics, and the corresponding PSSA in reading and mathematics comprised the data set. To provide context, I referenced the NWEA MAP assessments by the season and grade in which they were taken, e.g. the Fall 2011 NWEA MAP was coded Fall 6. Similarly, I referenced the PSSA by grade level.

Table 13

Class of 2018 Cohort Assessments

	Grade 6	Grade 7	Grade 8
Class of 2018	MAP Fall 2011	MAP Fall 2012	MAP Fall 2013
	MAP Winter 2011	MAP Winter 2012	MAP Winter 2013
	MAP Spring 2012	MAP Spring 2013	MAP Spring 2014
	PSSA-6 2012	PSSA-7 2013	PSSA-8 2014

Course Grades

In one part of the analysis, I included end-of-course grades for the 7th grade mathematics and Reading and English Language Arts (RELA) in the data set. As shown in Table 14, students took one of 4 math courses. The majority of students took an on grade level mathematics course, Course 2 Math, roughly 40% of students were accelerated above grade level taking Algebra 1 or Pre-Algebra depending on their level of acceleration. A small number of students took Math Seventh, a below grade level course that served special education students.

Table 14

Course Distribution

Course	Number of Students	Percent of Cohort	Average Course Grade
Mathematics			
Algebra 1	72	21.43	3.38
Pre-Algebra	69	20.54	3.35
Course 2 Math	186	55.36	3.30
Math Seventh	9	2.68	3.08
Reading and English Language Arts (RELA)			
RELA 7	333	95.14	3.34
RELA 7 CYBER	4	1.14	3.55
RELA Seventh	13	3.71	3.17

More than 95% of students took the grade level RELA course with a small number of students having participated in the cyber version of the class. As with mathematics, a small number of students participated in a below grade level RELA course, RELA Seventh, designed to serve special education students. Course grades were continuous data based upon a four-point

scale (0.0-4.0). The variety of courses that students took, posed a possible limitation until average grade calculations revealed strong similarity. I included course membership in the original regression model but found it to not be a factor and therefore it was excluded from the final analysis.

Missing Data

The Class of 2018 cohort consisted of 405 students, though the specific data set varied in each step of the analysis due to missing data. For the specific analysis in each subject area, mathematics and reading, I constructed the data set using student records that had scores for each assessment. As shown in Table 15, I encountered three types of missing data in this study. Approximately 10% of the original cohort had multiple, consecutive missing assessment data

Table 15

Starting Cohort, Missing Data, and Final Cohort

		Missing Data			
Math	Starting Cohort	Transfers	Absences	No Grade	Final Cohort
RM ANOVA	405	45	35	N/A	325
Growth – Descriptive	405	45	35	N/A	325
Movement – Descriptive	405	28	12	N/A	365
Multiple Regression	405	41	14	14	336
Reading					
RM ANOVA	405	45	32	N/A	328
Growth – Descriptive	405	45	32	N/A	328
Movement – Descriptive	405	40	6	N/A	359
Multiple Regression	405	41	8	6	350

points, suggestive of a transfer in or out of the district. Other student records lacked either a single assessment record or multiple, non-consecutive assessment records suggesting school absence. A small number of student records lacked end-of-course grades. No explanation was available for the absence of these grades in student records. In total, fourteen student records or

3.5% of the total cohort lacked course grades for mathematics and 8 student records (1.9%) lacked grades for reading.

Data Analysis

This section reviewed each research question, identified the corresponding hypothesis, and detailed the specific analysis employed for answering each research question. Additionally, this section discussed the justification for the specific statistical analysis selected. The generalized purpose of this study was to investigate the degree to which NWEA MAP interim assessments informed instruction to promote improved student achievement and predicted performance on the PSSA accountability assessments in advance of these assessments. Henderson et al., (2007, 2008) noted the practical impossibility of isolating the variable of informed instruction. To answer the overarching question, to what extent did NWEA MAP interim assessments inform instruction to improve student achievement, this study instead analyzed the longitudinal student growth of the cohort.

Question 1 – Student Growth

To investigate within-year and across-year student growth over time, I employed a three-part analysis using both inferential and descriptive statistics. I considered mathematics and reading separately in Questions 1a and 1b respectively. For the inferential analysis, I used a repeated measures (RM) analysis of variance (ANOVA). RM-ANOVA is the best analysis for this research question as it tested the variance among means of a dependent variable over time. Each student in the cohort was exposed to a qualitative variable, in this case, instruction informed by prior and ongoing interim assessment data over time and their achievement was measured on nine occasions by NWEA MAP assessments. The dependent variables for the repeated measures ANOVA were the mathematics or reading RIT scores from the Fall, Winter,

and Spring NWEA MAP administrations taken by the Class of 2018 during their middle school years.

The null hypothesis for this analysis was that no significant difference existed among the mean RIT scores for each administration of the NWEA MAP. Stated symbolically, H_0: $\mu_{6Fall} = \mu_{6Winter} = \mu_{6Spring} = \mu_{7Fall} = \mu_{7Winter} = \mu_{7Spring} = \mu_{8Fall} = \mu_{8Winter} = \mu_{8Spring}$, where μ represented the overall means by grade level (6, 7, and 8) and test season (Fall, Winter, and Spring). The alternate hypothesis was that mean RIT scores would increase over time.

For inclusion in the RM-ANOVA data set, students needed to have taken each of the 9 NWEA MAP interim assessments, Fall, Winter, and Spring for grades 6, 7, and 8. Incomplete test results, those that did not have test scores for each administration of the NWEA MAP, were removed from the population. As shown in Table 15, the mathematics data set included 325 student records with scores for each of the nine administrations of the NWEA MAP interim assessment. I removed 80 student records that were missing scores; 45 of these had missed multiple, consecutive assessments suggestive of a transfer in or out of the school. The remaining 35 students missed a single test or more than one test but not consecutive assessments suggestive of school absence. Similarly, for the reading data set, the starting cohort of 405 students decreased by 45 transfers and 32 absences resulting in a final cohort of 328 student records.

This study used the Statistical Package for Social Science (SPSS) to complete the RM-ANOVA analysis. RM- ANOVA assumed normality and sphericty, homogeneity of variance, in the data set, and required continuous data (Field, 2009). To validate the assumption of normality, I inspected the data using histograms and identified missing data or outliers. The assumption of sphericity was analyzed in the ANOVA analysis using Mauchly's test of sphericity. NWEA MAP RIT scores were interval data and well suited for RM-ANOVA.

The output from the RM-ANOVA would indicate whether an overall significant difference existed in the mean RIT scores over time, but it will not assess growth from administration to administration. Additional analysis was needed to determine whether growth was significant relative to prior assessment data. The RM-ANOVA output was augmented to include Bonferroni post-hoc analysis and contrasts.

Table 16

Student Count and Group Mean by Performance Level Descriptor (PLD)

PLD	Fall 6		Winter 6		Spring 6		Fall 7		Winter 7		Spring 7		Fall 8		Winter 8		Spring 8	
4	n_{6f}	$\bar{x}_{6f}$	n_{6w}	$\bar{x}_{6w}$	n_{6s}	$\bar{x}_{6s}$	n_{7f}	$\bar{x}_{7f}$	n_{7w}	$\bar{x}_{7w}$	n_{7s}	$\bar{x}_{7s}$	n_{8f}	$\bar{x}_{8f}$	n_{8w}	$\bar{x}_{8w}$	n_{8s}	$\bar{x}_{8s}$
3	n_{6f}	$\bar{x}_{6f}$	n_{6w}	$\bar{x}_{6w}$	n_{6s}	$\bar{x}_{6s}$	n_{7f}	$\bar{x}_{7f}$	n_{7w}	$\bar{x}_{7w}$	n_{7s}	$\bar{x}_{7s}$	n_{8f}	$\bar{x}_{8f}$	n_{8w}	$\bar{x}_{8w}$	n_{8s}	$\bar{x}_{8s}$
2	n_{6f}	$\bar{x}_{6f}$	n_{6w}	$\bar{x}_{6w}$	n_{6s}	$\bar{x}_{6s}$	n_{7f}	$\bar{x}_{7f}$	n_{7w}	$\bar{x}_{7w}$	n_{7s}	$\bar{x}_{7s}$	n_{8f}	$\bar{x}_{8f}$	n_{8w}	$\bar{x}_{8w}$	n_{8s}	$\bar{x}_{8s}$
1	n_{6f}	$\bar{x}_{6f}$	n_{6w}	$\bar{x}_{6w}$	n_{6s}	$\bar{x}_{6s}$	n_{7f}	$\bar{x}_{7f}$	n_{7w}	$\bar{x}_{7w}$	n_{7s}	$\bar{x}_{7s}$	n_{8f}	$\bar{x}_{8f}$	n_{8w}	$\bar{x}_{8w}$	n_{8s}	$\bar{x}_{8s}$

Additionally, I conducted two descriptive analyses to determine the movement of students among performance level descriptor categories over time. Movement among performance levels is important when working with high-stakes, categorical data such as PSSA performance levels. I calculated group means by performance level descriptor for each NWEA MAP assessment (Fall, Winter, and Spring) taken by the Class of 2018 cohort. To identify movement between performance levels, I used tabular representation as shown in Table 16, where n_{ij} and $\bar{x}_{ij}$ represents the number of students, *n*, and group mean, $\bar{x}$, scoring at a given performance level descriptor for each year, *i*, and season, *j*. The data set for this analysis was identical to the data set from the RM-ANOVA, 325 student records for mathematics and 328 student records for reading.

Furthermore, to show student movement among categories from the fall administration to the spring administration, I employed a more detailed descriptive analysis. Defining group membership by the grade 7 fall NWEA MAP mathematics RIT scores, I tracked movement

among performance level descriptors from the fall administration through the spring administration for the grade 7 assessments of the 2018 cohort. The data set for this descriptive analysis of within-grade movement required that students had taken each of the NWEA MAP interim assessments in grade 7. As shown in Table 15, the mathematics data set was decreased by 28 transfers and 12 absences for a final cohort of 365 student records. The reading cohort had 40 transfers and 6 absences for a final cohort of 359 student records.

Question 2 – Predictive Utility

To answer the question, to what extent do repeated administrations of NWEA MAP mathematics interim assessments contribute to the overall utility to predict performance on the mathematics PSSA, I used a multiple regression. Multiple regression is the best analytical tool for this research question because it tests the significance of a linear combination of the independent variables to determine whether these variables are collectively predictive of the dependent variable. The null hypothesis would be that the repeated administrations of the NWEA MAP (Fall, Winter, and Spring) do not contribute to the predictive utility of the model based upon student demographic data and final course grades. A significant result from the multiple regression analysis would cause rejection of the null hypothesis. I hypothesized that each administration of the NWEA MAP significantly and individually contributed to the prediction of PSSA achievement. As with Question 1, this study considered mathematics and reading separately in Questions 2a and 2b, respectively.

For inclusion in the data set for the multiple regression, students needed to have taken the Fall, Winter, and Spring NWEA MAP assessment in grade 7 and the PSSA in grades 6 and 7. The mathematics data set began with 405 student records and was decreased by 41 transfers, 14 absences, and 14 missing course grades for a resultant data set of 336 student records. The

reading data set similarly started with 405 student records from which records for 41 transfers, 8 absences and 6 missing course grade were removed, resulting in 350 student records in the final cohort.

To assess the addition of each category of data and each individual NWEA MAP interim assessment, predictor variables were entered as blocks in the multiple regression. The predictor variables for the regression were categorized in blocks as student demographic data (school membership, economic disadvantage, and special education status), existing student achievement data (grade 6 PSSA scaled score and the teacher assigned final course grade in the grade 7 course), and successive NWEA MAP assessments as shown in Table 17.

Table 17

Block-wise Independent Variables for Multiple Regression

Block	Independent Variables
1	School Membership, Economic Disadvantage, Special Education Status
2	6th Grade PSSA, 7th Grade End of Course Grade
3	7th Grade Fall NWEA MAP RIT Score
4	7th Grade Winter NWEA MAP RIT Score
5	7th Grade Spring NWEA MAP RIT Score

The school demographic data were dichotomous data. The teacher assigned final course grades were interval data, expressed on a four-point scale carried out to the hundredths place. The dependent variable was the scaled score on the grade 7 PSSA. Seventh grade was selected because baseline grade 6 PSSA data were available and the majority of grade 7 students typically took the same course whereas in grade 8, advanced math students' course enrollment diverged such that no single course represented a majority.

The output from the multiple regression model included correlation and regression analysis to identify the degree to which each administration of the NWEA MAP assessments in grade 7 contributed to the prediction of the grade 7 PSSA. The relative values of the

coefficients, β, informed the directionality and relative power of each administration of the NWEA MAP to predict the PSSA 7. Additionally, the output included a measure of how much variation can be explained by the models, R^2, and the change, ΔR^2, in the variation explained by addition of each additional NWEA MAP administration.

Question 3 – Variation among Subjects

To answer question 3, do the changes in NWEA MAP scores and predictive utility of NWEA MAP scores vary by subject, I compared the statistical analysis of mathematics and reading from the first two research questions. To assess variation across subjects in student growth, I compared overall means for each NWEA MAP administration. Additionally, I compared the descriptive analysis by using the yearly growth as a percentage of NWEA MAP school growth norms (NWEA, 2015). To compare the movement among performance levels, I synthesized the categorical movement across grade 7. To analyze the variance between subjects relative to predictive utility, I compared the model summaries for the multiple regression. Furthermore, I compared the regression coefficients across all five models.

CHAPTER 4

Results

This study sought to analyze the utility of NWEA MAP interim assessments to improve student academic growth through informed instruction and to predict performance on the PSSA. This chapter presents the analysis, organized in three sections, with each section devoted to one research question. Following the presentation of the results, each section relates the results of the analysis to the research question.

Question One: Do NWEA MAP RIT Scores Differ Over Time?

The first question sought to measure the longitudinal student growth through informed instructional practice. The analysis employed a repeated measures analysis of variance (RM-ANOVA) and descriptive statistics. The null hypothesis for this analysis was that there would be no significant difference among the mean RIT scores for each administration of the NWEA MAP. Stated symbolically, H_0: $\mu_{6Fall} = \mu_{6Winter} = \mu_{6Spring} = \mu_{7Fall} = \mu_{7Winter} = \mu_{7Spring} = \mu_{8Fall} = \mu_{8Winter} = \mu_{8Spring}$ where μ represented the group means by grade level (6, 7, and 8) and test season (Fall, Winter, and Spring). The alternate hypothesis was that mean RIT scores would increase significantly over time. I conducted the analysis separately for mathematics, Question 1a, and reading, Question 1b.

