Summer 2009

Do Benchmark Assessments Increase Student Achievement on State Standardized Tests?

Patrick Hefflin

DO BENCHMARK ASSESSMENTS INCREASE STUDENT ACHIEVEMENT ON STATE STANDARDIZED TESTS?

A Dissertation

Submitted to the School of Education

Interdisciplinary Doctoral Program for Education Leaders

Duquesne University

In partial fulfillment of the requirements for

the degree Doctor of Education

By

Patrick Hefflin, M.Ed.

August 2009

DUQUESNE UNIVERSITY
SCHOOL OF EDUCATION
INTERDISCIPLINARY DOCTORAL PROGRAM FOR EDUCATIONAL LEADERS

Dissertation

Submitted in Partial Fulfillment of the Requirements
For the Degree of Doctor of Education (Ed.D.)

Presented by:

Patrick Hefflin
M.S., Administration and Management, Duquesne University, 1998
M.S., Science Education, Duquesne University, 1990
B.A., Physics, Mansfield University, 1989

June 23, 2009

DO BENCHMARK ASSESSMENTS INCREASE STUDENT ACHIEVEMENT ON STATE STANDARDIZED TESTS?

Approval

_____, Chair
Carol Parke, Ph.D.
Associate Professor, Department of Foundations & Leadership
Duquesne University

_____, Member
Connie M. Moss, Ed.D.
Clinical Associate Professor, Director, CASTL
Duquesne University

_____, Member
Stefan Biancaniello, Ph.D.
Instructor, Reading Clinic
Duquesne University

_____, Member
John Meighan, Ed.D.
Superintendent, Kiski Area School District

Program Director
James E. Henderson, Ed.D.
Professor of Educational Leadership and
Director, Interdisciplinary Doctoral Program for Educational Leaders
Duquesne University School of Education

ABSTRACT

DO BENCHMARK ASSESSMENTS INCREASE STUDENT ACHIEVEMENT ON STATE STANDARDIZED TESTS?

By

Patrick Hefflin

August 2009

Dissertation Supervised by Carol Parke, Ph.D.

Since 4Sight Benchmark Assessments are being promoted by politicians and educational departments throughout many states, this study was needed to determine the correlation between students' scaled scores on 4Sight and state standardized tests (PSSA). This study also uncovered teachers' perceptions of 4Sight and the extent that they used 4Sight data to modify classroom instruction.

In summary this research discovered a strong correlation between 4Sight and the PSSAs. There were significant statistical results that supported a connection between the scaled scores on 4Sight Benchmark Assessments and student scaled scores on the PSSA for this rural school district. Even though the results for seventh grade were much different from eighth grade, the 4Sight Benchmark Assessments were a much stronger predictor of PSSA math scores in seventh grade compared to the eighth grade math scores. Ultimately, the results of this study support the use of both PSSA scaled scores

and 4Sight Benchmark Assessments scaled scores to determine the assessed levels of the students and to help the teachers make more informed decisions concerning classroom instruction.

Based on the results from the teacher interviews, teachers generally tended to use tools like data analysis in their classroom if they felt comfortable, confident, and had a sense of reassurance that the tool directly impacted student achievement in a positive manner. Once the teachers developed a sense of assurance when working with student data, interpreting the results, and developing classroom activities that addressed students' weak skills, teachers tended to support the use of 4Sight data analysis in their classrooms.

DEDICATION

This dissertation is dedicated to my loving daughter, Davina T. Hefflin, who is my anchor, my foundation, and my friend. You supported and believed in my abilities when no one else did. You pushed me to work on this document when I didn't have the strength to work, you encouraged me to persevere when I wanted to give up, and you gave me a reason to finish when I wanted to quit. You will always be my phenomenal woman, phenomenally.

TABLE OF CONTENTS

Page

CHAPTER 3: Methodology

CHAPTER 4: Results

CHAPTER 5: Summary, Interpretations, and Recommendations

LIST OF TABLES

Page

CHAPTER 1

INTRODUCTION

We would like to believe that every child can learn and achieve; that no child should be left behind as the result of being a casualty of our educational system. With this belief, every professional who is dedicated to the educational field is trying to find the missing pieces that will resolve the academic deficiency that too many of our children face. Administrators, teachers, and policy makers are scrambling to determine where each child is performing academically, identify the academic gaps and develop a solution to address these discrepancies. Starting in the early 1900s, the U.S. educational system began utilizing various forms of assessments to identify the academic level of each child and therefore identify the weak or missing academic skills. Today, educational mandates have resulted in the wide use of assessments as a way to identify student needs with the anticipation that this will result in increased student achievement. Unfortunately, "we are a nation obsessed with the belief that the path to school improvement is paved with better, more frequent and more intense standardized testing. The problem is that such tests, ostensibly developed to 'leave no student behind', are in fact causing major segments of our student population to be left behind because the tests cause many to give up in hopelessness – just the opposite effect from that which politicians intended" (Stiggins, 2002, p.759).

The United States' public educational system is in a state of crisis. This has been the perception of many for at least a century. Due to this view, politicians have tried repeatedly to restructure the educational system to fit the demands and needs of this country. This perception could partly be due to the global demands to produce gains in

fields involving science and math which could lead to great economic strides and possibly push this nation ahead of all other nations. This global demand of being the best is challenging nations around the world to develop young minds that will invent, produce, or develop systems that will help them lead all other nations. This race is causing politicians all over the world to look very closely at the quality of education for children. Rightfully so, it is these young minds who will either make or break every country in this world.

This great race did not begin recently. It actually started decades ago as diplomats from various countries saw the opportunity to forge ahead with inventions and business monopolies that would place their nation as the leader of many. During the 1900s, politicians and educators in the U.S. agreed that in order to gain strides in educating our children, we must first require children to attend school (Peariso, 2006). States developed compulsory attendance laws that required all children no later than the age of eight until the age of seventeen, or until the child graduates from high school (whichever comes first) to attend school. Since children were now mandated to attend school, educators were mandated to develop programs to keep the children engaged, learning and acquiring skills that would benefit the child and ultimately this country. As the world economies continued to change and much duress was being experienced all over the world, the educational system also continued its own metamorphosis. During the World War I era, the United States Armed Forces developed a standardized aptitude test that assisted in assigning duties to soldiers based on their ability and strengths (Peariso, 2006). Politicians and educators saw this as a means to determine the ability and strengths of the children. These standardized aptitude tests resulted in the schools',

districts', states', and nations' testing programs. During the late 1930s colleges began using admission tests and later used other standardized tests in the 1950s (Peariso, 2006). These published standardized and admission tests, not only determined the acceptance of those entering into higher institutions of learning, but these tests began the era of their being used as means of accountability. This form of accountability resulted in school systems being rated on how well they prepared students for higher education. During 1965, the Elementary and Secondary Education Act (ESEA) was signed into law by President Lyndon Johnson as a means to address the growing problem of reading and math deficiency in U.S. public schools. ESEA was designed to be reauthorized once every five years.

"During the 1970's and 1980's, rapid breakthroughs in technology and increased pressures from global competition caused business leaders to begin questioning the preparedness of American graduates and the rigor of the curriculum of public school systems across the United States" (Tankersley, 2007, p. 7). Politicians, business leaders, educators, as well as the rest of America, watched as lawmakers and Presidents built the foundation needed to improve public education. During the 1970's, high-stakes testing was introduced. These tests held the schools, districts, and states accountable for the achievement of students. In 1983, 'A Nation at Risk' was published by the National Commission on Excellence in Education, which required schools, districts, and states to be held even more accountable for the achievement of students (National Commission on Excellence in Education, 1983). In 1994, President Bill Clinton signed "the $300 million School-to-Work Opportunities Act followed by the more aggressive Improving America's Schools Act of 1994 (Goals 2000) legislation" (p.7) which led to the

establishment of subject content standards and uniform national curriculum standards. This assertive stance afforded other Presidents and lawmakers the justification needed to address the apparent needs of our educational system. With the introduction of mandates, laws, and regulations, school districts across this country were reintroduced to the term of accountability and states were required to implement systems that held all educational entities accountable for the academic achievement of each child. For the first time, at every level, from the national level to the state level and including the local level, everyone held a sense of responsibility to ensure the best education for all children. Continuing with the five year reauthorization of ESEA, the No Child Left Behind Act was established in 2001 under the presidency of President George W. Bush. This act mandated for all states to establish challenging academic standards that all public schools will adhere to with the goal that all children be proficient in these standards by the year 2014 (No Child Left Behind Act of 2001. Public Law 107-110, 2002).

Pre-dating the introduction of *No Child Left Behind* (Public Law 107-110, 2002), the citizens of the United States have dedicated many hours, as well as millions of dollars to improve education for all children. As a result, many saw the paradigm shift from everything being acceptable, including practices that allowed many students to fail and/or drop out of school to the inclusion of more research-based teaching strategies. All of this has led to the race of quickly assessing students in order to identify skills that need to be addressed prior to the administration of high-stakes state assessments. Due to the increased political pressures for all children to be proficient in reading and math by the year 2014, educators across the United States are mandated to improve student academic achievement (NCLB, 2003). This mandate has forced educators to deliberately seek

instructional tools that will help them quickly identify curricular inadequacies and implement research-based, good instructional practices that will help their students master skills based on the state standards.

Many believe that American education is the cornerstone of this great nation. With the concentrated efforts of educators and political leaders, we have taken a closer look at the quality and equity of the educational services that are rendered to all children. For many years, the quality of educational service was accepted under the restrictions that surrounded the neighborhood school. Today, American education has undergone major changes. Due to the direct support of politicians through dedicated federal and state funding, American education has been catapulted from the past, restructured in the present, and is being prepared for the future. Unlike many other countries, we believe that every child living in the United States has a right to a free and appropriate education. Politicians, including President George W. Bush, believe that the future of America relies on the success of our schools and the quality of the education that our children receive. President Bush has forever marked his presidency with the law, No Child Left Behind (NCLB), which focuses on "accountability for states, school districts, and schools; greater choice for parents and students, particularly those attending low-performing schools; more flexibility for States and local educational agencies (LEAs) in the use of Federal education dollars; and a stronger emphasis on reading, especially for our youngest children" (Executive Summary, 2002, p.1). President Bush described this law as helping to reshape the educational system and will forever be the "cornerstone of my administration" (p.1).

No Child Left Behind (NCLB) of 2001 was finally introduced in both houses of Congress and on January 8, 2002 was signed by President Bush into law. As a reform of the Elementary & Secondary Education Act of 1965 (ESEA), NCLB improved federal spending on educational improvements, required stronger accountability for results, especially in reading and math, and encouraged greater flexibility with more local control. This flexibility and increased local control gave school districts more power to provide the best education to all students. School districts, under NCLB, are able to implement more flexible programs and to adopt research-proven, best instructional practices. The greatest flexibility is the ability of school districts to spend up to half of their federal education funds in manners that best fit their schools (O'Neill, 2007). NCLB also empowered parents by allowing them to have other options if their child is attending consistently low performing, low achieving schools. Ultimately, NCLB emphasizes the need for schools to adopt and implement research-based instruction that strides to improve student achievement, especially in reading and math. Since school districts are more accountable for the performance of their students, it is no longer acceptable for students to fail. Educators must consider the achievement of their student population as a whole, as well as look at the achievement of their students who belong to subgroups. The subgroup populations include those students living in poverty, race, ethnicity, disability, and limited English proficiency. According to President Bush, "[t]hese reforms express my deep belief in our public schools and their mission to build the mind and character of every child, from every background, in every part of America" (Overview, 2002, p.1).