Mathematics

To answer the question of whether NWEA MAP interim assessment scores differed significantly over time, this study analyzed NWEA MAP RIT scores using RM-ANOVA. RIT scores from 325 students who had each taken all three NWEA MAP interim assessments, Fall, Spring, and Winter in mathematics for each grade 6, 7, and 8 comprised the data set for the RM-ANOVA. All nine NWEA MAP mathematics RIT scores were entered into SPSS as within-

subjects factors and analyzed using RM-ANOVA. As shown in Figure 1, overall test administration mean scores tended to increase throughout each school year and declined from spring to fall administrations.

Figure 1

Overall Mean RIT Scores for NWEA MAP Mathematics by Administration

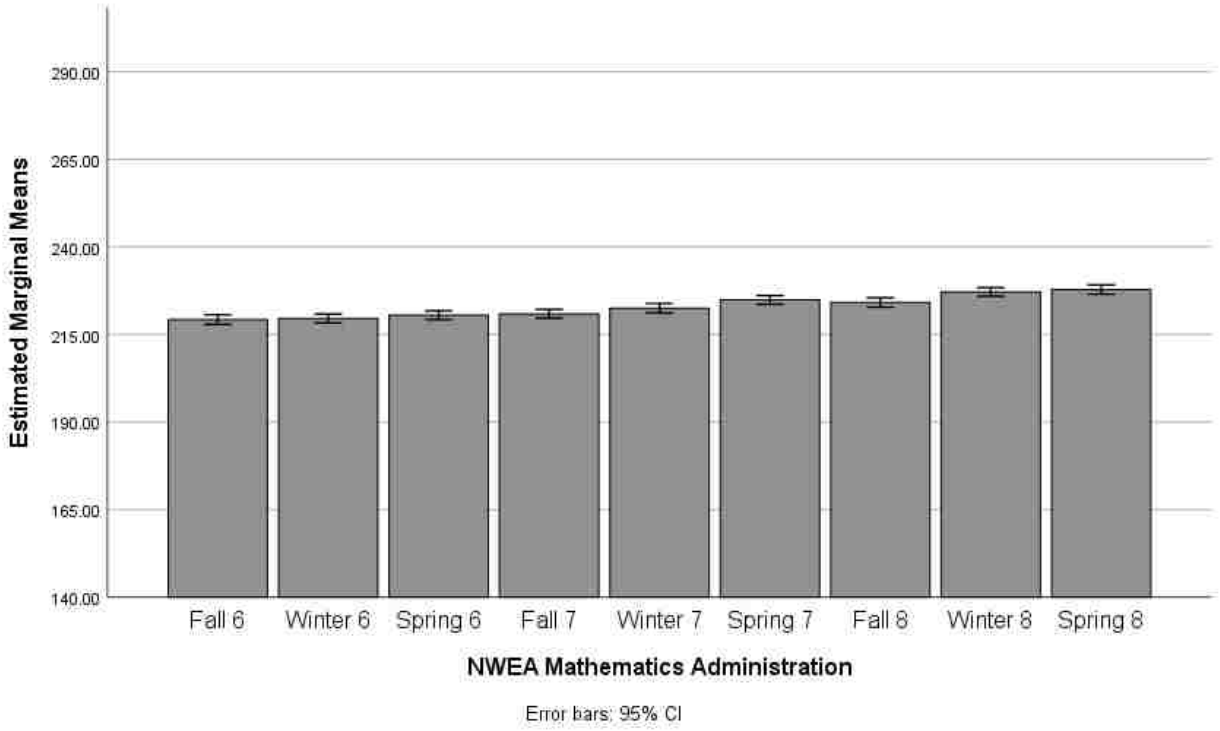


Table 18 showed the overall means and standard deviations for each administration of the NWEA MAP mathematics interim assessment. Grade 7 demonstrated the largest gains in overall group means, 7.68 RIT units, compared to 5.52 and 5.90 for grade 6 and grade 8, respectively. However, the 2.16 RIT units of nominal growth advantage shown during grade 7 mirrored the decline in overall group means of 2.65 RIT units from Spring 6 to Fall 7. Thus the within-year growth observed over grade 7 included recapture of the decline from Spring 6 to Fall 7. The overall mean declined slightly from Spring 7 to Fall 8.

Table 18

NWEA MAP Mathematics Overall Means by Administration

NWEA MAP	Group Mean	Standard Deviation
Fall 6	228.41	12.76
Winter 6	230.79	12.93
Spring 6	233.93	12.51
Fall 7	231.28	12.87
Winter 7	234.33	12.81
Spring 7	238.96	13.52
Fall 8	238.36	14.28
Winter 8	240.77	13.49
Spring 8	244.26	15.07

RM-ANOVA Results. I conducted a one-way, RM-ANOVA to compare the effect of informed instruction over time on student achievement as measured by the NWEA MAP mathematics interim assessment over the course of grades 6, 7, and 8. I inspected the data from each NWEA MAP test administration using a histogram to validate the assumption of normality and found the data be reasonably normally distributed. The data failed the assumption of sphericity as Mauchly's Test of Sphericity was found to be significant, $\chi^2(35) = 160.63, p < .001$. To correct for the deviation in sphericity, I interpreted the significance of the data using the Greenhouse-Geisser adjustment.

Table 19

Pairwise Mean Differences in NWEA MAP Mathematics Overall Means

	1	2	3	4	5	6	7	8
1. Fall 6								
2. Winter 6	2.385**							
3. Spring 6	5.526**	3.142**						
4. Fall 7	2.877**	.492	-2.649**					
5. Winter 7	5.926**	3.542**	.400	3.049**				
6. Spring 7	10.557**	8.172**	5.031**	7.680**	4.631**			
7. Fall 8	9.957**	7.572**	4.431**	7.080**	4.031**	-.600		
8. Winter 8	12.366**	9.982**	6.840**	9.489**	6.440**	1.809**	2.409**	
9. Spring 8	15.852**	13.468**	10.326**	12.975**	9.926**	5.295**	5.895**	3.486**

*** p < .001*

The results of the RM-ANOVA showed significant variance in NWEA MAP group means, $F(6.90,2234.13) = 367.48,\ p < .001,\ \omega^2 = .53$. Thus, I rejected the null hypothesis that no significant variance among test administration existed. As shown in Table 19, Bonferroni post-hoc analysis revealed significant pairwise growth, $p < .001$, in all but three parings, Winter 6 and Fall 7, Spring 6 and Winter 7, and Spring 7 and Fall 8.

To investigate the significance of the growth in overall means, the analysis included a planned contrast which compared each NWEA MAP administration mean to the average of previous test administration means. The first line (Table 20) compared the means of the first two NWEA MAP administrations, Winter 6 and Fall 6. After the first line, each of following lines related the means of the subsequent test administration to the aggregated means of the previous NWEA MAP test administrations. For example, line two compared the overall mean from the Spring 6 administration to the combined means from the previous two administrations, Winter 6 and Fall 6. Each contrast except for the Fall 7, $F(1, 324) = .78,\ p > .05$, was found to be significant, $p < .001$.

Table 20

Tests of Within-Subject Contrasts – Mathematics Administration Versus Combined Previous

		Mean Square	F	Sig.	Partial Eta Squared	Observed Power
NWEA MAP	Winter 6 vs. Fall 6	1848.08	46.00	.000	.124	1.000
	Spring 6 vs Previous	6104.22	218.69	.000	.430	1.000
	Fall 7 vs Previous	18.72	.78	.380	.002	.142
	Winter 7 vs Previous	3389.08	113.69	.000	.260	1.000
	Spring 7 vs Previous	16914.31	479.96	.000	.597	1.000
	Fall 8 vs Previous	9518.45	279.33	.000	.463	1.000
	Winter 8 vs Previous	16143.75	588.91	.000	.645	1.000
	Spring 8 vs Previous	30284.12	895.46	.000	.734	1.000

Descriptive Analysis. This study employed a two-part descriptive analysis. First, to analyze longitudinal growth, this study tracked NWEA MAP mathematics RIT scores across grades 6, 7, and 8 disaggregated by performance level descriptor. NWEA (2010) developed concordance cut scores, minimum scores for membership in the performance level descriptor, relating NWEA MAP scores by season to PSSA scores. NWEA (2010) noted that the minimum score for each range for the Fall and Spring administration represented the lowest score that corresponded "to a 50% probability of achieving that performance level" (p. 4). In their 2010 linking study, NWEA did not publish cut scores for the Winter administration. This study interpolated Winter cut scores based upon the cut scores available for Fall and Spring. I used the NWEA MAP Fall and Spring cuts scores and the interpolated Winter cuts scores to define RIT Ranges for each performance level. Using NWEA MAP cut scores, I converted NWEA MAP mathematics RIT scores into performance level projections.

As shown in Table 21, I tallied group membership, *n*, by performance level and calculated group means, $\overline{x}$, for each performance level (Advanced, Proficient, Basic, and Below Basic). Group membership fluctuated across test administrations indicating movement among groups. For example, membership in the Advanced performance level varied from a low of 196 students in Fall 7 to a high of 231 students in Spring 7. While group membership varied for individual performance levels, the number of students who scored Proficient or above remained relatively constant, $\bar{x} = 286.3, SD = 3.94$.

Group means showed positive growth over time in every performance level descriptor from Fall 6 through Spring 8. However, growth of group means was non-linear and exhibited instances of decline between Spring 6 and Fall 7 in the Advanced, Proficient, and Basic performance levels. The earlier analysis of overall group means suggested this pattern.

Similarly, the cut score defined by NWEA declined by one RIT unit in Advanced and 3 RIT units in Basic from Spring 6 to Fall 7. NWEA derived these cut scores as a result of a norming study which used a population of 6,000 students over 15 districts and did not note potential causes for this decline. The decline in cut scores suggested that the decline from Spring 6 to Fall 7 observed in this study reflected the larger population of Pennsylvania NWEA MAP test takers.

Table 21

NWEA MAP Mathematics Longitudinal Movement by Performance Level

		Fall			Winter			Spring		
Grade	Performance Level	RIT Range	*n*	$\bar{x}$	RIT Range*	*n*	$\bar{x}$	RIT Range	*n*	$\bar{x}$
6	Advanced	224-300	211	235.98	227-300	207	238.60	230-300	211	241.21
	Proficient	213-223	73	218.73	216-226	76	221.79	218-229	80	224.28
	Basic	206-212	26	209.77	208-215	26	212.23	210-217	23	215.04
	Below Basic	140-205	15	201.27	140-207	16	202.69	140-209	11	204.09
7	Advanced	229-300	196	239.68	231-300	217	241.44	233-300	231	245.69
	Proficient	218-228	83	223.35	220-230	68	225.10	222-232	61	227.11
	Basic	207-217	37	212.54	209-219	29	215.52	210-221	24	217.79
	Below Basic	140-206	9	198.67	140-208	11	200.73	140-209	9	203.00
8	Advanced	233-300	219	245.98	235-300	220	248.09	237-300	227	251.91
	Proficient	223-232	67	227.85	225-234	67	229.49	226-236	63	231.51
	Basic	214-222	26	218.96	216-224	30	221.27	217-225	22	221.64
	Below Basic	140-213	13	203.00	140-215	8	207.13	140-216	13	210.77

N = 328.

Note: *Winter RIT Range interpolated from Fall and Spring (NWEA, 2010).

To further analyze the performance by category membership, this study tracked the within-year movement during the 7th grade year. As shown in Table 22, students were characterized by their membership in a performance level descriptor category based upon their score on the Fall 7 NWEA MAP. Student scores were tracked across the three 7th grade NWEA MAP administrations. Overall, 70 students (21.5%) increased their spring performance level from their fall performance level, 242 students (74.5%) finished at the same level, and 19

Table 22
7th Grade NWEA MAP Mathematics Movement by Performance Level

Fall	Winter	Spring	Students	Percent of PLD	Percent of Cohort
Advanced	Advanced	Advanced	181	92.35	55.69
		Proficient	4	2.04	1.23
	Proficient	Advanced	9	4.59	2.77
		Proficient	1	0.51	0.31
	Basic	Basic	1	0.51	0.31
			196	100.00	60.31
Proficient	Advanced	Advanced	21	25.30	6.46
		Proficient	10	12.05	3.08
		Basic	1	1.20	0.31
	Proficient	Advanced	12	14.46	3.69
		Proficient	22	26.51	6.77
		Basic	5	6.02	1.54
	Basic	Advanced	1	1.20	0.31
		Proficient	8	9.64	2.46
		Basic	3	3.61	0.92
			83	100.00	25.54
Basic	Proficient	Advanced	6	16.22	1.85
		Proficient	9	24.32	2.77
		Basic	2	5.41	0.62
		Below	2	5.41	0.62
	Basic	Proficient	3	8.11	0.92
		Basic	6	16.22	1.85
		Below	3	8.11	0.92
	Below	Advanced	1	2.70	0.31
		Proficient	3	8.11	0.92
		Below	2	5.41	0.62
			37	100.00	11.38
Below	Basic	Proficient	1	11.11	0.31
		Basic	2	22.22	0.62
		Below	1	11.11	0.31
	Below	Basic	4	44.44	1.23
		Below	1	11.11	0.31
			9	100.00	2.77

students (5.8%) declined one or more levels. The middle two performance levels, where both movement up and down was possible, better captured movement among categories. Of the students who performed at the Proficient level on Fall 7, 34 students (41.0%) improved to

Advanced, 40 students (48.2%) remained at the Proficient level, and 9 students (10.8%) declined to Basic. Of the 37 students who performed at the Basic level on the Fall 7 MAP, 22 students (59.5%) improved by one or more levels, 8 students (21.0%) remained at the Basic level, and 7 students (18.9 %) declined to Below Basic.

Table 23

Mathematics PSSA Membership and Group Means by Performance Level

	PSSA 6			PSSA 7			PSSA 8		
	n	$\bar{x}$	%	n	$\bar{x}$	%	n	$\bar{x}$	%
Advanced	211	1727.20	65.53	226	1737.26	70.19	220	1692.12	68.32
Proficient	67	1390.55	20.81	67	1387.37	20.81	58	1362.93	18.01
Basic	38	1248.89	11.80	17	1240.41	5.28	30	1235.27	9.32
Below Basic	6	1088.67	1.86	12	1135.67	3.73	14	1093.86	4.35

N = 322

To compare the performance on NWEA MAP to the Pennsylvania accountability assessments, Table 23 tallied the group membership by performance level for the PSSA. Three students from the 325 student sample did not take the PSSA 8 and were therefore removed from

Table 24

Mathematics Student Performance Level Movement PSSA 6 through PSSA 8

	PSSA 6	Spring 6	Fall 7	Winter 7	PSSA 7	Spring 7	Fall 8	Winter 8	PSSA 8
Advanced	211	210	194	215	226	229	217	218	220
Proficient	67	79	83	68	67	61	67	67	58
Basic	38	23	37	29	17	23	26	30	30
Below Basic	6	10	8	10	12	9	12	7	14
Percent Advanced & Proficient	86.34	89.75	86.02	87.89	90.99	90.06	88.20	88.51	86.34

this calculation. The number of students who achieved proficiency increased from PSSA 6 to PSSA 7 by 15 students but then declined by 15 students for PSSA 8. Viewed longitudinally from grade 6 through grade 8, the number of students who achieved proficiency, 278, did not change.