Each state, under NCLB, must demonstrate academic improvement throughout all of the school districts. Each state must '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' (O'Neill, 2007, p.1-4). During designated time, all Pennsylvania public school districts must administer the Pennsylvania System of Schools Assessment (PSSA) in reading, math, writing, and, the most recently added, science. Based on the results of these tests, school districts, along with each school, are issued report cards that rate how well their students are performing on the annual high-stakes state assessments. The assessments and annual report cards have resulted in many educators being more attentive to what is occurring in each classroom. This wake-up call has resulted in educators addressing the quality and effectiveness of the classroom instruction. Some schools, after receiving results that indicated their students were not proficient in identified standards, began to look at ways they could improve instruction based on valid research. Educators are critiquing their services and began collaborating with colleagues to deliberately improve instruction and require the teachers in their schools to be highly qualified. Since NCLB has required every school district to look within its walls to rate and address its professional conduct, many school districts are requiring their teachers and administrators to be highly qualified in their content areas, as well as, require decisions be based on reliable research. Instead of adopting the latest popular instructional practices, educators are now evaluating programs for rigor and relevance (Datnow, 2007). School districts are expecting their professional employees to be life-long learners, who read and research best practices in their content areas. Educators are

now driven to make more informed decisions that are grounded in research by credible organizations.

Identification of the Problem

By being more knowledgeable in what the research says about student achievement, one major complaint of many educators is the fairness of high-stakes assessments like the Pennsylvania System of School Assessment (PSSA) (Gulek, 2003, Linn, 2000, Nichols, Glass, & Berliner, 2006). Teachers and administrators in the Commonwealth of Pennsylvania, as in other States, complained that they receive the results from the PSSAs too late to make any substantial impact in their instruction. The PSSAs are generally administered in the spring and the results are not available until late summer or early fall. Teachers complained that the beginning of the new school year is too late to address instructional issues that may have only been an issue with the cohort from the previous year. Since each state is held accountable, just as each school district is held accountable, the Secretary of Pennsylvania Education, Dr. Gerald L. Zahorchak, endorsed and promoted the use of 4Sight Benchmark Assessments (4Sight) to assist educators in making informed instructional decisions. With this endorsement comes the permission to allow schools the option of using state funds to pay for the benchmark assessments. This reaches back and supports an important aspect of NCLB: to provide greater flexibility for schools to spend funds to support and promote student achievement. Many schools have resorted to benchmarking their students to help improve education by predicting student performance levels, along with identifying non-proficient areas. Due to the need to know where the students are according to the standards, schools have

embraced the use of benchmark assessments (Lutz-Doemlinger, 2007, Olson, 2005a, Pottieger, 2008).

4Sight Benchmark Assessments were developed and field-tested by Success For All Foundation (SFA), led by Dr. Robert Slavin, director of the Center for Research and Reform in Education at Johns Hopkins University. The 4Sight Benchmark Assessments were designed to be administered five times during the school year; the first (baseline) should be given within the first week of the new school year. The remaining four assessments should be administered at the conclusion of each grading period. Educators use these results to analyze the curricula, the instructional practices, and communicate the results with students and parents. According to SFA, if used appropriately, teachers are able to know the predicted achievement level, identify the possible educational gaps for each student, modify the instructional delivery in the classroom, and lead to increased student achievement (Success for All Foundation, 2007). The deliberate instructional adjustment should ultimately impact student achievement in a positive manner (Stiggins, 2006).

4Sight Benchmark Assessments were written according to the PSSA blueprint in order to mirror the PSSA and provide a good prediction of how well the child would perform if the PSSA were administered at that moment. Most importantly, by conducting an item analysis, teachers are able to identify specific areas in which the students are weak and by addressing these areas in the classroom, they are able to improve academic achievement. As many educators embrace benchmark assessments, there are some who feel the assessments do not make such a significant impact that warrant giving up instructional time for additional assessments (Olson, 2005b). This dialogue has led to

many questioning the validity of benchmark assessments and if the data that is acquired from these assessments is worth the instructional time that is sacrificed in order to administer them. Teachers and administrators want to know if 4Sight Benchmark Assessments will make a difference in their school, with their students. There exists a need for additional research to determine if the use of benchmark assessments (4Sight Benchmark Assessments) will help improve student achievement on standardized assessments (PSSA).

This dialogue has also opened the unfortunate realization that many teachers do not fully understand the power and usefulness of assessments. Teachers tend to view assessments as a means to assign a grade, and nothing else. Administrators know they are accountable for improving student achievement, but are not always clear what changes need to be made to make a significant impact. Some teachers and administrators do not see the power in using data to drive daily instruction. The average teacher's response correlated with what the experts say for the average classroom, "the main means of teaching was lecture, and the main assessment of performance was a set of tests that measured [the students'] recall and basic understanding of the facts taught in the course" (Sternberg, 2007, p.20). Teachers and administrators need to be retrained and educated, not only on the various forms of assessments, but the purpose, use, and power of each form of assessment. This leads to another identified problem. Once the teacher has benchmark data, what do they do with the data? "Analyzing data and acting on data are two different steps in the school improvement process (Lutz-Doemling, 2007, p. 73)."

Purpose of Study

As the room gets more crowded with the loud voices addressing the educational dilemma of 'to-test-or-not-to-test', more and more professionals are diligently trying to find the answer. At times it appears that the room is significantly divided between those who believe benchmarking student academic progress can be valuable information to a teacher as they allow student data to drive classroom instruction and those who argue that the excessive assessments do not raise academic achievement. Researchers like Joan Herman and Eva Baker state that "good benchmark testing can encourage instruction on the full depth and breadth of the standards and give students opportunities to apply their knowledge and skills in a variety of contexts and formats" (Herman & Baker, 2005, p. 49). While other researchers, like W. James Popham, believe that educators are pressured to use various assessments and unknowingly use the results inappropriately. W. James Popham states that due to the various forms of assessments, "many students are receiving educational experiences that are far less effective" (Popham, 2001, p. 1). The primary purpose of benchmark assessments is to inform teaching and learning. It is tool to be used by teachers, administrators, parents, and students on the effectiveness of the instruction and the learning. Benchmark assessments should be formative and must be embedded in the daily instruction and curriculum. It is not effective if it is imposed from outside (Datnow, 2007). It must become a central part of the school culture. Unfortunately, many teachers and administrators are not very clear what this looks like as it impacts the classroom.

The purpose of this study was to add to the current research by determining the effectiveness of the use of benchmark assessments as a tool to improve student

achievement on state standardized tests and to determine the extent to which teachers actually use 4Sight Benchmark data to modify classroom instruction. Is it really worth the instructional class time to incorporate benchmark assessments? By addressing this purpose we discovered some empirical evidence that determined whether it is worth the loss of instructional time to administer benchmark assessments and whether benchmark assessments increased student achievement. In addition, this study added to the knowledge base of helping educators have a clearer understanding of the various forms, the usefulness, and the power of assessments.

Need of Study

As States and school districts are faced with the reality of increased accountability requirements, it is imperative that we determine if current assessment practices are the most effective way to educate our children. Researchers have dedicated more time to investigate the reliability, validity, and the effects of current assessment tools (Lutz-Doemling, 2007, Potteiger, 2003, & Protheroe, N, ERC, & NAESP, 2001). Since 4Sight Benchmark Assessments are being promoted by politicians and educational departments throughout many states, we need to determine if the administration of these assessments help educators positively impact student performance on the state assessments. Teachers tend to view effective assessments as merely a way to evaluate if the student learned the required materials (Arter, dng, Black, & Wiliam, 1988, Brandt, 1998, Chappius & Chappius, 2007, Gibbs, & Simpson, 2004, Leahy, Lyon, Thompson, & Wiliam, 2005, Popham, 2003, Schaffer, Burry-Stock, Cho, Boney, & Hamilton, 2000, Stiggins, 2002, Tomlinson, 2007, & Wiliam, Lee, Harrrison, & Black, 2004). Many educators view the

main purpose of classroom assessment as a means to assign a grade and to establish a closure to the specific chapter. Many educators feel that the PSSA's and benchmark assessments are useless since they do not place an immediate grade for a specific content area. Teachers view the high-stakes assessment, also known as summative assessments, as only valuable to administrators, as a way to anticipate if the school would meet adequate yearly progress (AYP). Other than that, many teachers feel it is a waste of time since it does not help address the curriculum they are required to teach (Olson, 2005b). If this is true, are we falsely assuming that teachers are utilizing the benchmark assessments appropriately in their classrooms and does there exists adequate research showing the effectiveness of 4Sight?

Assessment is a large focus of every school district in the United States due to the mandated accountability for the academic achievement of all students (O'Neill, & Johnson, 2007). As I pondered the need for this study, I am ever more convinced that educators everywhere need to remember the lost information they may have studied in their statistics or assessment educational courses. The only way we can address academic deficiencies is by identifying the weak or missing skills and modifying what we are teaching in the classroom so it becomes meaningful and applicable to all children. This study was needed to not only add to the research to make sure there exists a high correlation between benchmark assessments (focusing on 4Sight Benchmark Assessments) and student achievement on standardized tests (PSSA); it was also needed to determine if benchmark assessments help improve student achievement. Educators need to have a form of validation that what they are doing in their classrooms and schools

can directly impact the academic achievement of all students and by utilizing authentic student data, they can greatly impact their students' lives.

World-known educator and assessment expert, Carol Ann Tomlinson, admitted that early in her educational career she viewed assessments merely as a means to assign a grade and due to her lack of understanding of the role of assessment, she ignored assessment when she could and only did it when she had to. Tomlinson reflected that she began to see "assessment as judging performance, then as informing teaching, and finally as informing learning" (Tomlinson, 2007, p.13). My vision is that this research will help other educators, including those not willing to admit it (as Carol Ann Tomlinson did), discover the power of assessment to help all children.

Research Questions

This study answered the following questions:

1. What is the nature of the relationship between the use of benchmark assessments (4Sight Benchmark Assessments) and student achievement on state standardized tests (PSSA) for a rural school district?
2. To what extent are teachers implementing and using the 4Sight Benchmark Assessments results in their classrooms?

Objectives

The objectives of this study were as follows:

1. To verify that a strong correlation exists between 4Sight Benchmark Assessments and student achievement on the PSSAs.

2. To determine if it is worth instructional time and district funds to benchmark students' academic progress through the use of 4Sight Benchmark Assessments.
3. To determine the extent of how teachers are using the student data in their classrooms.

Definition of Terms

Academic Achievement – is the acquiring of knowledge and skills that are defined and identified as useful and appropriate.

Academic Achievement Standards / Academic Content Standards – terminology used in No Child Left Behind Act 2001 to identify standards stipulated by the state and district authorities.

Adequate Yearly Progress (AYP) – is the indicator that shows if districts and schools are making yearly progress toward reaching the goal set by NCLB that all children will be proficient in reading and mathematics by the year 2014.