The PSSA accountability calendar, that is PSSA to PSSA, did not align with the NWEA MAP grade level assessment designations. Students took the NWEA Spring MAP in May of each school year after the corresponding PSSA which students took in March of that year. For example, starting with the PSSA 6, students took Spring 6, Fall 7, and Winter 7, before taking the PSSA 7. To investigate how NWEA MAP growth compared to PSSA growth, this study reviewed the same NWEA MAP longitudinal growth but applied the accountability calendar, PSAA to PSSA (Table 24). Longitudinal data on group membership by performance category showed fluctuations in group membership between assessments but ultimately lacked clear evidence of growth.

Summary of Mathematics Growth. The analysis of whether NWEA MAP mathematics scores differed significantly over time found evidence of statistically significant growth via the RM-ANOVA. Because this analysis sought to investigate the utility of repeated administrations of the NWEA MAP with regard to their utility to inform instruction and ultimately, improve student outcomes, this RM-ANOVA employed several analytics to investigate trends. This study employed Bonferroni post-hoc analysis for a pairwise comparison of means, contrasts to compare administration means to the previous aggregated means, and descriptive analysis on longitudinal growth and movement by performance level.

This descriptive analysis of longitudinal trends identified several findings that need further investigation. From the simplest descriptive statistics, the data showed a decline from Spring to Fall coincident with the absence of instruction during summer months. Bonferroni pairwise analysis supported the significant decline from Spring 6 to Fall 7 and further noted the non-significant growth between three pairs including two no-consecutive pairs, Winter 6 to Fall 7 and Spring 6 to Winter 7. The non-significant growth in non-consecutive test administrations

showed the nominal gains in RIT scores did not register a statistical, let alone a practical, significance in student achievement. Comparison of administration means to the aggregate means of previous administrations supported this finding as Fall 7 produced a non-significant result.

The additional descriptive analysis of movement by performance level, noted longitudinal increases in group means in each performance level and despite fluctuations between administrations, net positive movement of six students scoring Proficient or Advanced from the Fall 6 administration to the Spring 8 administration. Furthermore, the analysis of grade 7 movement of students who scored in the Basic and Proficient performance levels found 56 of the 120 students (48.7%) improved by one or more category by Spring 7, while 48 students (40%) persisted in the same category, and 16 students (13.3%) declined.

Ideally, the analysis of NWEA MAP RIT scores would have revealed growth across group means and movement in group membership at both Advanced and Proficient levels. NWEA MAP gains would have been evidenced in growth on the PSSA. However, while NWEA MAP RIT scores noted statistically significant gains, the percent of students who scored Proficient and Advanced from PSSA 6 to PSSA 8 did not reflect this growth.

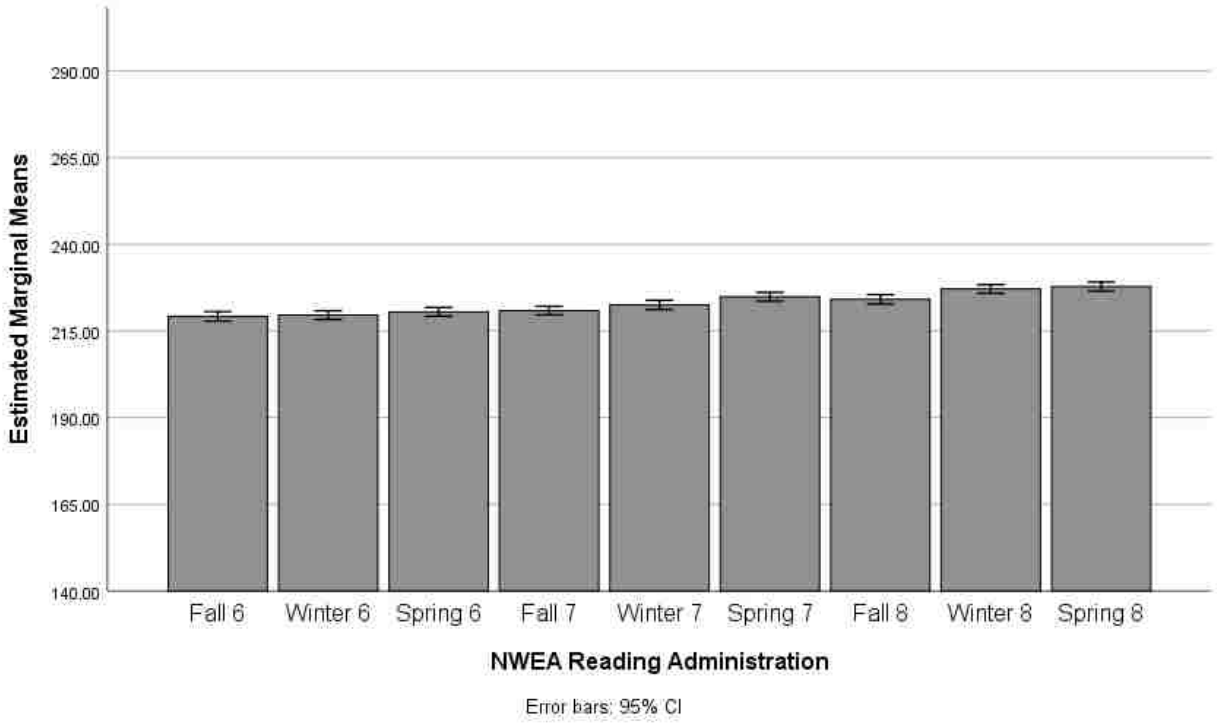
Reading

To analyze whether NWEA MAP reading interim assessment scores differed significantly over time, I applied the same analysis employed for mathematics. The data set for the RM-ANOVA was comprised of 328 students who had each taken all three NWEA MAP interim assessments, Fall, Spring, and Winter in reading for each grade 6, 7, and 8. I entered all nine NWEA MAP reading RIT scores into SPSS as within-subjects factors for analysis using

RM-ANOVA. As shown in Figure 2, overall test administration scores tended to increase throughout each school year.

Figure 2

Overall Mean RIT Scores for NWEA MAP Reading by Administration



RM-ANOVA Results. I conducted a one-way, repeated measures ANOVA to compare the effect of informed instruction over time on student achievement as measured by the NWEA MAP reading interim assessment over the course of grades 6, 7, and 8. I inspected the data for each NWEA MAP test administration using a histogram to validate the assumption of normality and found the data to be reasonably normally distributed. The data failed the assumption of sphericity, as Mauchly's Test of Sphericity was found to be significant, $\chi^2(35) = 83.46,\ p < .001$. To correct for the deviation from sphericity, I interpreted the significance using the Greenhouse-Geisser adjustment.

Table 25

NWEA MAP Reading Overall Means by Administration

NWEA MAP	Group Mean	Standard Deviation
Fall 6	219.23	12.59
Winter 6	219.58	11.74
Spring 6	220.52	11.66
Fall 7	220.89	11.47
Winter 7	222.50	12.42
Spring 7	224.87	11.60
Fall 8	224.15	12.39
Winter 8	227.14	11.32
Spring 8	227.82	12.09

Table 25 showed the overall means and standard deviations for each administration of the NWEA MAP reading interim assessment. Overall means for the first 4 administrations of the NWEA MAP reading interim assessment, Fall 6 through Fall 7, showed minimal growth of less than 2 RIT units. NWEA MAP did not decline from Spring 6 to Fall 7 but did decline from Spring 7 to Fall 8.

Table 26

Pairwise Mean Differences in NWEA MAP Reading Overall Means

	1	2	3	4	5	6	7	8
1. Fall 6								
2. Winter 6	.345							
3. Spring 6	1.284	.940						
4. Fall 7	1.659*	1.314*	.375					
5. Winter 7	3.265**	2.921**	1.982*	1.607**				
6. Spring 7	5.637**	5.293**	4.354**	3.979**	2.372**			
7. Fall 8	4.918**	4.573**	3.634**	3.259**	1.652*	-.720		
8. Winter 8	7.909**	7.564**	6.625**	6.250**	4.643**	2.271**	2.991**	
9. Spring 8	8.585**	8.241**	7.302**	6.927**	5.320**	2.948**	3.668**	.677

*** p < .001, *p < .05*

The results of the RM-ANOVA showed statistically significant variance in NWEA MAP group means, $F(7.45,2436.26) = 106.81,\ p < .001,\ \omega^2 = .25$. Thus, I rejected the null hypothesis that no significant variance existed among administration means. As shown in Table

26, Bonferroni post-hoc analysis revealed five pairings with non-significant pairwise growth, $p > .05$. These non-significant pairwise comparisons were concentrated in the first four test administrations, Fall 6 to Winter 6, Winter 6 to Spring 6, Fall 6 to Winter 6, Spring 6 to Fall 7, and Winter 8 to Spring 8.

To investigate the significance of the growth in overall means, the analysis included a planned contrast which compared each NWEA MAP administration mean with the average of previous test administration means. As shown in Table 27, each contrast except for Winter 6, Spring 6, and Fall 7 was found to be significant, $p < .001$.

Table 27

Tests of Within-Subject Contrasts – Reading Administration Versus Combined Previous

		Mean Square	F	Sig.	Partial Eta Squared	Observed Power
NWEA MAP	Winter 6 vs. Fall 6	38.93	.70	.404	.002	.133
	Spring 6 vs Previous	405.06	8.64	.004	.026	.834
	Fall 7 vs Previous	408.40	10.41	.001	.031	.896
	Winter 7 vs Previous	1958.54	42.63	.000	.115	1.000
	Spring 7 vs Previous	6140.64	184.85	.000	.361	1.000
	Fall 8 vs Previous	2732.25	66.40	.000	.169	1.000
	Winter 8 vs Previous	9795.12	265.60	.000	.448	1.000
	Spring 8 vs Previous	9772.69	279.04	.000	.460	1.000

Descriptive Analysis. This study employed a two-part descriptive analysis. First, to analyze longitudinal growth, this study tracked NWEA MAP reading RIT scores across grades 6, 7, and 8 disaggregated by performance level. I employed the identical process described in the mathematics analysis to create RIT ranges for each performance level. As with the mathematics sections, I used the available cut scores for Fall and Spring and interpolated cut scores for the Winter administrations. Using NWEA MAP cut scores, I converted RIT scores into performance level projections.

Table 28

NWEA MAP Reading Longitudinal Movement by Performance Level

		Fall			Winter			Spring		
Grade	Performance Level	RIT Range	*n*	$\bar{x}$	RIT Range*	*n*	$\bar{x}$	RIT Range	*n*	$\bar{x}$
6	Advanced	222-300	147	229.89	224-300	133	230.84	225-300	134	231.27
	Proficient	208-221	130	215.51	210-223	129	216.67	211-224	133	217.79
	Basic	198-207	35	203.06	200-209	50	204.98	201-210	46	206.20
	Below Basic	140-197	16	186.94	140-199	16	194.94	140-200	15	192.53
7	Advanced	226-300	111	232.66	227-300	135	233.79	228-300	138	235.58
	Proficient	213-225	146	219.78	214-226	119	220.08	215-227	128	221.70
	Basic	205-212	46	209.07	206-213	50	210.62	207-214	42	211.21
	Below Basic	140-204	25	196.88	140-212	24	195.71	140-206	20	199.90
8	Advanced	223-300	192	232.37	224-300	206	234.02	225-300	205	235.20
	Proficient	212-222	92	217.66	213-223	92	218.79	214-224	90	219.40
	Basic	206-211	18	209.22	207-212	17	209.59	208-213	15	210.80
	Below Basic	140-205	26	196.73	140-206	13	200.08	140-207	18	200.06

N = 328.

Note: *Winter RIT Range interpolated from Fall and Spring (NWEA, 2010).

As shown in Table 28, I calculated group membership, *n,* by performance level and group means, $\bar{x}$, for each performance level (Advanced, Proficient, Basic, and Below Basic). The within-year variance of group membership in Proficient and Advanced showed a decline during Grade 6 and increases in Grades 7 and 8. Across-year variance from Fall 6 to Spring 7 showed the number of students who performed at the Proficient level or above decreased from 177 students in Fall 6 to 166 students in Spring 7. From Spring 7 to Fall 8, the group membership in the Advanced category increased by 52 students, an improvement of 37.7%, while the group mean declined by more than 3 RIT units.

Closer inspection of the RIT ranges revealed a 5 unit drop in the Advanced cut score and a 3 unit drop in the cut score for Proficient. The gain in group membership in the Advanced level almost certainly reflected a categorical artifact rather than real growth in student outcomes.

The decrease in the group means for Advanced and Proficient further suggested that the categorical growth resulted from adjustment in cuts scores rather than improved student achievement. The lower cut score persisted throughout grade 8, with the Spring 8 Advanced cut score 3 units lower than the Spring 7 cut score.

To further analyze the performance by category membership, this study tracked the within-year movement during the 7th grade year. As shown in Table 29, students were characterized by their membership in a performance level descriptor category based upon their score on the Fall 7 NWEA MAP. Student scores were tracked across the three 7th grade NWEA MAP administrations.

Table 29

7th Grade NWEA MAP Reading Movement by Performance Level

Fall	Winter	Spring	Students	% of PLD	% of Cohort
Advanced	Advanced	Advanced	82	73.87	25.00
		Proficient	9	8.11	2.74
	Proficient	Advanced	11	9.91	3.35
		Proficient	7	6.31	2.13
	Basic	Advanced	1	0.90	0.30
	Below Basic	Below Basic	1	0.90	0.30
			111	100.00	33.84
Proficient	Advanced	Advanced	26	17.81	7.93
		Proficient	13	8.90	3.96
		Basic	1	0.68	0.30
	Proficient	Advanced	15	10.27	4.57
		Proficient	50	34.25	15.24
		Basic	11	7.53	3.35
	Basic	Proficient	16	10.96	4.88
		Basic	8	5.48	2.44
		Below Basic	1	0.68	0.30
	Below Basic	Proficient	2	1.37	0.61
		Basic	2	1.37	0.61
		Below Basic	1	0.68	0.30
			146	100.00	44.51
Basic	Advanced	Proficient	3	6.52	0.91
	Proficient	Advanced	2	4.35	0.61
		Proficient	14	30.43	4.27

		Basic	3	6.52	0.91
	Basic	Advanced	1	2.17	0.30
		Proficient	6	13.04	1.83
		Basic	7	15.22	2.13
		Below Basic	3	6.52	0.91
	Below Basic	Proficient	1	2.17	0.30
		Basic	1	2.17	0.30
		Below Basic	5	10.87	1.52
			46	100.00	14.02
Below	Advanced	Proficient	1	4.00	0.30
	Proficient	Proficient	3	12.00	0.91
		Basic	2	8.00	0.61
		Below Basic	1	4.00	0.30
	Basic	Proficient	1	4.00	0.30
		Basic	4	16.00	1.22
		Below Basic	2	8.00	0.61
	Below Basic	Proficient	2	8.00	0.61
		Basic	3	12.00	0.91
		Below Basic	6	24.00	1.83
			25	100.00	7.62

Overall, 81 students (24.7%) increased their performance from Fall 7 to Spring 7 by one or more categories, whereas 198 students (60.4%) remained in the same category and 49 students (14.9%) declined by one or more performance levels. In the middle two categories, Proficient and Basic, 68 students (35.4%) improved, 92 students (47.9%) persisted in the same category, and 32 students (16.7%) declined. Students in the Basic category exhibited more categorical movement than those in the Proficient. Of the 146 students who scored Proficient on the Fall 7 administration, 41 students (28.1%) moved to Advanced, whereas 81 (55.5%) remained Proficient and 24 students (16.4%) declined one or more levels. Of the 46 students who scored Basic on the Fall 7 administration, 27 students (58.7%) increased one or more levels, 11 students (23.9%) remained at the Basic level, and 8 students (17.4%) declined to Below Basic.

To compare the performance on NWEA MAP to the Pennsylvania accountability assessments, Table 30 tallied the group membership by performance level for the PSSA. Three

students from the 328 student sample did not take the PSSA 8 and were therefore removed from this calculation. The number of students who achieved proficiency increased from PSSA 6 to PSSA 7 by 33 students, an increase of more than 10%. Similarly, the number of students who achieved proficiency increased from PSSA 7 to PSSA 8 by 26 students. In addition to the increase in proficiency from PSSA 7 to PSSA 8, group membership increased dramatically in the Advanced category by 76 students which represented a 46.6% increase. Viewed longitudinally from grade 6 through grade 8, the number of students who achieved proficiency increased by 59 students including an increase of 92 students in the Advanced category.

Table 30

Reading Student Performance Level Movement PSSA 6 through PSSA 8

	PSSA 6	Spring 6	Fall 7	Winter 7	PSSA 7	Spring 7	Fall 8	Winter 8	PSSA 8
Advanced	146	133	110	135	163	137	190	205	239
Proficient	104	132	145	117	120	127	92	91	70
Basic	59	46	46	50	34	42	18	17	10
Below Basic	16	14	24	23	8	19	25	12	6
Percent Advanced & Proficient	76.92	81.54	78.46	77.54	87.08	81.23	86.77	91.08	95.08

N = 325

Summary of Reading Growth. The analysis of whether NWEA MAP reading scores differed significantly over time found evidence of statistically significant growth via the RM-ANOVA. Because this analysis sought to investigate the utility of repeated administrations of the NWEA MAP to inform instruction and ultimately, improve student outcomes, this RM-ANOVA employed several analytics to investigate trends. This study employed Bonferroni post-hoc analysis for a pairwise comparison of means, planned contrasts to compare administration means

to the previous aggregated means, and descriptive analysis on longitudinal growth and movement by performance level.

The RM-ANOVA results found significant growth in the mean RIT scores from Fall 6 through Spring 8. The most significant growth in group means (80% of total growth) occurred from Fall 7 to Spring 8 with very little growth observed over the first four NWEA MAP administrations, Fall 6 to Fall 7. Bonferroni post-hoc and the planned contrasts within the RM-ANOVA further evidenced this asymmetrical growth pattern. Bonferroni pairwise analysis noted non-significant growth, $p > 0.5$, between four pairs including the non-consecutive pair, Fall 6 to Spring 6. Additionally, using the more demanding $p < .001$ significance level, Fall 6 to Fall 7 yielded a non-significant result. Therefore, non-significant growth from Fall 6 through Fall 7 spanned the entre Grade 6 and showed that the nominal gains in RIT scores did not register a statistical, let alone a practical, significance for the entire grade 6 year and into the fall of grade 7. Comparison of administration means to the aggregate means of previous administrations further supported non-significant growth from Fall 6 through Fall 7.

The additional descriptive analysis of across-year movement by performance level noted longitudinal increases in group means in each performance level. Despite fluctuations between administrations, movement showed a net positive increase of 18 students scoring Proficient or Advanced from the Fall 6 administration to the Spring 8 administration. Anomalous changes in the RIT ranges for the grade 8 performance levels confounded interpretation of growth in student outcomes. NWEA (2010) noted an 18 point drop in percentile at which the Advanced cut score was set for Fall 8. NWEA set percentiles for the Advanced cut scores for Fall 6 and Fall 7 at the 77^{th} and 76^{th} percentiles respectively, whereas the Advanced cut score for Fall 8 corresponded to the 58^{th} percentile. While not specifically referencing the PSSA 8 Reading assessment, NWEA

noted that state accountability categorical designations and year-to-year difficulty of the NCLB accountability exams varied significantly and that these variations were reflected in NWEA cut scores. In the analysis of grade 7 within-year movement in the Proficient and Basic performance levels, roughly half of students persisted at the same performance levels. Of the 100 students who scored Basic or Proficient and moved performance levels, increases in performance outnumbered declines by a 2 to 1 ratio.

Question 2: Predictive Utility of NWEA MAP

To analyze the extent to which repeated administrations of NWEA MAP interim assessments contributed to the utility to predict performance on the PSSA this study employed a block-wise multiple regression of student demographics, existing student achievement data, and NWEA MAP interim assessment data. This multiple regression took the general form, $PSSA\ 7_i = b_0 + b_1 Predictor_{1i} + \cdots + b_n Predictor_{ni}$, where b represented the coefficients of a predictor variable from the multiple regression for each student, *i*. This study evaluated three dichotomous and five continuous variables to predict the continuous PSSA 7 outcome. The analysis was considered separately for mathematics and reading in Questions 2a and 2b respectively.

Question 2a: Mathematics

To analyze the extent to which repeated administrations of NWEA MAP mathematics assessments contributed to the overall utility to predict performance on the mathematics PSSA, I used a multiple regression. This study hypothesized that each administration of the NWEA MAP mathematics interim assessment would individually and significantly contribute to the overall predictive value of the model. As displayed in Table 31, the preliminary correlations showed significant correlations, $p < .001$, for seven of the eight predictor variables, with only

school membership showing non-significant results. The correlations among the NWEA MAP administrations were high (r >.800) and therefore required a collinearity analysis. Variance inflation factor (VIF) for each predictor variable was found to be less than 10 and therefore suitable (Meyers, 1990).

I completed an inspection of potential outliers in the data set. The data set showed eight data points that lay outside two standard deviations from the predicted values including three data points that lay more than three standard deviations outside of predicted values. These eight data points represented 2.3% of the data set thus falling below the expected 5% distribution. Table 32 showed the means and standard deviations for the continuous variables.

Table 31

Pearson Correlation Predictor Variables - Mathematics

	PSSA7	School	IEP	EconDis	PSSA6	Grade	Fall7	Winter7
School	.140**							
IEP	-.371*	.010						
EconDis	-.306*	-.046	.222*					
PSSA6	.841*	.095	-.374*	-.280*				
Grade	.523*	-.262	-.263*	-.204*	.478*			
Fall7	.839*	.112	-.348*	-.331*	.838*	.456*		
Winter7	.855*	.099	-.401*	-.354*	.820*	.464*	.883*	
Spring7	.867*	.087	-.381*	-.352*	.810*	.488*	.861*	.867*

Note: *$p < .001$, ** $p < .05$

This study employed a block-wise multiple regression. As shown in Table 33, the multiple regression calculated five successive regression models starting with demographic information, then adding student achievement data, and finally individual NWEA MAP administrations. All models were found to be significant, $p < .001$.

Table 32

Means and Standard Deviation - Mathematics

Variable	Mean	Standard Deviation
PSSA 7	1618.46	240.091
PSSA 6	1592.87	245.119
7th Grade Course Grade	3.32	.338
NWEA MAP RIT Fall 7	231.68	12.354
NWEA MAP RIT Winter 7	234.62	12.427
NWEA MAP RIT Spring 7	239.36	13.223

N = 335

The first model, comprised of demographic data, school membership, IEP status, and economic disadvantage status explained 20.8 percent of the variation in PSSA 7 scores. Model 2 augmented these student demographics data with existing student achievement data, PSSA 6 mathematics scores and mathematic end-of-course grades. The addition of student achievement data added predictive power, $\Delta R^2 = .537$, with 74.5% of the variation of PSSA 7 scores explained. As the regression model added each successive NWEA MAP in models 3, 4, and 5, comparatively small, decreasing gains in the predictive power of the overall model were noted. NWEA Fall explained an added 4.5%, whereas Winter and Spring added 2.4% and 2.2% respectively.

Table 33

PSSA 7 Mathematics Predictor Variables Model Summary

Predictor Variables Included	R^2	ΔR^2	F Change
School, IEP, Economic Disadvantage	.208	.208	28.970
6th Grade PSSA, 7th Grade EOC Grade	.745	.537	346.707
7th Grade Fall NWEA MAP	.790	.045	69.428
7th Grade Winter NWEA MAP	.814	.024	42.986
7th Grade Spring NWEA MAP	.836	.022	44.129

Note: $p < .001$.

Regression Coefficients. This multiple regression took the form, $PSSA\ 7_i = b_0 + b_1 School_i + b_2 IEP_i + b_3 EconDis_i + b_4 PSSA6_i + b_5 Grade_i + b_6 Fall7_i + b_7 Winter7_i +$

$b_8 Spring7_i$, where b represented the coefficients from the multiple regression for each student, i. Table 34 listed the coefficients for each predictor variable in each of the 5 models. Model 1 included only three demographic dichotomous variables, School Membership (0 – East, 1 – West), IEP status (0 – No, 1-Yes), and economic disadvantage (0 – No, 1 – Yes). Using only these demographic data, both IEP and Economic Disadvantage status had significant negative effects on the students' grade 7 PSSA scores. Inclusion of existing student academic achievement data as was the case in Model 2, prior year PSSA 6 and the end of course grade, changed the regression such that IEP and Economic Disadvantage were no longer significant. PSSA 6 score β = .704 and course grade β = .196 were both powerful predictor variables in Model 2.

In Model 3, the first of the NWEA MAP interim assessment data wer considered along with demographic information, existing student achievement data, PSSA 6 and course grade. Fall 7, $\beta_3 = 0.398$, was found to be significant, $p < .001$. This value indicated that, holding all other predictor variables constant, as a student's NWEA MAP Fall 7 interim assessment increased by one standard deviation (12.354 points), PSSA 7 mathematics score would increase by β standard deviations. The standard deviation for PSSA 7 mathematics scores was 240.091 points, therefore we would expect that a 12.354 point increase in NWEA MAP Fall 7 would yield a corresponding increase of 95.556 points in the PSSA 7 score, (0.398 x 240.091). Past performance on the PSSA remained the strongest predictor, PSSA 6 $\beta_3 = 0.402$.

In Model 4 with the addition of Winter 7 NWEA MAP, Winter 7 ($\beta_4 = 0.359$) was the most powerful predictor of PSSA 7. PSSA 6 remained a strong predictor with a standardized

Table 34

Multiple Regression Coefficients - Mathematics

Model	Predictor	B	Std. Error	Beta	Sig.
1	School	65.678	24.201	0.133	.007
	IEP	-259.047	40.364	-0.322	.000
	Econ Dis	-125.779	27.670	-0.228	.000
	(Constant)	1650.993	16.709		.000
2	School	60.675	14.783	0.123	.000
	IEP	-36.910	24.467	-0.046	.132
	Econ Dis	-29.054	16.177	-0.053	.073
	PSSA6	0.690	0.034	0.704	.000
	Grade	139.128	24.216	0.196	.000
	(Constant)	45.755	77.161		.554
3	School	49.184	13.520	0.100	.000
	IEP	-29.877	22.278	-0.037	.181
	Econ Dis	-8.537	14.924	-0.015	.568
	PSSA6	0.394	0.047	0.402	.000
	Grade	115.275	22.218	0.162	.000
	Fall7	7.739	0.929	0.398	.000
	(Constant)	-1198.490	165.007		.000
4	School	46.316	12.738	0.094	.000
	IEP	-8.589	21.225	-0.011	.686
	Econ Dis	2.880	14.159	0.005	.839
	PSSA6	0.312	0.046	0.318	.000
	Grade	104.633	20.982	0.147	.000
	Fall7	3.403	1.096	0.175	.002
	Winter7	6.940	1.058	0.359	.000
	(Constant)	-1659.74	170.549		.000
5	School	44.157	11.977	0.089	.000
	IEP	-2.098	19.974	-0.003	.916
	Econ Dis	10.592	13.359	0.019	.428
	PSSA6	0.257	0.044	0.263	.000
	Grade	87.796	19.884	0.124	.000
	Fall7	1.207	1.082	0.062	.266
	Winter7	4.436	1.064	0.230	.000
	Spring7	6.17	0.929	0.340	.000
	(Constant)	-1899.46	164.316		.000

Note: Dependent Variable: PSSA7 Mathematics

Beta, $\beta_4 = .318$. Fall 7, while still significant $p < .05$, decreased in its relative predictive power with a standardized Beta, $\beta_4 = 0.175$. With all other predictor variables held constant, an increase in Fall 7 of 12.354 points would represent an increase of 42.016 points on PSSA 7. In Model 5 with the addition of Spring 7, PSSA 6 was the most powerful predictor, β_5 = .270, followed by Spring 7, $\beta_5 = 0..231$, and Winter 7, $\beta_5 = 0..228$. Fall 7 dropped to $\beta_5 = .062$ leading to a non-significant increase of 14.885 in PSSA 7 with a corresponding increase of 12.354 in Fall 7.

Summary of Question 2a

The data from the multiple regression showed that each NWEA MAP mathematics administration independently and significantly improved the predictive power of the model. The addition of the first NWEA MAP data, improved the model more than the additions of additional NWEA MAP data, as evidenced by the decreasing changes in ΔR^2. Existing student achievement data, especially PSSA 6 data were found to be powerful predictors in Models 3-5, as evidenced by the large standardized Beta, $\beta_3 = .402$, $\beta_4 = .318$, and $\beta_5 = .263$. When 2 or more NWEA MAP interim assessments were included, such as Fall and Winter in Model 4, and Fall, Winter, and Spring in Model 5, Fall 7 lost significance.

Question 2b: Reading

To analyze the extent to which repeated administrations of NWEA MAP reading assessments contributed to the overall utility to predict performance on the reading PSSA 7, I conducted a block-wise multiple regression. This study hypothesized that each administration of the NWEA MAP mathematics interim assessment would individually and significantly contribute to the overall predictive value of the model. This study evaluated three dichotomous and five continuous variables to predict the continuous PSSA 7 outcome.