Alternative Assessment – applies to any and all assessments that differ from the multiple-choice, timed, one-shot approaches that characterize most standardized and many classroom assessments. (Marzano, 1993)

Assessment for Learning – occurs when teachers use the classroom assessment process and the continuous flow of information about student achievement that it provides in order to advance, not merely check on, student learning. (Stiggins, 2002)

Assessment of Learning – measures how much the students have learned, if standards were met, and if educators have done their jobs to educate the students. (Stiggins, 2002)

Authentic Assessment – conveys the idea that assessments should engage students in applying knowledge and skills in the same way they are used in the 'real world' outside of school. (Marzano, 1993)

Benchmark Assessments – are formative assessments given multiple times throughout the school year that show whether students are progressing toward achieving proficiency on state tests. (Herman & Baker, 2005)

Criterion-Referenced Grade – indicates measurement related to teaching objectives. (Zhang, 1996)

Formative Assessment – is a planned process in which assessment-elicited evidence of students' status is used by teachers to adjust their ongoing instructional procedures or used by students to adjust their current learning tactics. (Popham, 2008)

High-Stakes Assessment – also referred to as high-stakes summative assessments. Assessments that are mandated by law to judge the students' skills and knowledge and the information is used to rate the school district.

Informative Assessment – (a.k.a formative assessment) is assessment that is viewed as active learning for the teacher and the student instead of judging performance. (Tomlinson, 2008)

No Child Left Behind Act of 2001 (NCLB) - was introduced in both houses of Congress and on January 8, 2002 was signed by President Bush into law. This law was written as a reform of the Elementary & Secondary Education Act of 1965 (ESEA), which improved federal spending on educational initiatives, required stronger accountability for results, especially in reading and math, encouraged greater flexibility with more local control to enable school districts more power to provide more flexible programs and the best

education to all students, and empowered parents by allowing them to have other options if their child is attending consistently low performing/achieving schools. (No Child Left Behind Act of 2001, 2002)

Norm-Referenced Grade – indicates measurement of comparing a student's knowledge against other students. (Zhang, 1996)

Performance Assessment – is a broad term, encompassing many of the characteristics of both authentic assessment and alternative assessment. (Marzano, 1993)

Performance Standards – refer to the required level of proficiency students are expected to display when they have mastered a content standard. (Popham, 2003)

Proficiency Targets – measure whether the district and schools are making adequate annual progress toward the goal that all children will be proficient by the year 2014.

Summative Assessment – sometimes referred to as assessment of learning, typically documents how much learning has occurred at a point in time; its purpose is to measure the level of student, school, or program success. (Chappius & Chappius, 2007)

Test Preparation – teaching content that is known to be covered on a test. (Protheroe, 2008, Linn, 2000, Heubert, 1999, Duke & Richhart, 1997, Bushweller, 1997, Canner, 1992, Kilian, 1992, Smith, 1991)

Anticipated Limitations of the Study

The most obvious limitation of this study was the selected population of students involved in this study. I used one rural school district to add to the body of research conducted by other researchers involving benchmark assessments and state standardized assessments. It was anticipated that this research will one day add to the existing

research and allow for more generalized association with similar populations of students. Since this research made use of 4Sight Benchmark Assessments data that were administered and analyzed over three years, starting in 2005, it is very important to interview teachers who participated in the initial administration, analysis, and implementation of the assessments. It must be noted that the researcher of this study was the previous principal who implemented the use of 4Sight into the school, on the request of the Instructional Cabinet, who comprised of teachers and administrators. I believe that this comradery with the teachers brought forth the honest response during the interviews and provided valuable information on the effectiveness of benchmark assessments. I was inspired to conduct this study from conversations with teachers for the need to look at the available data in order to determine if benchmark assessments impacted student achievement.

CHAPTER 2

LITERATURE REVIEW

Introduction

Researchers all over the world are in agreement that high-stakes, summative assessments do not directly increase student achievement (Black, 1988; Popham, 2003; Stiggins, 2002). They could be beneficial to teachers and administrators in order to detect large-group increases or decreases. Ultimately, they serve to hold everyone accountable for the quality of education for each child. Unfortunately, "we are a nation obsessed with the belief that the path to school improvement is paved with better, more frequent and more intense standardized testing. The problem is that such tests, ostensibly developed to 'leave no student behind', are in fact causing major segments of our student population to be left behind because the tests cause many to give up in hopelessness – just the opposite effect from that which politicians intended" (Stiggins, 2002 p.759).

It is still true today as it was decades ago; teachers are not adequately trained to effectively use the various assessments. With the daily demands of educating children; "teachers rarely have the opportunity to learn how to use assessment as a teaching and learning tool" (p.762). The American Federation of Teachers (AFT), National Education Association (NEA), Council of Chief State School Officers, National Board for Professional Teaching Standards, and the National Council on Measurement in Education (NCME) developed standards in 1990 addressing teacher preparation and competence in student assessment. Disappointingly, nineteen years later, if you discuss the topic of assessment with many teachers, many describe it as tests that may occur at the end of the chapter in order to assign a grade based on the students' knowledge of the information.

(Black & Wiliam, 1988, Brandt, 1998, Chappius & Chappius, 2007, Crooks, 1988, Graham & Simpson, 2004, Leahy, & et. al., 2005, Marzano, 2006, Popham, 2001, Popham, 2003, Stiggins, 2002, Tomlinson, 2007, Wiliam, & et. al., 2004, Zhang, 1996) Many describe the dreaded high-stakes tests as having no value for their classroom. "Student achievement suffers because these once-a-year tests are incapable of providing teachers with the moment-to-moment and day-to-day information about student achievement that they need to make crucial instructional decisions" (Stiggins, 2002, p.759). When engaging teachers in discussions about the various types of assessments some will admit that they vaguely remember terms like formative and summative assessments from their one education assessment course. Some teachers are often quick to change the subject to something they feel more comfortable discussing, like classroom activities. Research supports that teachers need more training in effective use of assessments (Brookhart, 2001). Susan Brookhart states in her research that "[s]tudies have generally concluded that teachers' knowledge and skills regarding both classroom assessment and large-scale testing are limited" (p.2). Teachers are inadequately trained to effectively use student assessment data appropriately. Likewise, district and building administrators have not been adequately trained "to build assessment systems that balance standardized tests and classroom assessments" (Stiggins, 2002, p.759).

Paul Black and Dylan Wiliam (1988) state that they believe we have enough information from what is occurring in each classroom and what research states; we simply have to address the need to train all teachers in the area of effective assessment. Professional training of teachers must be carefully planned and adequately delivered. We will need to address cultural issues, misconceptions, and harmful beliefs before we can

start training teachers to think of all forms of assessment as valuable tools that can be used to help them understand the needs of their students. Richard Stiggins (2002) suggests an action plan to initiate the change that is needed to make the appropriate use of assessments be an important part of the school's culture. First and foremost, there must be a clear devotion to the professional development of assessment for learning. Teachers must be provided with a comprehensive, long-term professional development program to foster literacy in classroom assessments, allocating sufficient resources to provide them with the opportunity to learn and grow professionally. There is a need for professional development programs that address large-scale and classroom assessment for state, district, and building administrators that teach how to provide leadership in assessment. There warrants a change in professional certifications to include competence in both formative and summative assessments. And finally, require that all teachers' and administrators' preparation programs ensure that graduates are assessment literate in terms of promoting the use of assessment to document student learning (Stiggins, Arter, Chappius, & Chappius, 2006, Stiggins & Chappius, 2006).

Researchers will warn us, there is no quick fix to this problem. "[T]hese changes are hard to implement even in ideal conditions" (Wiliam, Lee, Harrison, & Black, 2004, p.49). It will require years of retraining, dedication of resources, and commitment of everyone to make this change occur. For the health of our children's educational experience, it is imperative that teachers are retrained to use assessments appropriately. Wiliam acknowledges that implementing research into the classroom is not an easy task. Wiliam and his colleagues chose two LEAs that were receptive to implementing formative assessment. The intervention was designed to build on the teachers'

professionalism. The conclusion of their research resulted in the school's performance being raised from the 25th percentile of achievement nationally into the upper half (Wiliam & et. al., 2004). This showed that if we dedicate a clear and precise plan to train and support teachers' use of formative and summative assessments, we can impact student achievement.

Summative Assessments vs. Formative Assessments

Since summative and formative assessments are greatly impacting the educational climate of every school, educators have a greater need to completely understand the benefits, strengths, effects, and the proper use of each form of assessment. According to educators like Sophie and James Chappius, summative assessments are "sometimes referred to as assessments of learning [which] typically documents how much learning has occurred at a point in time; its purpose is to measure the level of student, school, or program success" (Chappius & Chappius, 2007, p. 15). According to W. James Popham, formative assessments, also known as assessment for learning, "is a planned process in which assessment-elicited evidence of students' status is used by teachers to adjust their ongoing instructional procedures or used by students to adjust their current learning tactics" (Popham, 2008, p. 6). Educators agree there are benefits to incorporating both forms of assessments into the classroom. However, many agree that it must be done carefully with clear expectation and clear understanding of the results (Bloom, Hastings, & Madaus, 1971, Chappius & Chappius, 2007, Sternberg, 2007).

The most common form of summative assessments is the yearly standardized assessments that are mandated by state and federal law to evaluate the academic progress

of all students in the schools. Educators will agree that even though it may be difficult to quickly implement various techniques due to the reality of managing a classroom full of very different individuals; summative assessments "have an important role to play in securing public confidence in the accountability of schools" (Black & Wiliam, 1988, p.147). Unfortunately, if not used appropriately, summative assessments can cause irreversible damage. "[S]ome contend that they have exacerbated the problem by forcing increases in dropout rates and declines in graduation rates, especially among minorities…have caused as many chronic low achievers to give up in the face of what they believe to be unattainable achievement standards" (Stiggins & Chappius, 2006, p.13).

Large-scaled, high-stakes, summative assessment can provide educators with valuable information about instructional programs and services and can help educators make informed decisions about the programs' quality (Potteiger, 2008). Summative assessments can provide insight to instructional areas that may need additional attention, to those that may need minor to major reconstruction, and those programs that are ineffective. Summative assessments promote the system of accountability to ensure the education of all children, the method to evaluate instructional practices, and the professional collaboration that is required to address educational deficiencies.

Formative assessments can provide teachers with more information concerning individual student learning. Even though some researchers will readily interchange the terms formative assessment and assessment for learning (Wiliam, 2004), other researchers insists on distinguishing the difference between the two terms. Assessment for learning, according to Richard Stiggins (2002), must go a step further than formative

assessment and involve students. Formative assessment is a planned process in which assessment-elicited evidence of students' status is used by teachers to adjust their current learning tactics (Popham, 2008). Assessment for learning requires the students' use of the data from the formative assessments to evaluate their own learning and make any needed adjustments in order to improve their learning process. Assessment for learning is different from formative assessment because it requires the students to be actively participating in the analysis of their learning.

Ideally, we need to help teachers to consistently make formative assessment a part of their classroom culture and to keep the students as key participants. By actively engaging the students in the process of their learning, it "opens the assessment process and initiates students in as partners, monitoring their own levels of achievement" (Stiggins & Chappius, 2006, p.13). In order for formative assessment to work in the classroom, there needs to be a focus on a clear purpose, provide accurate and meaningful reflections on achievement, provide students with prompt feedback along with suggestions for improvements (teachers should refrain from judgmental feedback), and allow the student to be an important participant in the assessment process (Stiggins & Chappius, 2006, Tomlinson, 2007, Wiliam, Lee, Harrison, & Black, 2004).