As displayed in Table 35, the preliminary correlations showed significant correlations, $p < .001$, for seven of the eight predictor variable,s with only school membership showing non-significant results. The correlations among the NWEA MAP administrations showed high correlations

Table 35

Pearson Correlation Predictor Variables - Reading

Predictor	PSSA 7	School	IEP	EconDis	PSSA 6	Grade	Fall 7	Winter 7
School	0.110							
IEP	-0.363*	0.002						
EconDis	-0.28*	-0.039	0.224*					
PSSA 6	0.761*	0.093	-0.314*	-0.251*				
Grade	0.658*	-0.02	-0.228*	-0.226*	0.628*			
Fall 7	0.719*	-0.007	-0.402*	-0.278*	0.725*	0.572*		
Winter 7	0.763*	0.03	-0.374*	-0.235*	0.690*	0.579*	0.759*	
Spring 7	0.776*	0.066	-0.320*	-0.269*	0.708*	0.619*	0.771*	0.810*

Note: *$p < .001$

and therefore required a collinearity analysis. I found variance inflation factors (VIF) for each predictor variable to be less than 10 and therefore suitable (Meyers, 1990). Furthermore, the data set showed 13 data points that lay outside two standard deviations from the predicted values including two data points that lay more than three standard deviations outside of predicted values. These 13 data points represented 3.7% of the data set thus falling below the expected 5% distribution.

Table 36

Means and Standard Deviation - Reading

Variable	Mean	Standard Deviation
PSSA 7	1479.24	192.304
PSSA 6	1423.11	194.203
7th Grade Course Grade	3.34	.343
NWEA MAP RIT Fall 7	220.96	11.026
NWEA MAP RIT Winter 7	222.33	12.180
NWEA MAP RIT Spring 7	224.89	11.432

N = 350

This study employed a block-wise multiple regression to answer this question. As shown in Table 37, the multiple regression calculated five successive regression models starting with demographic information then adding student achievement data, and finally, individual NWEA MAP administrations. This study found all models to be significant, $p < .001$. The first model, comprised of demographic data, school membership, IEP status, and Economic Disadvantage status explained 18.4% of the variation in PSSA 7 scores. Model 2 augmented these student demographics data with existing student achievement data, PSSA 6 reading scores and ELA end of course grades. The addition of student achievement data added predictive power, $\Delta R^2 = .470$, with 65.4% of the variation of PSSA 7 scores explained. Models 3-5, added NWEA MAP interim assessment data to the model and each successive addition provided a small increase $\Delta R^2 = .030$ for Model 3, $\Delta R^2 = .040$ for Model 4, and $\Delta R^2 = .014$ for Model 5. NWEA MAP Fall explained an added 3.0%, whereas Winter and Spring added 4.0% and 1.4% respectively.

Table 37

PSSA 7 Reading Predictor Variables Model Summary

Predictor Variables Included	R^2	ΔR^2	F Change
School, IEP, Economic Disadvantage	.184	.184	25.998
6th Grade PSSA, 7th Grade EOC Grade	.654	.470	233.874
7th Grade Fall NWEA MAP	.684	.030	32.570
7th Grade Winter NWEA MAP	.724	.040	49.951
7th Grade Spring NWEA MAP	.738	.014	18.053

Note: $p < .001$.

Regression Coefficients. This multiple regression took the form, $PSSA\ 7_i = b_0 + b_1 School_i + b_2 IEP_i + b_3 EconDis_i + b_4 PSSA6_i + b_5 Grade_i + b_6 Fall7_i + b_7 Winter7_i + b_8 Spring7_i$, where *b* represented the coefficients from the multiple regression for each student, *i.* Table 38 listed the coefficients for each predictor variable in each of the five models.

Table 38

Multiple Regression Coefficients - Reading

Model	Predictor	B	Std. Error	Beta	Sig.
1	School	41.124	19.379	0.103	.035
	IEP	-195.908	30.762	-0.317	.000
	EconDis	-90.978	22.16	-0.205	.000
	(Constant)	1508.088	13.155		.000
2	School	26.349	12.773	0.066	.040
	IEP	-75.025	20.895	-0.122	.000
	EconDis	-23.744	14.801	-0.053	.110
	PSSA 6	0.514	0.042	0.519	.000
	Grade	164.531	23.053	0.294	.000
	(Constant)	203.038	65.451		.002
3	School	32.459	12.272	0.081	.009
	IEP	-45.918	20.638	-0.074	.027
	EconDis	-16.148	14.227	-0.036	.257
	PSSA 6	0.366	0.048	0.370	.000
	Grade	138.936	22.514	0.248	.000
	Fall 7	4.697	0.823	0.269	.000
	(Constant)	-546.708	145.542		.000
4	School	30.398	11.483	0.076	.008
	IEP	-30.151	19.435	-0.049	.122
	EconDis	-17.141	13.310	-0.039	.199
	PSSA 6	0.293	0.046	0.296	.000
	Grade	111.938	21.405	0.200	.000
	Fall 7	1.888	0.867	0.108	.030
	Winter 7	5.237	0.741	0.332	.000
	(Constant)	-896.712	144.877		.000
5	School	26.217	11.250	0.066	.020
	IEP	-34.608	18.997	-0.056	.069
	EconDis	-13.681	13.016	-0.031	.294
	PSSA 6	0.267	0.045	0.270	.000
	Grade	95.115	21.262	0.170	.000
	Fall 7	0.773	0.885	0.044	.383
	Winter 7	3.600	0.819	0.228	.000
	Spring 7	3.880	0.913	0.231	.000
	(Constant)	-1065.317	146.858		.000

Note: Dependent Variable: PSSA7 Reading

Model 1 included only three demographic dichotomous variables, School Membership (0 – East, 1 – West), IEP status (0 – No, 1-Yes), and economic disadvantage (0 – No, 1 – Yes). Using only these demographic data, both IEP and Economic Disadvantage had significant negative effects on the students' grade 7 PSSA scores. Inclusion of existing student academic achievement data as was the case in Model 2, prior year PSSA 6 and the end of course grade, changed the regression such that IEP and Economic Disadvantage lost significance. PSSA 6 score β = .519 and course grade β = .294 were both powerful predictor variables in Model 2.

Model 3 included the first of the NWEA MAP interim assessment data along with demographic information, existing student achievement data, PSSA 6 and course grade. Fall 7, β_3 = .269, was found to be significant, $p < .001$. This value indicated that, holding all other predictor variables constant, as a student's NWEA MAP Fall 7 interim assessment increased by one standard deviation (11.03 points), PSSA 7 mathematics score increased by β standard deviations. The standard deviation for PSSA 7 mathematics scores was 192.30 points therefore we would expect that an 11.03 point increase in NWEA MAP Fall 7 would yield a corresponding increase of 51.73 points in the PSSA 7 score (0.269 x 192.30). Past performance on the PSSA remained the strongest predictor, PSSA 6 β_3 = .370.

In Model 4 with the addition of Winter 7 NWEA MAP, Winter 7 (β_4 = .332) was the most powerful predictor of PSSA 7. PSSA 6 remained a strong predictor with a standardized $\beta_4 = .296$. Fall 7, while still significant $p < .05$, decreased in its relative predictive power with a standardized $\beta_4 = .108$. With all other predictor variables held constant, an increase in Fall 7 of 11.03 points would represent an increase of 20.77 points on PSSA 7. In Model 5 with the addition of Spring 7, PSSA 6 was the most powerful predictor, β_5 = .270, followed by Spring 7,

$\beta_5 = .231$, and Winter 7, $\beta_5 = .228$. Fall 7 dropped to $\beta_5 = .044$ and was not a significant predictor in the model.

Summary of Question 2b

The data from the multiple regression showed that each NWEA MAP reading administration independently and significantly improved the predictive power of the model. As shown in the model summary (Table 37), the addition of each NWEA MAP interim assessment data improved the model as shown by a positive ΔR^2. This study found existing student achievement data, especially PSSA 6, to be a powerful predictor in each of the models in which these data were included. PSSA 6 data were the strongest predictor, as evidenced by the largest standardized Beta score, in Model 3 with NWEA MAP Fall present and in Model 5 with all three NWEA MAP assessments included. When two or more NWEA MAP assessments were included, as in Model 4 (NWEA Fall and Winter) and Model 5 (NWEA Fall, Winter, and Spring), the impact of the Fall NWEA as a predictor of PSSA 7 lost its significance, $p > .001$ in Model 3 and $p > .05$ in Model 5.

Question 3: Variation by Subject

Student Growth.

To determine whether the changes in NWEA MAP scores over time and the predictive utility of NWEA MAP scores varied by subject, I compared the data from the mathematical and reading analyses. I investigated student growth through a RM-ANOVA and a two-part descriptive analysis. Both mathematics and reading NWEA MAP interim assessment data showed statistically significant growth, $p < .001$, in the overall means across the nine administrations of the NWEA MAP interim assessment. I tabulated the salient results from the

RM-ANOVA and descriptive analysis of the overall NWEA MAP administration means in Table 39.

Table 39

Growth – NWEA MAP Administration Overall Means

	Mathematics	Reading
Trend Across-Grades	Non-linear, positive	Non-linear, positive
	Gain 15.85 RIT Units	Gain 8.59 RIT Units
	Statistically Significant**	Statistically Significant**
Decline	Spring 6 to Fall 7**	Spring 7 to Fall 8
	Spring 7 to Fall 8	
Pairwise Non-significant growth*	Winter 6 to Fall 7	Fall 6 to Winter 6
	Spring 6 to Winter 7	Fall 6 to Spring 6
		Winter 6 to Spring 6
		Spring 6 to Fall 7
		Winter 8 to Spring 8
Contrasts	Fall 7 to Previous**	Winter 6 to Fall 6*
		Spring 6 to Previous**
		Fall 7 to Previous**

Note: ** non-significance at $p > .001$, *non-significance at $p > .05$

This study found overall growth to be non-linear and positive for both subjects. Bonferroni pairwise post-hoc analysis revealed a decline in overall mean RIT scores between Spring 7 and Fall 8 administrations for both mathematics and reading. Additionally, mathematics declined Spring 6 to Fall 7 whereas reading evidenced non-significant growth over the same period. These declines between Spring and Fall coincided with the interruption of instruction that occurred in the summer months when school was not in session. Both mathematics and reading exhibited additional instances of pairwise non-significant growth. The non-significant growth in reading was localized to the 6th grade year and into the Fall 7 administration.

NWEA (2015) published school group growth norms to facilitate comparison among schools and relative to the larger population of NWEA MAP test-takers. Relative to the

percentage of the school group growth norms, students in the cohort did not meet growth norms for 6th grade but exceeded norms for both 7th and 8th grade in both mathematics and reading. Additionally, both subjects showed longitudinal growth in the course means by performance level. In the

Table 40

RIT Growth as Percent of NWEA MAP School Growth Norm (2015)

	Fall 6 - Spring 6	Fall 7 - Spring 7	Fall 8 - Spring 8
Mathematics			
RIT Mean Growth	5.52	7.68	5.90
School Growth Norm	7.71	5.95	4.63
Percent of Norm	71.59	129.07	127.43
Reading			
RIT Mean Growth	1.29	3.98	3.67
School Growth Norm	4.76	3.71	2.83
% of Norm	27.10	107.28	129.68

Thum & Hauser, 2015

descriptive analysis of student movement among categories over 7th grade, in both subjects, students tended to persist at the categorical performance level (Table 41). Students who scored at the Basic level were the notable exception, with upward trends in both mathematics and reading. At the Proficient level, mathematics students were more likely to increase than in reading.

Table 41

7th Grade Student Movement Fall to Spring by Performance Level - Percent (Number)

	Mathematics				Reading			
PLD	*n*	Increased	Stasis	Declined	*n*	Increased	Stasis	Declined
Advanced	196		96.9	3.1	111		84.7	15.3
Proficient	83	41.0	48.2	10.8	146	28.1	55.5	16.4
Basic	37	59.5	21.6	18.9	46	58.7	23.9	17.4
Below	9	77.8	22.2		25	52.0	48.0	

Predictive Utility

With regard to predictive utility, both subjects displayed similarities in the relative predictive power of NWEA MAP assessments. The multiple regression in each subject area found that the addition of each NWEA MAP interim assessment independently and significantly improved the predictive utility of the model. As shown in Table 42, in both subject areas the addition of the successive NWEA MAP interim assessment added relatively small increases in the power of the model as evidenced by the change in ΔR^2.

Table 42

Subject Comparison of Multiple Regression Model Summaries

	Mathematics		Reading	
Predictor Variables Included	R^2	ΔR^2	R^2	ΔR^2
School, IEP, Economic Disadvantage	.208	.208	.184	.184
6th Grade PSSA, 7th Grade EOC Grade	.745	.537	.654	.470
7th Grade Fall NWEA MAP	.790	.045	.684	.030
7th Grade Winter NWEA MAP	.814	.024	.724	.040
7th Grade Spring NWEA MAP	.836	.022	.738	.014

Note: $p < .001$.

Examination of the regression coefficients in Table 40, showed that in both subjects, when models included multiple NWEA MAP assessments, as in Model 4 (Fall and Winter) and Model 5 (Fall, Winter, and Spring), the Fall administration lost significance to predict PSSA 7. Furthermore, PSSA 6 remained a strong predictor despite the inclusion of successive NWEA MAP data. In the final model, PSSA 6 was the strongest predictor of PSSA 7 reading scores. While in mathematics, PSSA 6 was not the strongest individual predictor, it remained a strong predictor in each model.

Notable Findings

From this analysis several notable findings emerged. First, this study found mixed results regarding student growth. Second, NWEA MAP interim assessments added little practical

significance to the overall predictive utility of the model especially when models included multiple administrations of the NWEA MAP. Third, the addition of the Spring administration of NWEA MAP interim assessment resulted in a loss of significance of the Fall NWEA MAP administration. Fourth, existing student achievement data, especially PSSA 6 data remained a powerful predictor of PSSA 7 performance even when models included multiple NWEA MAP interim assessments. Fifth, student demographic information, IEP status and economic disadvantage, made no statistically significant contribution to the predictive model for PSSA 7 once student achievement had been added to the model.