Carol Ann Tomlinson is known world-wide for her commitment to help train teachers in the effective use of formative assessments. One of the reasons why educators are willing to listen and carefully consider what Tomlinson says, is that she openly admits her struggles with assessments early in her teaching career. She placed herself out on the ledge by openly admitting that she viewed assessments merely as a means to assign a grade and due to her lack of understanding of the role of assessments, she

ignored assessment when she could and only did it when she had to. After many years of looking at the one aspect of classroom instruction she dreaded, she began to see "assessment as judging performance, then as informing teaching, and finally as informing learning" (Tomlinson, 2007, p.13).

Carol Ann Tomlinson embraces the use of the term informative assessment because it informs the teachers and students of the direction of student learning. While others would say informative and formative assessments are the same, Carol Ann Tomlinson would say they are different because it is an assessment that is viewed as active learning for the teacher and the student instead of judging performance (Tomlinson, 2007). Carol Ann Tomlinson describes methods of gathering data from informative assessments simply as viewing "virtually all student products and interactions" (p. 10). Tomlinson gave an overview of informative assessment as:

> Understanding 1: Informative assessment isn't just about tests.
> Understanding 2: Informative assessment really isn't about the grade book.
> Understanding 3: Informative assessment isn't always formal.
> Understanding 4: Informative assessment isn't separate from the curriculum.
> Understanding 5: Informative assessment isn't about "after."
> Understanding 6: Informative assessment isn't an end in itself.
> Understanding 7: Informative assessment isn't separate from instruction.
> Understanding 8: Informative assessment isn't just about student readiness.
> Understanding 9: Informative assessment isn't just about finding weaknesses.
> Understanding 10: Informative assessment isn't just for the teacher.
>
> (Tomlinson, 2007)

Ultimately, informative assessment solidifies the need for differentiation in the classroom. "Informative assessment is not an end in itself, but the beginning of better instruction" (p. 11).

The use of both summative and formative assessments has a unique position in the classroom. When teachers are able to effectively use both forms of assessments, they will be better able to make informed decisions that can greatly impact student achievement. Even though we are mandated to monitor and measure student achievement, it was never meant to be done in isolation. The responsibility required to administer, analyze, evaluate, and communicate results must be done as a team. The administrators must work with teachers to use the assessments to make decisions concerning the types of programs needed to help students master the required skills. The teachers must work with students and parents to help them understand the direction of the child's education. Policymakers and politicians need to communicate to the American public the truth of assessment data in order to address the conditions of public education.

Test Preparation

Test preparation practices are becoming more common in classrooms as educators try to get a handle on their world filled with assessments. Over the years, test preparation practices have evolved from teachers helping students prepare for standardized tests by reminding them to get plenty of rest prior to the test, to teaching test-taking skills and providing opportunities for students to experience taking formal tests, to states releasing sample test questions that closely mirror the objectives and questions that will appear on the high-stakes tests and teachers feeling pressured to teach only those objectives (Shepard, 1989, Smith, 1991). Ideally, test preparation practices should be used by educators to help prepare the students to demonstrate their knowledge to the best of their ability by eliminating factors that could affect the results that are not directly connected

to academic achievement (Mele-McCarthy, 2007, Miyasaka, 2000). Unfortunately, Lorrie Shepard's (1989) study on test preparation showed "repeated practice or instruction geared to the format of the test rather than the content domain can increase scores without increasing achievement" (Shepard, 1989, p.17).

With the increasing pressure for students to demonstrate proficiency, test preparation programs and practices are becoming more and more popular in classrooms. Researchers warn educators to be careful integrating test preparation activities into their classroom instruction. Joan Mel-McCarthy (2007) warns that the classroom instruction must maintain "focus on student learning, not on a test performance" (p. 11). Joan Mel-McCarthy goes on to warn us that "[w]hile teaching test content may result in better test scores, it does not ensure the broad range of knowledge necessary to apply skills to new situations" (p.13). Besides being careful not to focus entirely on the high-stakes tests, teachers and administrators do need to be familiar with the design of both the test prep and the state assessment; they need to know the depth of knowledge that is required by their state, and they need to know what and how much a student needs to know in order to be considered proficient and advanced. Unfortunately, teachers tend to deliberately guide their classroom instruction to address the state standards on which the students have failed to demonstrate mastery while taking the preparation test. By doing this, educators describe test preps as merely teaching to the test, which could result in narrowing the curriculum (Muir, 2001; Olson, 2005).

Jeanne Miyasaka (2000), based on her research, identified the following five guidelines to help educators use test preparation in a manner that will promote student achievement:

Guideline #1: "Test preparation should be embedded in and focus on teaching the entire curriculum objective domain which may include state content standards and appropriate norm-referenced test objectives" (p. 7);

Guideline #2: "Test preparation practices should include a wide variety of assessment approaches, e.g., multiple-choice items, short answer items, extended response performance tasks, especially those that are included in the test. Practices should also include a variety of item formats within each assessment approach, e.g., different types of multiple-choice item formats" (p.10);

Guideline #3: "Test preparation should include instruction in and practice of test-taking strategies" (p.11);

Guideline #4: "Test preparation should take place throughout the year" (p.12);

Guideline #5: "Test preparation practices should help students understand the importance of doing their best on the test without feeling inappropriately pressured" (p.13).

Test preparation is essential if we want to prepare our students to demonstrate their knowledge in multiple authentic ways. It would not be in the best interest of our students if we educate them in one manner and evaluate them in another manner that they are not accustomed to experiencing. However, it is clear from research that this preparation should not detract from the duty of educators to teach the needed skills that will help our students become successful. "Understanding the critical link between test preparation and high-quality teaching may help educators refocus their efforts on finding ways to better understand the curriculum objectives and expand their repertoire of teaching strategies that truly increase student learning" (Miyasaka, 2000, p.15). Research has supported that test preparation does not need to hinder the teachers' pedagogical practices in the classroom. Teachers should balance the need to cover content and teach

test-taking skills required to demonstrate acquired knowledge on standardized tests. Once this balance is achieved, educators can minimize the danger of inflating test scores (Diamond, 2007, Koretz, 2005, Linn, 2006).

The Use and Effectiveness of Benchmark Assessments

Benchmark assessments are formative assessments given multiple times throughout the school year that show whether students are progressing toward achieving proficiency on state tests (Herman & Baker, 2005). Since the need to assess students throughout the year has become more urgent, textbook publishers and companies that specialize in developing state-specific benchmark assessments like the Success For All Foundation, have developed assessments that are aligned to state standards and provide practice for students to take high-stakes assessments. We know some assessments could be used as both formative and summative; therefore, the manner in which teachers use the data will dictate the purpose and type of assessment. If the data is used only to give a grade at the end of a chapter, then that assessment instrument is considered a summative assessment. But, if the assessments are used to communicate to the students and parents in a non-judgmental and meaningful way with the next step involving the teachers and students using the data to improve instruction and learning, then this assessment is formative. The developers and authors of benchmark assessments state clearly how their assessments are intended to be formative in nature. They stress that benchmark assessments are more effective if educators immediately analyze, share the results with students and parents, and modify classroom instruction and learning based on student data (Olson, 2005a, Success For All, 2004, 2007).

Joan L. Herman and Eva L. Baker (2005) admit that "[b]enchmark testing should be worth the time and money that schools invest in it. Well-designed benchmark tests can contribute to as well as measure student learning. But if such tests are not well designed, they can waste students' and teachers' valuable time and energy, ultimately detracting from good teaching and meaningful learning" (p.54). Lynn Olson (2005), after reviewing what experts and tests vendors claim, cautions educators on the rush to assume that all benchmark assessments are adequate indicators for how the students will perform on the summative, high-stakes tests. Lynn Olson goes on to say that many test vendors rushed in to address the demand of a lucrative market for benchmark assessments, without meeting the requirements of having a valid, quality assessment that could be used by teachers appropriately.

The Pennsylvania Department of Education and the U.S. Department of Education have approved the use of 4Sight Benchmark Assessments as well as the use of state funds to purchase the assessments. Robert Slavin from Johns Hopkins University is one of the developers of the Success For All Foundation and the main catalyst for the development and implementation of 4Sight Benchmark Assessments. "4Sight Reading and Math Benchmarks were created by the Success For All Foundation to provide a formative evaluation of student progress that predicts how a group of students would perform if the PSSA were given on the same day" (Success For All Foundation, 2007, p.3). The format was carefully mirrored and correlated statistically to the Pennsylvania System of School Assessment (PSSA) in order to provide teachers and administrators with valuable insight to what the students know at the time the assessments are administered. The assessments were designed to be administered at the beginning of the

school year (benchmark) and repeated at the conclusion of each grading quarter. The purpose of these quarterly benchmarks is to help schools and districts "use the assessment results to inform instruction and track progress toward proficiency during the course of a school year" (p.3). Even though many teachers and administrators are unsure to the validity of these assessments, it is yet unknown what impact these assessments will have on instruction and learning.

Description of 4Sight Benchmark Assessments

4Sight Benchmark Assessments were developed by the Success For All Foundation under the direction of Dr. Robert Slavin, director of the Center for Research and Reform in Education at Johns Hopkins University. The assessments were formed to "provide a formative evaluation of students progress that predicts how a group of students would perform if the PSSA were given on the same day" (Success for All, 2007, p.3). "Blueprints for specific PSSA assessments as well as released tests and Assessment Anchors were carefully studied and analyzed in order to provide a blueprint for the development of the 4Sight Reading and Math Benchmarks for Pennsylvania" (Appendix A, B; Success For All, 2007, p.3). 4Sight Benchmark Assessments were approved by States all over the U.S. as a research-based, school assessment program that could be purchased using state and federal funds. Success For All has developed benchmark assessments that correlate with practically every U.S. state standardized assessment. In the state of Pennsylvania, the correlation reported by Success For All Foundation was developed using "linear regression to provide an estimated performance of students on the state's high-stakes…assessment" (p.13). The 2007 correlation for "reading ranged

from .75 to .88 and for math ranged from .68 to .76" (Appendix C, D; Success For All Foundation, 2007, p.18). The 2008 correlation for "reading ranged from .74 to .89, and for math ranged from .86 to .91" (Appendix E, F; Success For All Foundation, 2008, p. 18). By having a high correlation between 4Sight Benchmark Assessments and the Pennsylvania System of School Assessment, educators can easily predict the performance level of each student. The benchmark assessments are comprised of both multiple choice and open-ended questions that mirror the types of questions students will be exposed to on the standardized assessments. Spokespeople of the Success for All Foundation are very clear that the 4Sight Benchmark Assessments are not to be used as a Test Prep. They advise teachers that they should not expose the students to the test questions prior to the administration of the assessments and they should not teach to the test. However, the teachers should utilize the information from the student reports and item analysis report to determine the weaknesses in the curriculum, to determine if and when the assessment anchor is covered in their classroom instruction, and to determine if the students need remedial help to master the eligible content. "The data from an assessment is only good when the assessment is used as it was designed. Use 4Sight data: To monitor student progress towards proficiency on the PSSA for their enrolled grade level. Identify strengths and weaknesses on PA standards and/or reporting categories" (Success for All, 2004, slide 3).

"Success For All Foundation does not consider the administration of the 4Sight Benchmark Assessments to be an effective intervention strategy in isolation. Instead, the Success For All Foundation recommends the administration of the benchmark assessments as one part of a more comprehensive school improvement process" (Lutz-

Doemling, 2007, p. 72). Researchers have discovered that if 4Sight is used in addition to other assessments, like Dibels Oral Reading Fluency (DORF), that are also proven to have a high correlation to state standardized assessments, educators will have the most consistent prediction of student achievement level (Shapiro, Solari, & Petscher, 2008). Even though research is limited, there are a few published researches showing the high correlation between the 4Sight Benchmark Assessments and the PSSAs (Lutz-Doemling, 2007; Potteiger, 2008; Success For All Foundation, 2006; Success For All Foundation, 2007). The existing research shows that the 4Sight Benchmark Assessments predict the achievement level for the chosen experimental student population. Christina Lutz-Doemling (2007) conducted a study to determine if there was a significant relationship between sixth grade PSSA reading and math scaled scores and the scores from 4Sight Benchmark Assessments. Christina Lutz-Doemling, during her study of sixth graders, determined a strong positive and highly significant Pearson correlation coefficient (r) in reading ($r=.77$, $p<.0005$) and math ($r=.77$, $p<.0005$) (Lutz-Doemling, 2007, p.64). Cheryl A. Potteiger (2008) conducted a similar research to determine if 4Sight Benchmark Assessments in mathematics could be used to predict PSSA math scores for students in third, fifth, and eighth grades. Cheryl Potteiger also determined there exists a significant Pearson correlation between 4Sight Benchmark Math Assessments and PSSA math scores of students in third grade ($r=.74$, $p<.05$), fifth grade ($r=.85$, $p<.05$) and eighth grade ($r=.88$, $p<.05$) (Potteiger, 2008, p.37-38). Success For All reported in 2007 their inter-form reliability, using the Pearson Correlation procedure, "ranged from .65 to .75 for reading and from .74 to .81 for math (first edition), indicating the reliability of the 4Sight Benchmarks" (Appendix G; Success For All Foundation, 2007, p.19). Success

For All updated their inter-form reliability in 2008 to "range from .69 to .78 for reading, and from .74 to .85 for math, indicating the reliability of the 4Sight Benchmarks" (Appendix H; Success For All Foundation, 2008, p.19).

However, some researchers, including those who are critics of the Success For All Foundation, believe that all research should be replicated for different populations before making such bold generalized statements of effectiveness (Pogrow, 2000a, 2000b, 2002). Unlike Success For All, both researchers Lutz-Doemling and Potteiger recognized that their studies could not make broad generalized statements since the studies were conducted with small populations of students. Stanley Pogrow has openly criticized Success For All's research tactics and has accused researchers at Success For All and Dr. Slavin of "creating the appearance of success in a way that masks failure" (Pogrow, 2002, p. 463). Even though much of Dr. Pogrow's complaint of research integrity was in response to other school-wide reform programs developed by SFA, it does tend to bring up questions if the same accusations could be attributed to 4Sight's reported success. In order to validate the present research results, student population should be larger than those chosen in the initial research conducted by Success For All before making generalized claims of the benchmark's effectiveness. There should also be more research conducted to investigate the application of 4Sight student data in the classroom. This additional research can possibly determine if it's the 4Sight Benchmark Assessments that are affecting the student performance or is it good instructional practices that are the root cause.

CHAPTER 3

METHODOLOGY

Introduction

Due to the focus and mandates to improve public education in the United States, many educators are deliberately seeking ways to help all students achieve proficiency on state standardized assessments. In order to stay in compliance with No Child Left Behind (NCLB) of 2001 and ensure that 100% of all students become proficient in math and reading by the year 2014, school districts are using various assessments to help them determine the academic needs of children prior to the administration of the high-stakes assessments. This research made use of data acquired from 4Sight Benchmark Assessments, PSSA student data, and data resulting from math teachers' and administrators' interviews in order to address the following research questions:

1. What is the nature of the relationship between the use of benchmark assessments (4Sight Benchmark Assessments) and student achievement on state standardized tests (PSSA) for a rural school district?
2. To what extent are teachers implementing and using the 4Sight Benchmark Assessments results in their classrooms?

Significance of the Study

By conducting this research, we provided additional insight to the effectiveness of benchmarking students and the possible ramifications of the additional use of assessments. This research provided additional results to determine if there exists a strong relationship between 4Sight and the PSSA for this population of students. By

adding to the current research, we may be able to one day generalize if there exists the reported high correlation for all students. This research went further to determine if the act of benchmarking provided a short term increase in student achievement or does it really increase the students' acquisition of skills. Are the teachers utilizing the results from these benchmark assessments to make informed instructional decisions to improve achievement or are they simply administering the assessments based on a directive from administrators? We need to determine if the student achievement is based on the mastery of skills that can be demonstrated in other ways over time, or are they simply performing better on standardized tests due to being exposed to similar questions.

The School Setting

This study focused on the intermediate-level grades (7 and 8) in a rural school district located in Westmoreland and Armstrong Counties in the state of Pennsylvania. This rural school district encompasses 102.5 square miles and serves a total population of 30,000 residents in nine municipalities. The district has seven elementary schools, one intermediate school, and one high school. The intermediate school contains approximately 800 students in grades seven and eight. The student demographic consists of 95% Caucasian/non-Hispanic, 4.4% African American, and .6% members of other races. Low economic sub-group population is approximately 29%, while 13% of the population consists of students with special-needs, having individualized educational plans (IEPs). The average teacher to student ratio is one to seventeen. There are 45 teachers, two counselors, and two administrators at the intermediate level.

This study required the input of the math teachers. There are seven math teachers. Two math teachers are male while the remaining five teachers are female. All teachers are white. Three teachers taught over ten years, two teachers taught between three and nine years, and the remaining two teachers taught between one and two years. One teacher was present under the supervision of Principal A, B, and C and therefore has experience teaching prior to the implementation of 4Sight Benchmark Assessments, during the full implementation of 4Sight, and during the year where the test results were not stressed by the principal. Four teachers taught under the supervision of Principal B and C, while two teachers are new to the school, having only taught under the supervision of Principal C.

The Instruments

This study utilized several instruments in order to answer the research questions: the Pennsylvania System of School Assessment (PSSA) – Mathematics (spring administration), 4Sight Benchmark Assessment – Test 3 Mathematics (spring administration), and Teacher Interviews. In order to add to the current research in hopes of assisting in validating or refuting current results, this study concentrated on student data in mathematics. The PSSA Mathematics scores consisted of data from the 2004-2008 school years. Likewise, the 4Sight Benchmark Assessments in Mathematics consisted of data during the 2005-2008 school years. In order to have a benchmark year for comparison, the PSSA scores from the 2004 school year were utilized in order to build a foundation in which the scores were compared. Teachers were interviewed in order to determine their use of student data to make informed instructional decisions.

The results from the teacher interviews provided insight of how the student data was used to drive instruction.

Pennsylvania System of School Assessment (PSSA)

Initially under federal mandates, Pennsylvania, like many other states, developed an assessment system to measure student performance in public schools. The State Board of Education was required to:

> ". . . develop or cause to be developed an evaluation procedure designed to measure objectively the adequacy and efficiency of the educational program offered by the public schools of the Commonwealth . . . The evaluation procedure shall be so constructed and developed as to provide each school district with relevant comparative data to enable directors and administrators to more readily appraise the educational performance and to effectuate without delay the strengthening of the district's educational program. Tests developed . . . shall be used for the purpose of providing a uniform evaluation of each school district . . ." (Data Recognition Corporation, 2007, p.15)

The Department of Education formed an organization to develop appropriate measures and to engage in field testing questions. The PSSA, a criterion-referenced assessment, was instituted in 1992 to assess student performance (grades 3, 5, 8, & 11) in the adopted academic standards for Reading, Writing, Speaking and Listening, and Mathematics (Pennsylvania State Board of Education, 1999). With the implementation of No Child Left Behind Act, Pennsylvania began monitoring the performance levels of all students (grades 3-8, & 11) in 2006.

The broad purpose of the PSSAs is to "provide information to teachers and schools to guide the improvement of curricula and instructional strategies to enable students to reach proficiency in the academic standards" (Data Recognition Corporation, 2006, p.17). The PSSAs were developed to consist of both multiple-choice and open-

ended response questions. The multiple-choice questions were designed to measure the broad knowledge of the content standards and the open-ended questions were designed to require students to apply problem solving and written skills to solve more complex problems. "Psychometrically, multiple-choice items are very useful and efficient tools for collecting information about a student's academic achievement. Open-ended performance tasks are less efficient in the sense that they generally generate fewer scorable points in the same amount of testing time. They do, however, provide tasks that are more realistic and that better sample higher-level skills" (Data Recognition Corporation, 2006, p.20-21).

The PSSA-Mathematics assesses students in five reporting categories: Numbers and Operations, Measurement, Geometry, Algebraic Concepts, and Statistics and Probability. Based on the multiple-choice and open-response questions, students can obtain a performance level of Below Basic (inadequate academic performance), Basic (marginal academic performance), Proficient (satisfactory academic performance), or Advanced (superior academic performance).

The Validity of PSSA

The validity of any assessment is based on whether the assessment accurately measures the information it is intended to measure. The validity of the PSSA is evidenced in the content validity, convergent validity (relationship between student's performance on two tests) and discriminant validity studies conducted by Human Resources Research Organization (HumRRO), which included an extensive evaluation of test items and of statistical relationships of the PSSA, including convergent and

discriminant validity (Sinclair, Thacker, & HumRRO, 2005, Thacker, Dickinson, & Koger, 2004). This study documented the high correlation of .7 and .9 between the PSSAs and other comparison tests (e.g., GPA, CTBS/Terra Nova, CAT-5, SAT-9, Northwest Evaluation Association's Achievement Tests, and New Standards Reference Exam) (Sinclair, Thacker, & HumRRO, 2005, Thacker, Dickinson, & Koger, 2004).

The Reliability of PSSA

The reliability of a test is based on the consistency of obtaining the same results when taken by different students in other settings. The Cronbach's Alpha reliability indices were calculated and reported by Data Recognition Corporation using the traditional formula, the ratio of true score variance to total score variance, and the result existed between .92 and .93 (Appendix J, K, L; Data Recognition Corporation, 2007a, 2007b, 2008). The more reliable the test is the closer the calculation will compute to one.

4Sight Benchmark Assessments

4Sight Benchmark Assessments in mathematics and reading were created by Success For All Foundation for the main purpose of providing a "formative evaluation of student progress that predicts how a group of students would perform if the PSSA were given on the same day" (Success For All Foundation, 2007, p.3). In order for educators to maximize the use of 4Sight, Success For All Foundation and Pennsylvania Training & Technical Assistance Network (PaTTAN), an organization developed by the Pennsylvania Department of Education to work with local educational agencies to improve student achievement, provided training on the administration and analysis of

benchmark assessments. 4Sight Benchmark Assessments are designed to be administered at most five times per year. The first assessment, usually administered the first few weeks of the new school year, serves as a baseline for student achievement. The remaining four assessments could be administered at the end of each report quarter, or at least every nine weeks. Educators should use "the assessment results to inform instruction and track progress toward proficiency during the course of a school year" (p.3).

The 4Sight Benchmark Assessments are carefully designed based on the blueprints used to develop the PSSAs. "Blueprints for specific PSSA assessments as well as released tests and Assessment Anchors were carefully studied and analyzed in order to provide a blueprint or the development of the 4Sight Reading and Math Benchmarks for Pennsylvania" (p.3). Based on this blueprint, the standards addressed on all 4Sight Benchmark Assessments have the same weight as the PSSA for each grade level. In keeping with the main purpose of 4Sight, which is to mirror the PSSA, if the weighting or focus of the PSSA change; 4Sight Benchmark Assessments would change as well. "4Sight Benchmarks mapped the specific type of item, an item description, the item stem, the state standard and Assessment Anchor to which the item was tied, and the number of items of each type needed to mirror the proportion of these items on the state assessment" (p.5). The scoring of the 4Sight Benchmark Assessments, especially the open-response questions, utilizes the Pennsylvania rubrics or scoring guides for the PSSA.