Table 43

Comparison by Subject - Multiple Regression Coefficients

		Mathematics		Reading	
Model	Predictor	Beta	Sig.	Beta	Sig.
1	School	0.133	.007	0.103	.035
	IEP	-0.322	.000	-0.317	.000
	EconDis	-0.228	.000	-0.205	.000
2	School	0.123	.000	0.066	.040
	IEP	-0.046	.132	-0.122	.000
	EconDis	-0.053	.073	-0.053	.110
	PSSA 6	0.704	.000	0.519	.000
	Grade	0.196	.000	0.294	.000
3	School	0.100	.000	0.081	.009
	IEP	-0.037	.181	-0.074	.027
	EconDis	-0.015	.568	-0.036	.257
	PSSA 6	0.402	.000	0.370	.000
	Grade	0.162	.000	0.248	.000
	Fall 7	0.398	.000	0.269	.000
4	School	0.094	.000	0.076	.008
	IEP	-0.011	.686	-0.049	.122
	EconDis	0.005	.839	-0.039	.199
	PSSA 6	0.318	.000	0.296	.000
	Grade	0.147	.000	0.200	.000
	Fall 7	0.175	.002	0.108	.030
	Winter 7	0.359	.000	0.332	.000
5	School	0.089	.000	0.066	.020
	IEP	-0.003	.916	-0.056	.069
	EconDis	0.019	.428	-0.031	.294
	PSSA 6	0.263	.000	0.270	.000
	Grade	0.124	.000	0.170	.000
	Fall 7	0.062	.266	0.044	.383
	Winter 7	0.230	.000	0.228	.000
	Spring 7	0.340	.000	0.231	.000

Note: Dependent Variable: PSSA7

CHAPTER 5

Discussion and Implications

Educational researchers have noted the widespread implementation of interim assessments (Marsh et al., 2006; Stecher et al., 2008) as educational leaders sought to leverage these instruments to improve student outcomes and meet accountability demands. Perie et al. (2009) identified three purposes of interim assessments, instructional, evaluative, and predictive, each possessing an intuitive appeal to improve student outcomes and meet accountability demands. Informed by Perie's categorization, this study employed a longitudinal, quantitative analysis to investigate the student growth, both within-year and across-year, and the predictive utility of repeated administrations of NWEA MAP interim assessments in a middle-school setting. This chapter begins by answering the research questions that guided this study. The chapter continues with a discussion of the strengths and limitations of this study. Lastly, this chapter concludes with a discussion of the notable findings and implications relative to practice and future research.

Answers to Research Questions

Student Growth. The first research question that guided this study asked whether NWEA MAP interim assessment scores varied significantly over time. In this study, I found clear evidence of statistically significant growth in overall NWEA MAP means measured longitudinally across grades 6-8 in both mathematics and reading. While NWEA MAP overall group means did exhibit a positive across-year trend, the inferential and descriptive analyses in this study showed mixed results. Consistent with prior research on interim assessments, this study found both evidence of statistically significant student growth (Slavin et al., 2013;

Konstantopolis et al., 2013) and evidence of no statistically significant growth (Henderson et al, 2007, 2008; Cordray et al., 2013).

The three grade levels and two subject areas in this study presented six different periods across which to evaluate within-year growth of NWEA MAP overall means. For five of these six periods, for all but grade 6 Reading, pairwise mean differences exhibited statistically significant within-year growth. In other within-year metrics, student growth exceeded NWEA MAP school norms in both subjects for grades 7 and 8 but lagged these norms in both subjects for grade 6. PSSA proficiency, as measured by the percent of students who scored Proficient or Advanced, exceeded PA statewide growth in both subject areas from PSSA 6 to PSSA 7. However, from PSSA 7 to PSSA 8 both mathematics and reading scores lagged state growth averages.

Viewing the across-year data for the evaluative purpose similarly showed mixed results, with evidence of growth and also evidence of stasis. The observed across-year growth, grades 6-8, of the group means exceeded the school growth norms from Thum and Hauser's (2015) NWEA MAP norming study in mathematics but not in reading. Conversely, the across-year trend in PSSA scores, as measured by the percent of students who scored Proficient or Advanced, increased favorably relative to PA state averages in reading but remained unchanged from PSSA 6 in mathematics.

In sum, while I found some evidence of student growth, comparison against student growth norms failed to show clear evidence of sustained growth. Ideally, within-year and across-year longitudinal growth would have shown increases in both group membership in the Advanced and Proficient performance levels marking movement from the lower categories and increases in the group means. Movement of students among performance level categories

trended positively but not as unilaterally positively as expected and a commonly occurring movement pattern was no movement at all.

Predictive Utility. The second question that guided this study asked to what extent repeated NWEA MAP interim assessments improved the utility to predict performance on the PSSA. I hypothesized that each administration of the NWEA MAP would be individually and statistically significant predictors of PSSA 7. In both mathematics and reading, this study indeed found each NWEA MAP interim assessment individually and significantly improved the predictive model. While the improvement in predictive utility registered statistical significance, the relatively minimal improvement over existing data called into question the practical significance.

The predictive model that used student demographic data and existing student achievement data, PSSA 6 and course grades, explained a surprisingly high percentage of the variation in PSSA 7, 74.5% and 65.4% in mathematics and reading respectively. The NWEA MAP mathematics assessment data explained an additional 4.5%, 2.4%, and 2.2% of variation with the addition of the Fall, Winter, and Spring administrations respectively for a combined contribution of 9.1% of additional variation explained. Similarly, the NWEA MAP reading assessments explained an additional 3.0%, 4.0%, and 1.4% of variation for the Fall, Winter, and Spring administrations respectively for a combined total of 8.4% of additional variation explained.

Variance between Subjects. The third question that guided this study investigated the variance in student growth and predictive utility of NWEA MAP interim assessments between mathematics and reading. This study found general agreement in student growth between mathematics and reading as observed in growth of overall means and mixed categorical growth

and stasis. I found similar patterns in predictive utility of NWEA MAP interim assessments in mathematics and reading. The purpose of this question was to provide a measure of validity that findings were not limited to a specific subject. While I found differences between the data for mathematics and reading, the trends were similar.

Strengths and Limitations

The longitudinal and quantitative design contributed greatly to the strength of the current study. This study sought to augment the existing research base with an analysis of the contribution of repeated administrations of interim assessments across a three year cohort in both mathematics and reading. Much of the prior research on interim assessments investigated within-year student outcomes over a single year. Several researchers have suggested that interim assessment research better fit with an across-year, DDDM model than with the shorter term, within-year, formative assessment model (Abrahms, Varier, & McMillan, 2012; Christman et al., 2009; Davidson & Frohbeiter, 2011; Shepard et al., 2012). This current study investigated within-year student outcomes through descriptive analysis and also across-year student outcomes. By following a cohort across several years in both mathematics and reading, this study analyzed not only the instructional purpose, but also the evaluative purpose of interim assessments by following longitudinal effects across years. For example, had this study followed this mathematics cohort for only 7^{th} grade, the data would have shown the growth over 7^{th} grade but failed to capture the decline that occurred over 8^{th} grade.

Another strength of the current study resulted from the block-wise defined model of multiple regression. By employing a block-wise model, this study disaggregated the individual contributions of student demographic data, existing student achievement data, and the interim assessment data. Furthermore, the block-wise model allowed for evaluation of the individual

contribution to the variance explained by the predictive model made by each successive NWEA MAP interim assessment.

This study contained several limitations. First, small sample size and the single study site used in this study limited the generalizability of the findings. For each year of this study, more than 120,000 students in 500 districts took the PSSA. The 405 student cohort used for this study represented a small percentage of test-takers. The single district source for the data may not have accurately reflected data from other districts. Additionally, the district that comprised the sample achieved at a high-level. The high achievement of the district limited the potential to generalize findings to lower achieving schools. Furthermore, the high proportion of students who had achieved at the highest performance level, Advanced, potentially constrained the ability to meaningfully interpret growth as it would not have resulted in category movement.

Second, as Henderson et al. (2007, 2008) noted, the virtual impossibility to isolate and therefore measure informed instruction limited this study. Employing a purely quantitative design, this study did not analyze whether, and in what ways, teachers used interim assessment data to inform their instructional practice. The current study compared student growth to NWEA MAP growth norms and cohort PSSA data to aggregated Pennsylvania state averages. Several studies examined through qualitative analysis how teachers used data to inform instruction (Abrams, Varier, & McMillan, 2012; Christman et al., 2009; Shepard, Davidson, & Bowman, 2011). Consistent with the iterative nature of DDDM frameworks, no significant growth can be expected simply through implementation of interim assessment, but rather must also include changes in teaching and learning informed by interim assessment data analysis. Since this study did not investigate changes in teaching and learning, this study cannot inform best practices in informed instruction.

A small number of previous studies attempted to isolate informed instruction through experimental design, in which study sites were matched pairs with one school participating in interim assessment and the matched pair not participating (Henderson, 2007, 2008; Konstantopoulos, et al. 2013). Such an experimental design would have provided a better basis for comparison of growth. However, while the presence of a matched pair, experimental design may have provided a better basis against which to measure growth, even in experimental design, the near impossibility of eliminating formative assessment practice rendered comparison with control groups virtually meaningless (Henderson, 2007, 2008).

Third, the grading practice of the study district may well have influenced the relative importance of grades as a predictor of accountability assessment outcomes. The study district employed a standards-based grading practice that emphasized mastery of standards. Bowers (2010) and other researchers identified the non-academic factors that limited the predictive validity of course grades relative to standardized assessment outcomes. The grading practices employed by the study site sought to eliminate non-academic factors from grades and instead represented only the students' demonstrated level of mastery of standards. To the degree that this standards-based grading practice succeeded, this study may have over-represented the predictive validity of grades.

Fourth, the descriptive analysis of growth contained within this study relied upon categorical designations to describe student movement. As noted by Porter, Linn, & Trumble (2005), categorical designations presented a potential threat to reliability and validity as small changes in thresholds can significantly alter categorical outcomes. Stated more simply, categorical data can mask actual growth or imply growth that did not exist.

Fifth, this study limited analysis to scores from a single interim assessment product, NWEA MAP. The study did not investigate other interim assessment products marketed by other organizations. Additionally, this study investigated NWEA MAP scores and did not consider other formative supports offered by NWEA and other educational organizations within the context of an interim assessment program.

Discussion

Perie et al. (2009) noted the "primary goal of an interim assessment designed to serve instructional purposes is to adapt instruction and curriculum to better meet student needs" (p. 15). One would expect these adaptations to instruction and curriculum to have resulted in improved student outcomes. Whether the district in the current study intended the primary purpose of the implementation of NWEA MAP interim assessments to be instructional or evaluative, neither the within-year nor the across-year data provided a clear determination that student growth occurred.

In addition to the instructional and evaluative purposes, Perie et al. (2009) noted the predictive purpose of interim assessments. Interim assessments appeal to educational leaders due to a perceived lack in actionable value of existing data sources. This study analyzed the contributions to predictive utility in a block-wise multiple regression. In successive blocks, I augmented existing student data with an increasing number of NWEA MAP interim assessment predictor variables as the models progressed. This analysis identified four important findings: NWEA MAP provided limited practical significance in the predictive model; Spring NWEA MAP eliminated the significance of Fall; existing student achievement data, especially PSSA 6, persisted as strong predictors of PSSA 7; and the addition of student achievement data eliminated the significance of demographic information to predict PSSA 7.

The model summary showed that NWEA MAP interim assessments contributed a surprisingly small increase to the overall predictive utility of the model. All three administrations of the NWEA MAP improved the percentage of variation explained in PSSA scores by only 9.1% in mathematics and 8.4% in reading. Considering that the demographic data and existing assessment data explained 74.5% of the variation in mathematics and 65.4% of the variation in reading, it is a fair question to ask if added power to explain variation justified the investment in lost instructional time for each administration of the NWEA MAP.

Examination of the regression coefficients revealed that the additional administrations of the NWEA eliminated the significance of the Fall 7 NWEA MAP as a contributor to the predictive model. In both mathematics and reading, the inclusion of additional NWEA MAP interim assessments marginally improved the overall percentage of variation explained by the model. However, the additional NWEA MAP interim assessment data rendered the predictive utility of the Fall NWEA MAP administration non-significant. Perie et al. (2009) theorized that effective formative use of interim assessment data could erode the predictive value of interim assessment when students score higher than predicted due to the effective formative use of the interim data. It would be reasonable to suggest that instructional practice informed by interim assessment and other data would exert a greater effect over time, resulting in greater growth, and theoretically reducing the utility of pre-existing data to predict performance. The NWEA MAP data seemed to support this theory as the most distant NWEA MAP, Fall 7, did not have significant predictive utility in the presence of the Winter 7 and Spring 7 assessments. Other data from this study did not conform to this theory.

Despite the distance from PSSA 7, PSSA 6 remained a strong predictor of PSSA 7 in every model. In both mathematics and reading, PSSA 6 ranked as either the first or second most

powerful predictor of student outcomes for PSSA 7 in every model. Of the assessment events included in the model, PSSA 6 and the three NWEA MAP interim assessments, PSSA 6 represented the most distant from the outcome measure, PSSA 7. As the most distant, the interval between PSSA 6 and PSSA 7 afforded the greatest opportunity for informed instruction, and therefore, more time for student growth. In contrast to the theory proffered by Perie et al. (2009), the year of informed instruction that elapsed between PSSA 6 and PSSA 7 did not erode the predictive power of PSSA 6. That the most distant predictor variable retained its predictive power relative to interim assessments that were more proximal to the PSSA 7 was a surprising finding.

The relative and persistent strength of PSSA 6 as a predictor for PSSA 7 was a particularly surprising finding especially when considered within the context of the decreased significance of the Fall administration of the NWEA MAP. Perie et al.'s (2009), observations anticipated the decline of the predictive significance of the most distant interim assessment, Fall 7, yet their observations do not explain the persistent strength of the PSSA 6. Reconsideration of the multiple regression with respect to the accountability calendar, PSSA to PSSA, would have placed Spring 6, Fall 7, and Winter 7 as the interim assessments taken between PSSA 6 and PSSA 7.

However, repeating the analysis with Spring 6, Fall 7, and Winter 7 yielded very similar results. The predictive power of PSSA 6 persisted as it remained the strongest predictor in the mathematics and second strongest in reading. These additional results call into question how the most distant predictors, PSSA 6 and the most distant NWEA MAP, differed significantly in predictive power. One possible explanation might lie in the assessments themselves. Brown and Coughlin (2007) refuted NWEA MAP's predictive validity, noting that "concurrent relationships

are adequate, but they do not provide the type of evidence necessary to support predictive judgements" (p. 8).

Course grades remained a significant factor in every model. Grades moderately correlated with PSSA 7 in mathematics ($r = .523$), even more strongly correlated in reading, ($r = .658$), and remained a significant predictor in each regression model. The medium to high correlations observed in this study conformed to Hoge and Coladarci's (1989) wide ranging correlations though somewhat underperformed those observed by Demray and Elliot (1998). Course grades, in contrast to the other academic factors included in the regression, did not represent an event but rather an amalgam of several assessment events over the course of the school year. As observed in this study, the combined power of prior year PSSA scores and course grades offer an alternative viewpoint to the contention that existing data sources lack the utility to inform instructional practice and predict performance.

IEP status and economic disadvantage significantly predicted performance on the PSSA 7 only in the absence of other student achievement data. In nearly every model that included student achievement data, IEP status and economic disadvantage were not significant contributors to the predictive model. School membership was not a factor in the predictive power of the model. DDDM researchers have noted that successful implementation of data informed instructional practice requires building characteristics including leadership and a data friendly culture (Means et al., 2010). The lack of significance of school membership suggested that implementation was consistent in both locations.