The Validity of 4Sight

Since 4Sight Benchmark Assessments were designed to mirror the PSSA, careful attention was given to provide evidence of the test content and internal structures to the PSSA during the development of the blueprint used to derive the test. The 4Sight test developers measured content validity and criterion validity by developing a correlation between pre-pilot and pilot student test scores and comparing those scores with the students' PSSA scores. The resulting math correlation ranged from .86 to .91 (Success For All Foundation, 2008).

The Reliability of 4Sight

The reliability of a test is based on the consistency of obtaining the same results when the same test is administered multiple times. Test developers for 4Sight used the Pearson Correlation procedure to determine the inter-form reliability. Those results reflected a range from .74 to .81 for the math assessments. This range indicates high reliability. In addition to the calculation of the inter-form reliability, Success For All Foundation also computed the inter-rater reliability. This computation is vital since there are multiple test items that require the scoring by educators. In order to increase the inter-rater reliability, Success For All, with the assistance of PaTTAN, provided training on the proper way to score the open-ended response questions using the scoring guides blueprinted after the PSSA scoring rubric. Additionally, Success For All Foundation requires multiple individuals to score the open-ended responses. The scores from the 4Sight Benchmark Assessments were correlated to the PSSA scaled scores. Success For

All Foundation calculated the inter-rater reliability to measure .74 to .85 for math (Success For All Foundation, 2007).

The Success For All Foundation is dedicated to providing benchmark assessments that can accurately provide student data that educators can have confidence in using to make informed instructional decisions to help all students achieve. "As the PSSA undergoes additional improvements, these changes will also be reflected in revised versions of the 4Sight Reading and Math Benchmarks for Pennsylvania and additional data will be collected from the schools to continue to provide correlated estimates of student performance on the PSSA, as well as continue to ensure the validity and reliability of the 4Sight Reading and Math Benchmarks" (Success For All Foundation, 2008, p.20).

Interviews

In order to gather more information on the use of student data, how it impacted daily instruction, the level of confidence working with student data and provide a description of the school culture, individual interviews were conducted with the professional employees. The interviews were conducted privately in order to ensure confidentiality by the researcher. The interviews were conducted with the anticipation that more insight can be obtained on how student data is used daily, how the professional employees used the data, whether the professional employees feel empowered by the student data and whether the culture of the school supported the use and implementation of benchmark assessment data.

The teachers interviewed were asked the following research questions:

Category I - Teacher's Background Information:

1. How many years have you been a math teacher?
2. How many years have you taught math at this school?
3. Were you present under the supervision of Principal A?
4. Were you present under the supervision of Principal B?

Category II - Teacher's Perception on Professional Development Activities:

5. Do you feel you were adequately trained to analyze 4Sight Benchmark Assessment data?
6. Do you feel you were adequately trained to use 4Sight Benchmark Assessment data to modify your classroom instructions?
7. Is on-going training provided to you to help you adequately analyze and implement 4Sight Benchmark data?

Category III - Teacher's Confidence Working With 4Sight Benchmark Assessments:

8. What is your confidence level analyzing and implementing 4Sight Benchmark data?
9. Do you feel confident sharing student results with students and parents?

Category IV - Teacher's Use of 4Sight:

10. Do you believe your use of 4Sight Benchmark data help improve your students' achievement?
11. Describe classroom activities used in classrooms that you developed which were directly based on 4Sight Benchmark Assessment data?
12. How frequently are these activities used in your classrooms?

13. Do you use any other activities that were developed by other teachers that resulted from their analysis of 4Sight Benchmark Assessment data?
    a. If yes, please describe the activities and the frequency of use.
14. Does benchmarking students impact your overall use of assessments?
    a. If yes, how?
    b. If no, why not?
15. Has benchmarking students change the way you view the usefulness of assessments?
    a. If yes, how?
    b. If no, why not?
16. Does benchmarking students provide you with greater empowerment to make more informed decisions that drive your instruction?
    a. If yes, how?
    b. If no, why not?
17. How frequently do you meet with other professionals to analyze benchmark assessment data?
    a. If meetings occur, who participates in these meetings?
18. How do you view benchmark assessment data (electronically, principal-provided results, or other)?
19. How frequently do you meet with other professionals to discuss the results of benchmark assessment data?
    a. If meetings occur, who participates in these meetings?

20. Do you share the data with students?
    a. If yes, describe how data is shared, when is data shared, and the how frequent if the data shared.
21. Do you share the data with parents?
    a. If yes, describe how data is shared, when is data shared, and the how frequent if the data shared.

Category V - Teacher's Perception of the Effectiveness of 4Sight:

22. Does benchmarking students impact your classroom instruction?
23. What was your opinion of the effectiveness of 4Sight Benchmark Assessments when it was first implemented in 2005?
24. What is your opinion today on the effectiveness of 4Sight Benchmark Assessments?
25. Describe how your opinion of the effectiveness of benchmarking students has changed over time.
26. Do you believe 4Sight Benchmark Assessments impact student achievement?
    a. If yes, how?
    b. If no, why not?

Category VI – Teacher's Overall Opinion of the Use of 4Sight Benchmark Assessments:

27. Would you recommend the continued use of 4Sight Benchmark Assessments?
    a. If yes, why?
    b. If no, why not?

Research Design and Participants

The use of 4Sight Benchmark assessments was implemented school-wide in year 2005. Due to this implementation, there was data from multiple years that was beneficial to analyze in order to determine the relationship between benchmarking students and whether there existed any increase in student achievement on the PSSA. I utilized 2003-2004 as the baseline year. The school was under the supervision of Principal A who retired at the end of this baseline year. Principal B took over the supervision of the school during 2004-2005 school year. 4Sight Benchmark was not used during these two years. However, the school experienced growth in academic achievement and made adequate yearly progress during the 2004-2005 school year. During the years 2005 through 2007 4Sight Benchmark Assessments were implemented under the supervision of Principal B. Principal B was promoted to central office and Principal C took over the supervision of the school. This data was significant because Principal B embraced the full implementation of 4Sight Benchmark Assessments and Principal C admitted to administering the benchmark assessments since it was mandated but did not make any use of the data during the entire year.

In order to answer the first research question, this study had two parts: Part One exclusively involved all math PSSA student data and Part Two involved the student math PSSA data, 4Sight Math Benchmark data, and information collected from teacher interviews. The information collected from the teachers was used to answer the second research question. The math scores were used in order for this study to add to the collection of research by providing additional insight that could, one day, lead to generalizing the effects of 4Sight Benchmark Assessments.

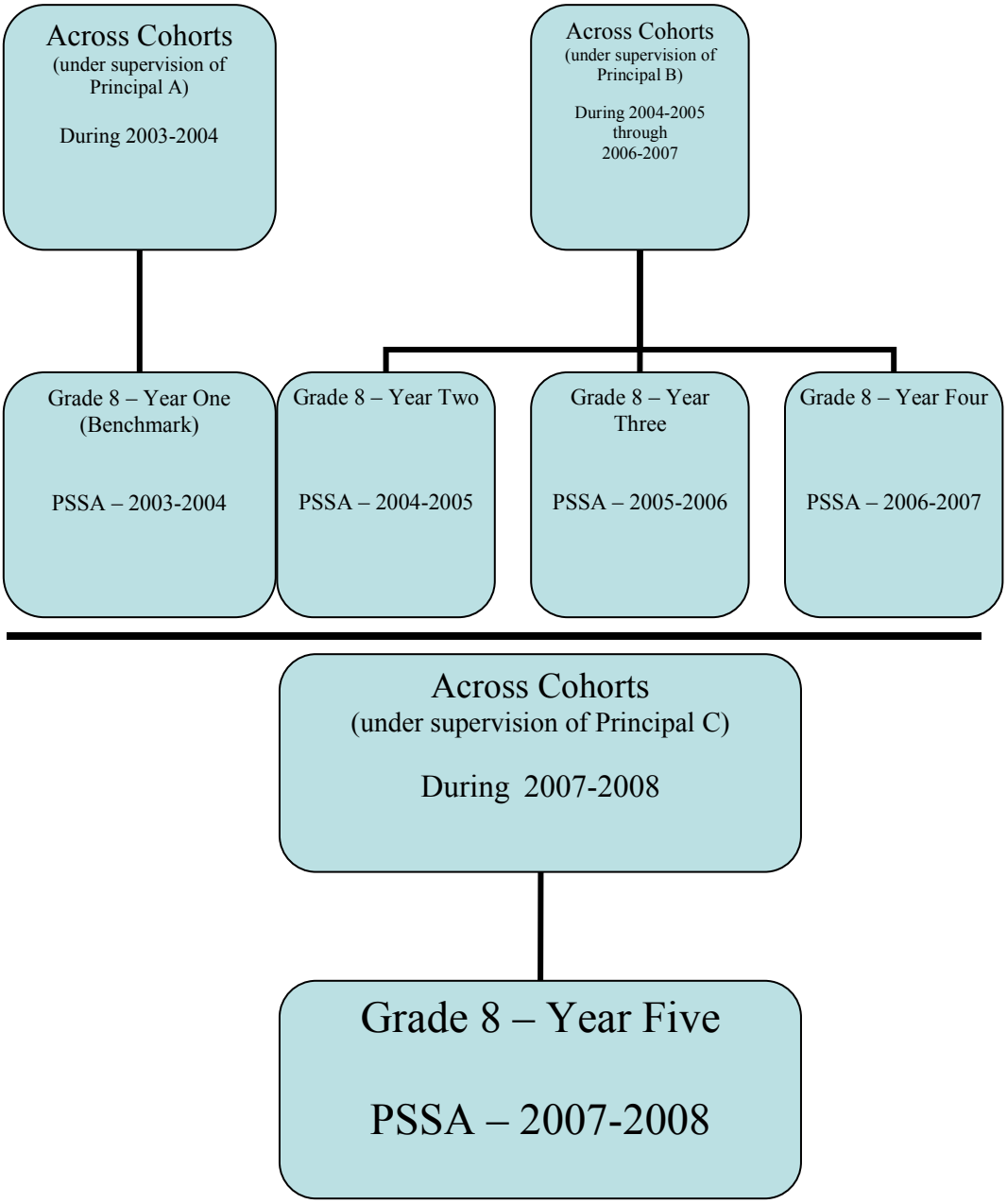
Part One of this study concentrated on the eighth grade student PSSA data from the benchmark year (2003-2004) through the fifth year (2007-2008). This portion of the study helped determine the rate of academic achievement, as scored on the Math PSSA, of eighth graders from 2004 through 2008. Since each year represented a different cohort of students, I titled this portion of the study to involve across cohorts. This information was vital because it provided insight to the achievement of eighth grade students according to the PSSA and the curriculum that was delivered by the eighth grade teachers.

Part Two of this study concentrated on two cohorts of students: Cohort #1 tracked the achievement of seventh grade students during the 2005-2006 school year to their eighth grade year during the 2006-2007 school year. Cohort #2 tracked the achievement of the seventh grade students during their 2006-2007 school year to their eighth grade year during the 2007-2008 school year. Since 4Sight Benchmark Assessments were administered during these years, 4Sight Math Benchmark Assessments scores were used in addition to the Math PSSA student scores and information collected from teacher interviews.

The data collected during Part One and Part Two of this study allowed the questions of this research to be answered: what is the nature of the relationship between the use of benchmark assessments (4Sight Benchmark Assessments) and student achievement on state standardized tests (PSSA) for a rural school district and to what extent are teachers implementing and using the 4Sight Benchmark Assessments results in their classrooms?

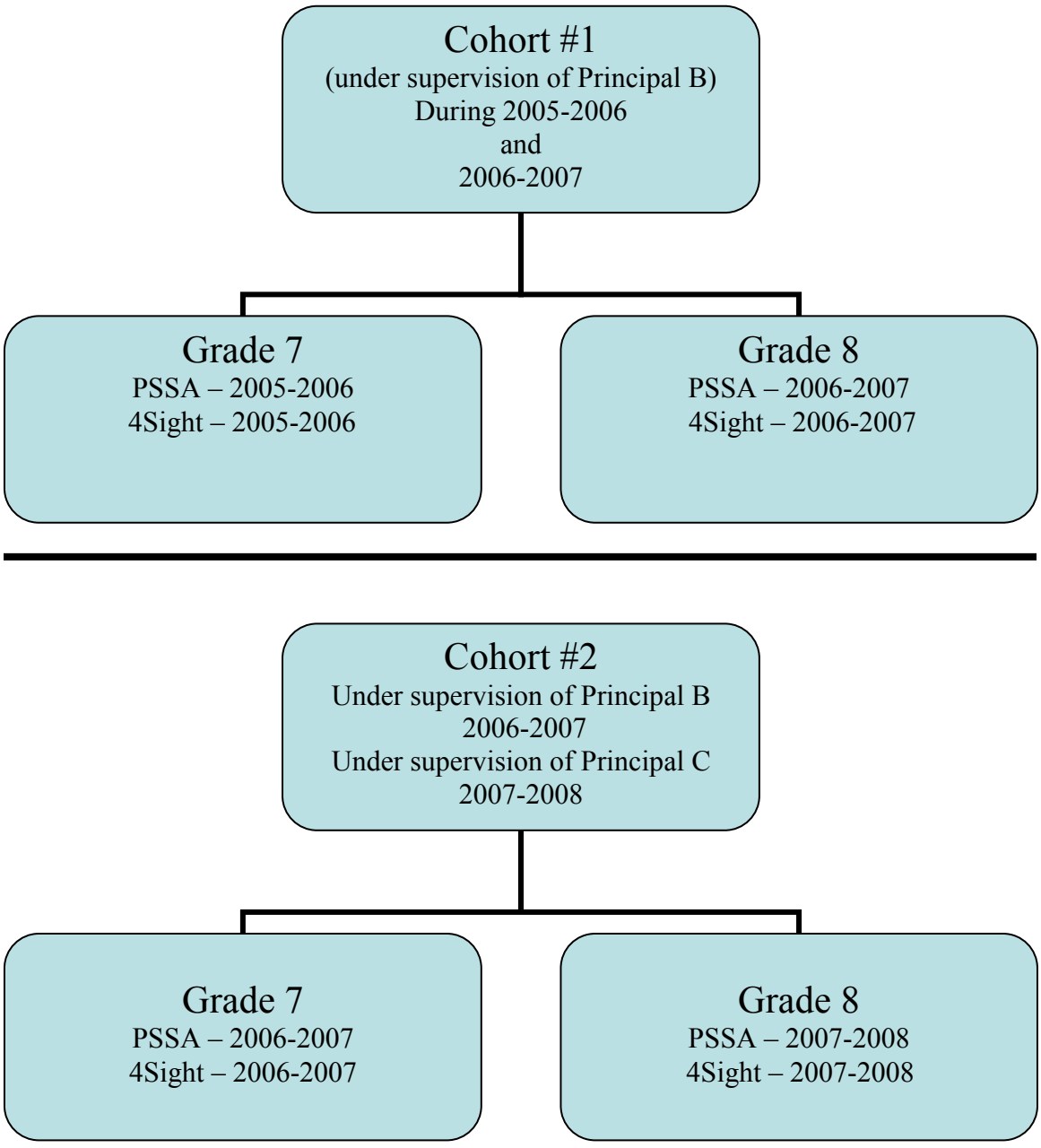
Part One – Across Cohorts

Figure 1. Across Cohorts – The involvement of eighth grade students across cohorts:



Part Two – Within Cohorts

Figure 2. Within Cohorts – The involvement of students from seventh grade to eighth grade:



Procedures

In accordance to the policies and procedures of Duquesne University Institutional Review Board, all required documents were obtained and regulations were followed to maintain that all student assessment data were de-identified and all teacher interview data were kept confidential. The following procedures were followed in order to collect all necessary data, the preparation of all student assessment data to ensure the anonymity, the confidentiality of teachers' interview data, and the analysis of all data to ensure proper research procedures:

Step 1: In order to maintain the integrity of this study, all students were assigned a random, unidentifiable number by a paid third party.

Step 2: Students' PSSA math data for the 2003-2004 school year through the 2007-2008 school year and the students' 4Sight Math Benchmark data for the 2005-2006 school year through the 2007-2008 school year were placed in an excel file by a paid third party in order to ensure the integrity of the study. The required information consisted of students' math PSSA scaled score, 4Sight math scaled score, 4Sight math correlated score to the PSSA, and their math PSSA and 4Sight performance levels. Once the data was accurately documented for each student and aligned with their classroom teacher, the paid third party replaced all names with the random, unidentifiable number assigned in step 1.

Step 3: Data was collected during teacher interviews in order to determine the

use of 4Sight Math Benchmark Assessment student data in their classrooms. All interviews were conducted by the researcher. The paid third party transcribed all interviews.

The student achievement data from the PSSA and the 4Sight Benchmark Assessments will remain secure at all times. The students' names were replaced with random numbers, assigned by a third party, in order for the researcher to analyze the correlation between students' PSSA math scores and their 4Sight Math Benchmark Assessment scores. The true identity of students will not appear anywhere in the data file or this dissertation. All data will be maintained according to requirements of Duquesne University for a minimum of five years at the completion of this dissertation.

The identity of the professional staff will remain confidential. In order to maintain the integrity of participants' identity, all participants will be assigned random numbers. All surveys and anecdotal notes taken during interviews will be maintained for a minimum of five years at the completion of this dissertation, according to the requirements of Duquesne University.

Data Analysis

The results from the data included both quantitative analysis and qualitative analysis. In order to answer the first research question, [What is the nature of the relationship between the use of benchmark assessments (4Sight Benchmark Assessments) and student achievement on state standardized tests (PSSA) for a rural school district?], this study required descriptive analysis where the means and standard deviations were

utilized. Inferential Statistics were used in order to analyze the achievement of students assigned to part one – Across Cohorts. One-factor ANOVA, where the independent variable is the five school years and the dependent variable is the eighth grade math PSSA scaled scores. The Null Hypothesis is that there is no significant difference in the mean scaled scores over time. If the Null Hypothesis is rejected, then a follow-up analysis, such as Tukey, was used to determine which years resulted in this significant difference. This allowed for determining if a trend existed across years. In order to address part two – Within Cohorts, inferential statistics were used by determining the correlation between the math PSSA students' scaled scores and the 4Sight Math Benchmark Assessments students' scores. Multiple regression analysis was used to determine if the 4Sight Math Benchmark students' scores significantly predict the math PSSA students' scaled scores. If it is discovered that the 4Sight Math Benchmark students' scores significantly predict the math PSSA students' scaled scores, then the variables contributing the most to predict the math PSSA students' scaled score will be identified.

In order to answer the second research question, [To what extent are teachers implementing and using the 4Sight Benchmark Assessments results in their classrooms?], qualitative analysis was used to organize and report the results from teachers' interviews. The responses were carefully organized and analyzed to determine to what extent the teachers were using the 4Sight Math Benchmark data in their classrooms and to determine if it is worth instructional time and district funds to benchmark students' academic progress through the use of 4Sight Benchmark Assessments. The researcher categorized the participants' responses to each question and conducted a contact analysis

of the interviews in order to summarize their responses, perceptions, and use of the benchmark assessments. In order to describe the teachers' responses, a rubric was used to rate and organize the qualitative data (Appendix M).

Summary

As schools are held more accountable for the academic achievement of all students, it is important for educators to know the instructional needs of the students. The strengths and weaknesses of the instruction are not always apparent and teachers are looking for additional ways to know what the students have already mastered and the areas that need to be addressed. Based on this need, providers of benchmark assessments are promising a quick analysis of student achievement so educators can address areas of deficiency prior to the administration of high-stakes state assessments. It is important to research and analyze the effectiveness of benchmark assessments. Even though there are limited studies on 4Sight Benchmark Assessments, the current studies tend to concentrate on students attending large urban school districts. It is important to test the effectiveness of 4Sight Benchmark Assessments in a rural school district.

CHAPTER 4

RESULTS

Introduction

The purpose of this research was to determine the effectiveness of the use of benchmark assessments as a tool to improve student achievement on state standardized tests and to determine the extent to which teachers actually use 4Sight Benchmark data to modify classroom instruction. With the use of this research, I wanted to determine if it was really worth the instructional time to incorporate benchmark assessments and whether benchmark assessments increased student achievement. This chapter culminates the answers to the two research questions that guided this study:

1. What is the nature of the relationship between the use of benchmark assessments (4Sight Benchmark Assessments) and student achievement on state standardized tests (PSSA) for a rural school district?
2. To what extent are teachers implementing and using the 4Sight Benchmark Assessments results in their classrooms?

Descriptive Statistics

Description of Sample

This research utilized student data and teacher interviews from a rural school district located in Westmoreland and Armstrong Counties in Pennsylvania. This study focused on the intermediate-level, grades seven and eight. The demographic identifiers were consistent throughout the years studied. Over the duration of five years used to conduct this study, from the school year 2003-2004 through the school year 2007-2008,

the percentage of students according to gender remained consistent with a mean of 52.5% male and 47.5% female (Table 4.1).

Table 4.1: Percent of Students According to Gender

YEAR	N	% Male	% Female
2004	404	50.2	49.8
2005	388	53.9	46.1
2006	367	52.6	47.4
2007	362	52.5	47.5
2008	333	53.2	46.8

The percentage of students who were identified as economically disadvantaged (SES) were also consistent with the exception of those reported during 2005, where it was reported that only 2.1% of the student population (n=388) was economically disadvantaged. This year may not adequately reflect the low economic population due to several possible factors. This was the first year under the supervision of Principal B and this was the first year the student demographic information was collected and collated electronically, rather than this information being bubbled directly on the test, as done in the past. The third possible explanation for the discrepancy could be accounted in the inaccurate reporting of student information within the electronic data received by the district. The remaining percentage of SES students is consistent with a mean of 24.9% (Table 4.2).

Table 4.2: Percent of Students According to Economic Disadvantaged (SES)

Year	n	%SES
2004	404	22.3
2005	388	2.1
2006	367	25.9
2007	362	22.4
2008	333	28.8

In addition to the consistency in gender and SES, the ethnicity variance is also consistent with the mean of 95% of students being white and 5% of students consisting of other ethnic groups like African American, Asian, Hispanic, and Native American (Table 4.3).