Considering the mixed evidence of student growth and the minimal improvement to the predictive model attributed to NWEA MAP interim assessments, NWEA costs must be justified. Consistent with NWEA MAP averages, the students in this study averaged approximately 60

minutes per assessment. Many of the students in this cohort tested in excess of 100 minutes per administration, therefore it may well have taken two class periods to complete a single NWEA MAP administration. Since these students took not only assessments in mathematics and reading, but also in language-use, assessment may well have consumed as many as 18 classes, two classes per assessment for each of the three subjects (mathematics, reading, and language-use), and each of the three administrations (Fall, Winter, and Spring).

The potential opportunity cost of these assessments in instructional time equated to more than 3% of the total instructional time for mathematics. Since this district taught reading and language-use during the same block, during Reading and English Language Arts (RELA), the opportunity cost for these assessments potentially consumed nearly 7% of the instructional time. The investment of instructional time in mathematics and RELA exceed the 2% maximum guideline suggested in ESSA (2015). In addition to the opportunity costs in instructional time, districts incurred financial costs for test acquisition, $13.50 per student, and for professional development that accompanied implementation (NWEA, 2015b).

Implications for Practitioners and Future Research

Practitioners

Educational leaders charged with improving student outcomes may well consider implementation of interim assessments. Based upon the findings from this study of NWEA MAP in two high-performing middle schools and the broader work on interim assessments by Perie, et al (2009), educational leaders should consider a number of factors. Among the first considerations in that decision-making process should be careful deliberation about the purpose of interim assessments under consideration. Once clear on the purpose, educational leaders need to vet existing sources of data. Existing data sources do not have opportunity costs in

instructional time and should be thoroughly leveraged for instructional, evaluative, and predictive utility before educators decide to augment these data with additional assessments.

Several researchers have noted a mismatch in intended purpose of interim assessments and the data generated, especially with intended instructional purpose. Problematic, over-reliance on multiple-choice formats in interim assessments, especially those marketed by test publishers, did not offer enough formative insight into why students did not understand (Christman et al., 2009; Shepard et al., 2011). As Black and Wiliam (2009) noted, formative assessment must provide student-level feedback to move the learner forward. Similarly, Perie et al. (2009) argued that for interim assessments to have within-year instructional value at the classroom level, they must contain questions that generate data specific to student misconceptions and these should include open-ended questions. Perie et al. (2009) further noted that few, if any, commercially available interim assessment products resembled the activities credited by formative assessment researchers for advancing student learning. Arguably, the instructional purpose of interim assessments could be better accomplished through formative assessment practice using classroom data.

For instructional and evaluative purposes, teacher acceptance of data sources matters and demands consideration. Several studies have observed that NCLB accountability assessment data did not provide educators actionable data to inform within-year instructional practice (Henderson et al., 2007, 2008; Herman & Baker, 2005; Marsh et al., 2006). Additionally, teachers did not value these NCLB accountability data (Guskey, 2007; Supovitz & Klein, 2003). However, this study suggests that these NCLB accountability assessment data may well have value as across-year evaluative and predictive instruments.

Supovitz and Klein (2003) noted that teachers highly valued classroom assessment data. While researchers noted potential formative value for classroom assessments, these assessments may be undervalued in their utility to predict performance on NCLB accountability assessments. This study showed course grades contributed significantly to the predictive model and in some cases more powerfully than repeated administrations of NWEA MAP interim assessments. Since course grades often contain non-academic components (Bowen, 2010), classroom assessments aligned to standards potentially have even greater predictive value than course grades. If educators intend to employ grades for either evaluative or predictive purposes, care should be exercised to align course grades to reflect what a student knows and can do. Further care should be exercised to not isolate grades alone but to consider NCLB accountability assessment data. Mandinach et al. (2006) warned that teachers focused only on classroom performance to the exclusion of NCLB assessment data tended to lose perspective on broader patterns aggregated to the class or grade. They also tended to disregard longitudinal patterns and quantitative analysis. Teacher "decision-making strategies often lacked systematicity, from student-to-student, class-to-class, and year-to-year and [were] unintentionally tinged with personal bias" (Mandinach et al., 2006, p. 2).

Datnow and Hubbard (2015) concluded in their review of DDDM research that teacher perceptions about data were critical to implementing change. Educational leaders must ensure that teachers possess the literacy to interpret and act upon the data. Data collected by educators who lack the ability to interpret and apply these data did not provide within-year instructional value. Despite the investment and acknowledgement of the importance of teacher assessment literacy, most classroom teachers have not been trained in how to interpret data (Mandinach et al., 2006; Herman & Gribbons, 2001; Supovitz, 2003; McMillan, 2000; Mason, 2006; Shanahan,

2005). Supovitz (2003) reported that 59% of teachers were characterized as lacking the necessary training to analyze assessment data and that 39% of administrators did not possess this skill. The deficit in classroom teachers' collective understanding of educational assessment data, especially as it pertained to analysis and interpretation, posed a barrier to widespread instructional use of data (Black & Wiliam, 1998b; Brookhart, 2011; Datnow & Hubbard, 2015; Grummer & Mandinach, 2015; Mandinach et al., 2006; Mason, 2006; Swan & Mazur, 2011). Additionally, Datnow and Hubbard (2015) found that most training in DDDM focused on how to interact with the technology in the data management system instead of how to use the data to improve instructional outcomes.

Lastly, if educators decide to implement an interim assessment program, then the implementation should take advantage of the low-stakes nature of interim assessments. Educators can and should experiment with the implementation especially with regard to the frequency of administration and the structure of the assessment. Leaders should evaluate the outcomes and continually reassess whether and to what extent interim assessments provide value beyond the acquisition costs and the opportunity costs in instructional time.

Future Research and Policy

This study added to the small body of research on interim assessments. With the wide scale adoption of interim assessments across the country, educational leaders would benefit from additional research to inform their decision making. This study did not attempt an investigation of whether, to what degree, or how classroom teachers used interim assessment data in their classrooms to inform instructional activities. While qualitative studies exist that investigated how teachers use data, a mixed method longitudinal study to investigate the instructional purpose of interim assessments would make a significant addition to educational practice.

Additionally, to further inform the question of how often to administer interim assessments while efficiently capturing predictive power, educational leaders would benefit from an experimental design in which the treatment varied the number of test administrations of interim assessments. Such a design would facilitate genuine comparison between treatment groups who had invested instructional time into interim assessments and a control group which had used the time engaged in instructional activities.

This study found that student demographic data were not significant predictors of achievement on the PSSA 7 once student achievement data were included in the model. IEP membership and economic disadvantage have typically predicted underperformance on accountability tests. The lack of significance of these factors in the presence of student achievement data represents a potential avenue for further study.

Lastly, research suggests limited predictive utility of classroom grades for students performing in the middle performance levels. In categorically defined accountability systems, these students take on a somewhat increased importance in their role as "bubble kids". Educational leaders in such a system would welcome a study of the predictive utility of grades and prior accountability assessment data to predict performance on future accountability assessments. This would be particularly welcome in lower performing schools where "bubble kids" make up a significant percentage of student populations.

Policymakers continue to demand accountability from educators for improved student outcomes. Despite limited research supporting interim assessments, many educational leaders have responded to these demands by implementing interim assessment programs. These interim assessment programs are designed to meet perceived data needs to inform instruction and predict performance on accountability assessments. It is incumbent upon policymakers to fund thorough

research on whether and to what extent interim assessments improve student outcomes. Furthermore, policymakers should investigate whether accountability assessments could be redesigned to provide more actionable data to educational leaders.

Summary

This study examined NWEA MAP interim assessments as instruments employed for instructional, predictive, and evaluative purposes. This analysis suggests that repeated administrations of the NWEA MAP interim assessments provided minimal improvements in the predictive value of existing student achievement data and therefore may not be justified based upon a predictive purpose. Viewed through an instructional or evaluative frame, the additional assessment data often replicated existing student achievement data and did not provide an overwhelming justification of student growth.

REFERENCES

Act, E. S. S. A. (2015). Pub. L. No. 114-95. In *114th Congress*.

Andersson, C. & Palm, T. (2017). The impact of formative assessment on student achievement: A study of the effects of changes to classroom practice after a comprehensive professional development programme. *Learning and Instruction, 49, 92-102.*

Andrade, H. L., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practice*, 27(2), 3-13.

Arter, J. A. (2010, May). Interim benchmark assessments: Are we getting our eggs in the right basket. In *Annual Meeting of the National Council of Measurement in Education. Denver, CO*.

Babo, G., Tienken, C. H., & Gencarelli, M. A. (2014). Interim testing, socio-economic status, and the odds of passing grade 8 state tests in New Jersey. *RMLE Online*, *38*(3), 1-9.

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science education*, *85*(5), 536-553.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5-25.

Bernhardt, V. L. (2004). Continuous improvement: It takes more than test scores. *ASCA Leadership*, November/December, 16-19.

Bidwell, A. (March 10, 2015). Opt-out movement about more than tests, advocates say. *US News.* Retrieved from http://www.usnews.com/news/articles/2015/03/10/as-students-opt-out-of-common-core-exams-some-say-movement-is-not-about-testing

Bonner, S. M. (2009). Investigating teacher use of practice tests for formative purposes. *Journal of Multi-disciplinary Evaluation*, 6(12), 125-136.

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

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5-31.

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

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

Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. Educational evaluation: new roles, new means: the 63rd yearbook of the National Society for the Study of Education, (part II), 26-50.

Bonner, S. M. (2009). Investigating teacher use of practice tests for formative purposes. *Journal of MultiDisciplinary Evaluation*, *6*(12), 125-136.

Boston, C. (2002). The concept of formative assessment. (ERIC Clearinghouse on Assessment and Evaluation No. ED470206).

Bowers, A. J. (2010). Grades and graduation: A longitudinal risk perspective to identify student dropouts. *The Journal of Educational Research*, *103*(3), 191-207.

Briggs, D. C., Ruiz-Primo, M. A., Furtak, E., Shepard, L., & Yin, Y. (2012). Meta-analytic methodology and inferences about the efficacy of formative assessment. *Educational Measurement: Issues and Practice,* 31(4), 13-17.

Brookhart, S. (2013). Comprehensive assessment systems in service of learning: Getting the balance right. In Lissitz, R. W. (Ed.), *Informing the practice of teaching using formative and interim assessments,* (165-184). Charlotte, NC: IAP Publishing.

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

Bushaw, W. J. & Calderon, V. J. (2014). Try it again, Uncle Sam. *Phi Delta Kappan,* 96(1), 9-20.

Chappuis, J. (2005). Helping students understand assessment. *Educational Leadership*, *63*(3).

Coburn, C. E., & Turner, E. O. (2011). The practice of data use: An introduction. *American Journal of Education*, *118*(2), 99-111.

Cronin, J., & Kingsbury, G. G. (2008). *Interim assessment and prediction*. Paper presented at the Technical issues in large scale assessment subgroup state collaborative on assessment and student standards council of chief state school officers.

Darling-Hammond, L., & Adamson, F. (2010). Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning.

Data Recognition Corporation, (2012). Technical report for the 2012 Pennsylvania system of school assessment. Retrieved from http://www.education.pa.gov/K-12/Assessment%20and%20Accountability/PSSA/Pages/PSSA-Technical-Reports.aspx#tab-1

Data Recognition Corporation, (2013). Technical report for the 2013 Pennsylvania system of school assessment. Retrieved from http://www.education.pa.gov/K-12/Assessment%20and%20Accountability/PSSA/Pages/PSSA-Technical-Reports.aspx#tab-1

Data Recognition Corporation, (2014). Technical report for the 2014 Pennsylvania system of school assessment. Retrieved from http://www.education.pa.gov/K-12/Assessment%20and%20Accountability/PSSA/Pages/PSSA-Technical-Reports.aspx#tab-1

Data Recognition Corporation, (2015). Technical report for the 2015 Pennsylvania system of school assessment. Retrieved from http://www.education.pa.gov/K-12/Assessment%20and%20Accountability/PSSA/Pages/PSSA-Technical-Reports.aspx#tab-1

Davis, D. R. (2007). A quality education? *Journal of philosophy and history of education*, 18.

Demaray, M. K., & Elliott, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly*, *13*, 8-24.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessments: The limited scientific evidence of the impact of formative assessments in education. *Practical Assessment Research & Evaluation*, 14(7), 1-11.

Faria, A. M., Heppen, J., Li, Y., Stachel, S., Jones, W., Sawyer, K., & Casserly, M. (2012). Charting Success: Data Use and Student Achievement in Urban Schools. *Council of the Great City Schools*.

Field, A. (2009). *Discovering statistics using SPSS*. London: Sage.

Franklin, C. A. & Snow-Gerono, J. L. (2007). Perceptions of teaching in an environment of standardized testing: Voices from the field. *The Researcher*, 21(1), 2-21.

Gaines, M. L., & Davis, M. (1990). *Accuracy of Teacher Prediction of Elementary Student Achievement*. Retrieved from https://files.eric.ed.gov/fulltext/ED320942.pdf

Goertz, M.E., Olah, L.N., & Riggan, M. (2009). From testing to teaching: The use of interim assessments in classroom instruction (*CPRE Research Report* No. RR-65). Philadelphia, PA: Consortium for Policy Research in Education.

Gewertz, C. (2017). National testing landscape continues to shift. *Education Week*, 36(21) p 1-8

Guskey, T. (2003). How classroom assessments improve learning. *Educational Leadership*, 60(5), 6-11

Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). *Using student achievement data to support instructional decision making* (NCEE 2009-4067). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practiceguides/.

Hamilton, L., Berends, M., & Stecher, B. (2005). Teachers' responses to standards-based accountability (WR-259-EDU). Santa Monica, CA: RAND.

Hamilton, L., & Stecher, B. (2004). Responding effectively to test-based accountability. *Phi Delta Kappan*, *85*(8), 578-583.

Harlen, W., & James, M. (1997) Assessment and learning: differences and relationships between formative and summative assessment. *Assessment in Education, Policy & Practice*, 4(3), 365-379.

Harris, E. A. (August, 12, 2015). 20% of New York state students opted out of standardized tests this year. *The New York Times.* Retrieved from http://www.nytimes.com

Hart, R., Casserly, M., Uzzell, R., Palacios, M., Corcoran, A., & Spurgeon, L. (2015). *Student testing in America's great city schools: An inventory and preliminary results*. Washington D.C.: Council of Great City Schools.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81-112.

Hefling, K. (January 7, 2015). Do students take too many tests? Congress to weigh question. *Associated Press.* Retrieved from http://www.pbs.org/newshour/rundown/congress-decide-testing-schools

Helwig, R., Anderson, L., & Tindal, G. (2002). Using a concept-grounded, curriculum-based measure in mathematics to predict statewide test scores for middle school students with LD. *The Journal of Special Education*, *36*(2), 102-112.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). Measuring How Benchmark Assessments Affect Student Achievement. Issues & Answers. REL 2007-No. 039. *Regional Educational Laboratory Northeast & Islands*.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2008). A Second Follow-Up Year for" Measuring How Benchmark Assessments Affect Student Achievement." REL Technical Brief. REL 2008-No. 002. *Regional Educational Laboratory Northeast & Islands*.

Heritage, M. (2007). Formative assessment: What do teachers need to know and do? *Phi Delta Kappan*, 89, 140-145.

Herman, J., & Baker, E. (2005). Making benchmark testing work. *Educational Leadership*, 63(3), 48-54.

Herman, J., Osmundson, E., Dai, Y., Ringstaff, C., Timms, M. & National Center for Research on Evaluation, Standards, & Student Testing (2011). Relationships between teacher knowledge, assessment practice and learning. *CRESST Report 809*.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research*, *59*(3), 297-313.

Hu, W. (2012, February 24). With teacher ratings set to be released, union opens campaign to discredit them. *The New York Times*, p. A22.

Ikemoto, G. S., & Marsh, J. A. (2007). Cutting through the "Data-Driven" mantra: Different conceptions of data-driven decision making. *Yearbook of the National Society for the Study of Education*, *106*(1), 105-131.

Kane, T. J., Staiger, D. O., Grissmer, D., & Ladd, H. F. (2002). Volatility in school test scores: Implications for test-based accountability systems. *Brookings papers on education policy*, (5), 235-283.

Kerr, J. C. & Lederman, J. (October 26, 2015). *This is how much time students actually spend taking standardized tests*. Retrieved from http://huffingtonpost.com/entry/standardized - testing-time

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

Konstantopoulos, S., Miller, S. R., & van der Ploeg, A. (2013). The impact of Indiana's system of interim assessments on mathematics and reading achievement. *Educational Evaluation and Policy Analysis*, *35*(4), 481-499.

Layton, L. (August 23, 2015). U.S. Schools are too focused on standardized tests, poll says. *Washington Post*. Retrieved from https://www.washingtonpost.com/local/education/us-schools-are-too-focused-on-standardized-tests-poll-finds/2015/08/22/4a954396-47b3-11e5-8e7d-9c033e6745d8_story.html

Lazarin, M. (October 2014). Testing overload in America's schools. *Center for American Progress*. Retrieved from https://cdn.americanprogress.org/wpcontent/uploads/2014/10/ LazarinOvertestingReport.pdf

Love, N. (2004). Taking data to new depths. *National Staff Development Council*, 25(4), 22-26.

Mandinach, E. B., Honey, M., & Light, D. (2006). *A theoretical framework for data-driven decision making*. Paper presented at the annual meeting of AERA, San Francisco, CA.

Marcos, C. (July, 8, 2015). House narrowly votes to renew No Child Left Behind after drama. *The Hill*. Retrieved from http://thehill.com/blogs/floor-action/house/247297-house-votes-to-renew-no-child-left-behind

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). *Making sense of data-driven decision making in education: Evidence from recent RAND research*. Santa Monica, CA: RAND Corporation.

McCaffrey, D. F., & Hamilton, L. S. (2007). Value-added assessment in practice: Lessons from the Pennsylvania value-added assessment system pilot project (Vol. 506). Rand Corporation.

McLeod, S. (2005). Data-driven teachers. Minneapolis, MN: UCEA Center for the Advanced Study of Technology Leadership in Education.

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research*, *95*(4), 203-213.

McMillan, J. H., Venable, J. C., & Varier, D. (2013). Studies of the effect of formative assessment on student achievement: So much more is needed. *Practical Assessment, Research and Evaluation*, 18(2).

Means, B., Padilla, C., & Gallagher, L. (2010). Use of Education Data at the Local Level: From Accountability to Instructional Improvement. *US Department of Education*.

Militello, M., & Heffernan, N. (2009). Which One Is Just Right? What Every Educator Should Know about Formative Assessment Systems. *International Journal of Educational Leadership Preparation*, *4*(3), n3.

Mintrop, H., & Sunderman, G. L. (2009). Predictable failure of federal sanctions-driven accountability for school improvement—and why we may retain it anyway. *Educational Researcher*, *38*(5), 353-364.

Moore, J. L., & Waltman, K. (2006, December). *Pressure to increase test scores in reaction to NCLB: An investigation of related factors*. Paper presented at the annual meeting for the American Educational Research and Evaluation Association, Center for Evaluation and Assessment: University of Iowa.

Nelson, H. (2013). Testing more, teaching less: What America's obsession with student testing costs in money and lost instructional time. Washington D. C.: American Federation of Teachers, 1-34.

Nichol, D. J., & McFarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2) 199-218.

Noble, J. P., & Sawyer, R. L. (2004). Is high school GPA better than admission test scores for predicting academic success in college? *College and University*, *79*(4), 17.

NWEA (May, 2010). A study of the alignment of the NWEA RIT scale with the Pennsylvania system of school assessments. Retrieved from www.nwea.org

NWEA (2012). Measures of academic progress: MAP basics overview. Retrieved from www.nwea.org

NWEA (2013). Proctor handbook for client-server MAP users. Retrieved from www.nwea.org

NWEA (2014). Make assessments matter: Students and educators want tests that support learning. Retrieved from www.nwea.org

NWEA (August, 2015). 2015 NWEA measures of academic progress normative data. Retrieved from www.nwea.org

NWEA (November, 2015). Measures of academic progress: Interim assessments for grades K-12. Retrieved from www.nwea.org

NWEA (February, 2016). Linking the Pennsylvania PSSA assessments to NWEA MAP tests. Retrieved from www.nwea.org

NWEA (February, 2016). MAP Reports reference for the web-based MAP system. Retrieved from www.nwea.org

NWEA (2016). How many items appear on MAP tests? Retrieved from https://legacy support.nwea.org/node/4649

Obama, B. H. (2015). An open letter to America's parents and teachers: Let's make our testing smarter. Retrieved from https://obamawhitehouse.archives.gov/blog/2015/10/26/open-letter-americas-parents-and-teachers-lets-make-our-testing-smarter

Olson, L. (2005). Benchmark assessments offer regular checkups on student achievement. *Education Week*, 25(13), 13-14.

Parke, C. S., & Lane, S. (2008). Examining alignment between state performance assessment and mathematics classroom activities. *The Journal of Educational Research*, *101*(3), 132-147.

Pennsylvania Bulletin. (2010, January). Rules and Regulations: Title 22 - Education [22 PA. Code Ch. 4], Doc.No. 10-56. Retrieved November 16, 2010 from http://www.pa bulletin.com/secure/data/vol40/40-2/56.html

Pennsylvania Department of Education (2007). PSSA and AYP fast facts. Retrieved from http://www.education.pa.gov/Documents/Data%20and%20Statistics/PSSA%20and%20AYP%20Results/2006%202007%20PSSA%20AYP/2006%2007%20PSSA%20and%20AYP%20Fast%20Facts.pdf

Pennsylvania Department of Education (2012). State Report: Status of Pennsylvania's public schools. Retrieved from http://paayp.emetric.net/StateReport#pie

Pennsylvania Department of Education (2013a). Rules and regulations: Title 22 – Education

Pennsylvania Department of Education (September, 2013b). *Educator Effectiveness Administrative Manual*. Retrieved from http://www.portal.state.pa.us/portal/server.pt/community/educator effectiveness_project/20903

Pennsylvania Department of Education (2015a). *Pennsylvania System of School Assessment*. Retrieved fromhttp://www.education.pa.gov/K12/Assessment%20and%20 Accountability/ PSSA/Pages/ default.aspx#.VoJwfPkrK00

Pennsylvania Department of Education (2015b). *Keystone Exams*. Retrieved from http://www.education.pa.gov/K-12/Assessment%20and%20Accountability/Pages/Keystone-Exams.aspx#.VoJxd_krK00

Pennsylvania Department of Education (2016a). Response to PVAAS misconceptions: District/school reporting. Retrieved from http://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PVAAS/Professional%20Development/PVAAS%20Misconceptions%20Booklet.pdf

Pennsylvania Department of education (2016b). PVAAS methodologies: Measuring growth & predicting achievement. Retrieved from http://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PVAAS/Methodology%20and%20Research%20Materials/PVAASMethodologies%20Measuring%20Growth%20Projecting%20Performance.pdf

Pennsylvania Department of Education (2017a). *School Performance Profile*. Retrieved from http://www.paschoolperformance.org/

Pennsylvania Department of Education (2017b). PSSA results. Retrieved from http://www.education.pa.gov/data-and-statistics/PSSA/Pages/default.aspx#tab-1

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

Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). The Role of Interim Assessments in a Comprehensive Assessment System: A Policy Brief. *Aspen Institute*.

Porter, A. C., Linn, R. L., & Trimble, C. S. (2005). The effects of state decisions about NCLB adequate yearly progress targets. *Educational Measurement: Issues and Practice*, *24*(4), 32-39.

Public School Code of 1949 – Omnibus Amendments, Act of June 30, 2012, P. L. 684-82 (2012).

Rivera, P. (2015) *The guide to your student's Pennsylvania student report*. Retrieved from http://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PSSA/Individual%20Student%20Reports/Report%20Guide%20English.pdf

Ruiz-Primo, M. A., & Furtak, E. M. (2004). Informal Formative Assessment of Students' Understanding of Scientific Inquiry. CSE Report 639. *Center for Research on Evaluation Standards and Student Testing CRESST*.

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

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice*, 5(1), 77-84.

SAS Institute (2016). SAS EVAAS Technical documentation of PVAAS analyses. Retrieved from http://www.education.pa.gov/Documents/K12/Assessment%20and%20Accountability/PVAAS/Methodology%20and%20Research%20Materials/SAS%20EVAAS%20Technical%20Documentation%20of%20PVAAS%20Analyses.pdf

Sawchuck, S. (May 13, 2009). Testing faces ups and downs amid recession. *Education Week*, 28(31), 16-17.

Sawyer, R. (2007). Indicators of usefulness of test scores. *Applied Measurement in Education*, *20*(3), 255-271.

Schoen, L., & Fusarelli, L. D. (2008). Innovation, NCLB and the fear factor: The challenge of leading 21st-century schools in an era of accountability. *Educational Policy*, 22, 181-203.

Scriven, M. (1967). The methodology of evaluation. En RE Stake (Ed.), AERA Monograph Series on Curriculum Evaluation N. 1. *Chicago: Rand Mc Nally*.

Shanahan, T., Hyde, K., Mann, V., & Manrique, C. (2005). Integrating curriculum guides, quarterly benchmark assessments, and professional development to improve student learning in mathematics. Paper presented at the Evaluation Summit: Evidence-Based Findings from the MSPs.

Shepard, L. A. (2005). *Formative assessment: Caveat emptor*. Paper presented at ETS Invitational Conference 2005 The Future of Assessment: Shaping Teaching and Learning. New York, (October 10-11, 2005).

Slavin, R. E., Cheung, A., Holmes, G., Madden, N. A., & Chamberlain, A. (2013). Effects of a data-driven district reform model on state assessment outcomes. *American Educational Research Journal, 50*(2), 371-396.

Stecher, B. M., Epstein, S., Hamilton, L. S., Marsh, J. A., Robyn, A, McCombs, J. S., Russell, J., & Naftel, S. (2008). *Pain and Gain: Implementing no child left behind in three states, 2004-2006.* Santa Monica, CA: RAND Corporation.

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

Success For All (2007). 4Sight reading and math benchmarks 2007-2008: Technical report for Pennsylvania. Retrieved from https://members.successforall.org/login.aspx

Success For All (2010). 4Sight benchmark assessments: Ordering policies and procedures and frequently asked questions , 2010-2011 school year.

Sunderman, G. L., Tracey, C. A., Kim, J., & Orfield, G. (2004). Listening to teacher: Classroom realities and no child left behind. Cambridge, MA: The Civil Rights Project at Harvard University.

Supovitz, J. A., & Klein, V. (2003). Mapping a course for improved student learning: How innovative schools systematically use student performance data to guide improvement. Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

Topol, B., Olson, J., Roeber, E., & Hennon, P. (2012). *Getting to higher-quality assessments: Evaluating costs, benefits, and investment strategies*. Stanford, CA: Standford University, Stanford Center for Opportunity Policy in Education.

U. S. Department of Education (2002). *Title I – Improving the Academic Achievement of the Disadvantaged*. Retrieved from https://www2.ed.gov/policy/elsec/leg/esea02/pg1.html

U. S. Department of Education (2003). No child left behind: A parent's guide. Retrieved from http://www2.ed.gov/parents/academic/involve/nclbguide/parentsguide.pdf

U. S. Department of Education (October 24, 2015). *Fact Sheet: Testing action plan*. Retrieved from http://www.ed.gov/news/press-releases/fact-sheet-testing-action-plan

Volante, L., & Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional development. *Canadian Journal of Education*, 30, 749-770.

Werner, E. (March 28, 2011). Obama says too much testing makes education boring. Associated Press. Retrieved from http://www.boston.com/news/nation/articles/2011/03/28/obama_says_standardized_tests_often_punitive/

White House (December, 2015). Every Student Succeeds Act: A progress report on elementary and secondary education. Washington D. C.: White House.

Wiliam, D. (2007). *Content then process: Teacher learning communities in the service of formative assessment.* In D. Reeves (Ed.), Ahead of the curve: The power of assessment to transform teaching and learning (pp. 182-204). Bloomington, IN: Solution Tree.

Wiliam, D., Kingsbury, G., & Wise, S. (2013). Connecting the dots: Formative, interim, and summative assessment. *Informing the practice of teaching using formative and interim assessment: A systems approach*, 1-19.

Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*, *39*(1), 1-37.

Zernike, K. (October 24, 2015). Obama administration calls for limits on testing in schools. The New York Times. Retrieved from http://www.nytimes.com/2015/10/25/us/obama-administration-calls-for-limits-on-testing-in-schools.html

Zhang, Z. & Burry-Stock, J. A. (2003) Classroom assessment practices and teachers' self-perceived assessment skills. Applied Measurement in Education, 16(4), 323-342.

VITA

David Finnerty, EdD

Education

Doctor of Education, Educational Leadership, 2018	Lehigh University, Bethlehem, PA
Master of Education - Secondary Math, 2007	Kutztown University, Kutztown, PA
Bachelor of Science - Materials Sci & Engineering, 1992	Lehigh University, Bethlehem, PA
Bachelor of Arts - Applied Science, 1992	Lehigh University, Bethlehem, PA

Experience

Quakertown Community High School – Principal	2013 - Present
William Allen High School – Assistant Principal	2011 - 2013
Schuylkill Haven Area High School – Math Teacher	2005 - 2011
Francis D. Raub Middle School – Math Teacher	2004 - 2005
Weiler Corporation, Cresco, PA	
Vice President - Northern Region	2000 - 2003
Northern Regional Sales Manager	1999 - 2000
Midwest Regional Sales Manager	1998 - 1999
District Sales Manager	1995 - 1998
Applications Engineer	1994 - 1995
Customer Service Representative	1993 - 1994
United States Army National Guard	1993 - 2001