Table 4.3: Percent of Students According to Ethnicity

YEAR	N	%White	%Other
2004	404	95.5	4.5
2005	388	94.6	5.4
2006	367	94.0	6.0
2007	362	95.3	4.7
2008	333	95.5	4.5

The teachers (n=7) who are responsible to teach mathematics to the seventh and eighth grade students consisted of five female and two male teachers. All teachers are white and three teachers teach only the seventh graders (Teachers A, B, & G), while two teachers teach only the eighth graders (Teachers C & D), and the remaining two teachers teach both seventh and eighth graders (Teachers E & F). Veteran teachers, those who have taught for over ten years, accounted for three out of seven, while those who have tenured, yet are considered fairly new to the educational field (teaching three to nine years), account for two out of the seven teachers. The remaining two teachers are not tenured, having taught less than three years. One teacher was present under the supervision of Principal A, B, and C, four teachers taught under the supervision of Principal B and C, and two teachers taught only under the supervision of Principal C. This is important since 4Sight Benchmark Assessments were not used by Principal A, were fully implemented by Principal B, and were administered but data was not used by Principal C.

Results

Research Question 1: *What is the nature of the relationship between the use of benchmark assessments (4Sight Benchmark Assessments) and student achievement on the state standardized tests (PSSA) for a rural school district?*

In order to answer this research question, this study was divided into two parts: Part One (Across Cohorts) exclusively involved all eighth grade PSSA student math data from the benchmark year, 2003-2004, through year five, 2007-2008, and Part Two

(Within Cohorts) involved students' PSSA math data, 4Sight Benchmark data, and information collected from teacher interviews. Part Two (Within Cohorts) consisted of two cohorts. Cohort #1 was defined by the students' seventh grade year during 2005-2006 through their eighth grade year during 2006-2007. Cohort #2 was defined by the students' seventh grade year during 2006-2007 through their eighth grade year during 2007-2008.

Part One: Across Cohorts Analysis for Research Question #1

In order to determine the rate of academic achievement on the math PSSA for the eighth grade students in part one, I organized my analysis by first taking a look at the eighth grade student PSSA data from the benchmark year (2003-2004) through the fifth year (2007-2008). The descriptive statistics for the eighth grade PSSA revealed that the average scaled scores increased 1.85 during the 2004 to the 2005 school year. This was during the transition between Principal A, who retired at the conclusion of the 2003-2004 school year, and the new supervision of Principal B into the intermediate school during the 2004-2005 school year. During both years, 4Sight Benchmark Assessments were not administered. However, when comparing the 2005 PSSA math data to the 2006 PSSA math data, it is revealed that the mean scaled score increased by 30.80. This was during the first year of 4Sight Benchmark implementation at the school. Continuing the comparison between 2006 PSSA math scaled scores and the 2007 PSSA math scaled scores, the mean scaled scores decreased by 25.29. Finally comparing the 2007 PSSA math scaled scores to 2008 PSSA math scaled scores revealed an increase of 44.16. During the 2007-2008 school year the school was under the supervision of Principal C.

Another interesting result was that the standard deviation decreased each year, indicating that the students' PSSA scores were more tightly distributed around the mean and that student achievement consistently became more homogeneous (Table 4.4).

Table 4.4: 8th Grade PSSA Mathematics Means and Standard Deviations

Year	N	Mean (M)	Standard Deviation (SD)
2004	404	1369.10	308.510
2005	388	1370.95	212.851
2006	367	1401.75	209.174
2007	362	1376.46	205.585
2008	333	1420.62	199.978

The purpose of examining PSSA scaled scores across eighth grade cohorts was to determine if the average score was significantly increased over time. Therefore, a one-way ANOVA was used. The null hypothesis was that the means of the scaled scores for students were equal across the five years. As shown by the data, there is significant difference between the eighth grade PSSA Math scaled scores over time (p=.01). Therefore the Null Hypothesis was rejected and additional analysis was used to determine which years resulted in this significant difference (Table 4.5).

Table 4.5: ANOVA Results

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Year	725706.120	4	181426.530	3.340	.010
Error	100,400,000	1849	54320.614		
Total	101,200,000	1853			

Because the assumption of homogeneity of variance was not satisfied for the ANOVA analysis, $p<.001$, alternative statistical tests were conducted to determine whether they produced results that were similar to the ANOVA. A modified version of ANOVA, robust tests of equality of means, revealed a Welsh p-value of .004 and Brown-Forsythe p-value of .009. These results were consistent with ANOVA since they both reject the null hypothesis of equal means, in fact the two modified test do so at a more stringent significant level. Both values indicated more significance between the scaled scores over time. Both values reflected the statistical test for the equality of group variances, measuring the statistical spread of scores.

As indicated earlier, the Post Hoc results reflect a comparison of each mean scaled score with the mean scaled scores for other years. I used the Games-Howell post hoc test because it accounts for the violation of the assumption of homogeneity of variance, thus it is a more valid test on this set of data than the more typical Tukey post hoc test. There was not a significant increase in the mean scaled scores when comparing 2003-2004 scores with 2004-2005 scores. However, when using 2003-2004 as a baseline

year, comparing the mean scaled scores with subsequent years, there were moderate significant difference between the benchmark year and the 2005-2006 mean scaled scores and higher significant difference during the 2007-2008 mean scaled scores (Table 4.6).

Table 4.6: Post Hoc Results

	2004	2005	2006	2007	2008
2004	------	1.85 (1.000)	32.65 (.414)	7.36 (.995)	51.53 (0.050)
2005	------	------	30.80 (.265)	5.51 (.996)	49.67 (.011)
2006	------	------	-------	-25.29 (.468)	18.88 (.740)
2007	------	------	-------	--------	44.16 (.034)

**Numbers in parentheses are the significance level.

Part Two: Within Cohorts Analysis for Research Question #1

The second part of the analysis needed in order to answer the first research question concentrated on comparing student achievement for two cohorts by tracking their achievement from the seventh grade year to their eighth grade year. Cohort #1 involved the seventh grade students during the 2005-2006 school year and their eighth grade results during the 2006-2007 school year. Cohort #2 involved the seventh grade students during the 2006-2007 school year and their eighth grade results during the 2007-2008 school year. Since 4Sight Benchmark Assessments were administered during these years, I analyzed those scores along with the PSSA math scaled scores for each year. In order to look at the relationships between scaled scores, I first obtained correlation coefficients and then conducted regression analysis to analyze the prediction of PSSA scores.

The results indicated there was statistical correlation between the eighth grade 4Sight Benchmark Assessments and the eighth grade PSSA scaled scores for both cohort #1 and cohort #2. Likewise there was also statistical correlation between the seventh grade PSSA and the eighth grade PSSA scaled scores for cohort #1 and cohort #2. When looking at the strength of all correlations, it was determined to only be moderate. It appeared that the correlation strength of the seventh grade PSSA to the eighth grade PSSA had the same strength as using the eighth grade 4Sight to correlate with the eighth grade PSSA scaled scores (Table 4.7).

Table 4.7: 8th Grade Pearson Correlation Results

	Cohort #1	**Cohort #2**
8th grade 4Sight and 8th grade PSSA	.517	.476
7th grade PSSA and 8th grade PSSA	.443	.531

**p<.001

The Pearson Correlation analysis represented a much stronger correlation between the seventh grade 4Sight and the seventh grade PSSA scaled scores (Table 4.8).

Table 4.8: 7th Grade Pearson Correlation Results

	Cohort #1	**Cohort #2**
7th grade 4Sight and 7th grade PSSA	.871	.901

**p<.001

In order to determine if the 4Sight Benchmark Assessments can be used to significantly predict the PSSA scaled scores, multiple regression analysis was used. Using eighth grade PSSA for cohort #1 as the dependent variable, I used the eighth grade 4Sight Benchmark Assessments and their seventh grade PSSA results as predictors. The F value, used for measurement of deviation between individual distribution scores, was found to be $F(2,359)=97.816$; $p<.001$. Both the 4Sight Benchmark Assessments results and the seventh grade PSSA results significantly predicted the eighth grade PSSA for cohort #1. When used alone, the 4Sight scores accounted for 26.8% of the variance in eighth grade PSSA, and the seventh grade PSSA accounted for an additional 8.5% of the variance over and above the 4Sight. Overall, the total amount of variance accounted for by both predictors was 35.3% (Table 4.9).

Table 4.9: Cohort #1 Regression Results

	R	**Change in R^2**
Eighth Grade 4Sight Benchmark	.517	.268
Seventh Grade PSSA	.594	.085
Total		.353

$p<.001$

Both predictors were significant ($p<.001$), but 4Sight Benchmark Assessments contributed more information (standardized Beta coefficient=.418) than the seventh grade PSSA (standardized Beta coefficient=.308).

Multiple regression was conducted in order to predict the seventh grade PSSA by using the seventh grade 4Sight Benchmark Assessments as the predictor. The measurement of deviation between individual distribution scores resulted in $F(1,360)=1136.293$; $p<.001$ and the correlation coefficient, R, the measurement of linear dependence or strength, resulted in 75.9% of variance in the seventh grade PSSA scores was accounted for by knowing the students' 4Sight Benchmark Assessments scores (Table 4.10).

Table 4.10: Cohort #1 Regression Results

	R	**Change in R^2**
Seventh Grade 4Sight Benchmark	.871	.759

$p<.001$

The same analyses were conducted for Cohort #2 using eighth grade PSSA as the dependent variable, while using the eighth grade 4Sight Benchmark Assessments and their seventh grade PSSA results as predictors. The F value, used for measurement of deviation between individual distribution scores, was found to be $F(2,330)=91.935$; $p<.001$. Both the 4Sight Benchmark Assessments results and the seventh grade PSSA results significantly predicted the eighth grade PSSA for cohort #2. When used alone, the 4Sight scores accounted for 22.7% of variance in the eighth grade PSSA, and the seventh grade PSSA accounted for an additional 13.1% of the variance over and above the 4Sight. Overall, the total amount of variance accounted for by both predictors is 35.8% (Table 4.11).

Table 4.11 Cohort #2 Regression Results

	R	**Change in R^2**
Eighth Grade 4Sight Benchmark	.476	.227
Seventh Grade PSSA	.598	.131
Total		.358

$p<.001$

Both predictors were significant ($p<.001$), but the PSSA from the previous year (7^{th} grade scores) contributed more information (standardized Beta coefficient=.401) than the eighth grade 4Sight Benchmark Assessments results (standardized Beta coefficient=.304).

Multiple regression was conducted in order to predict the seventh grade PSSA by using the seventh grade 4Sight Benchmark Assessments as the predictor. The measurement of deviation between individual distribution scores resulted in $F(1,331)=1432.236$; $p<.001$ and the correlation coefficient, R, the measurement of linear dependence or strength, resulted in 81.2% of variance in the seventh grade PSSA scores was accounted for by knowing the students' seventh grade 4Sight Benchmark Assessments scores (Table 4.12).

Table 4.12: Cohort #2 Regression Results

	R	**Change in R^2**
Seventh Grade 4Sight Benchmark	.901	.812

$p<.001$

Overall, there were significant statistical results that supported a connection between the use of 4Sight Benchmark Assessments and student achievement on the PSSA for this rural school district. Even though the results for seventh grade were much different from eighth grade, the 4Sight Benchmark Assessments were a much stronger predictor of PSSA math scores in seventh grade compared to the eighth grade math scores. However, by using both PSSA scaled scores and 4Sight Benchmark Assessments scaled scores to assist in addressing math instruction, the teachers have two tools that may assist them in determining the assessed levels of their students and for making more informed decisions concerning classroom instruction.

Research Question 2: *To what extent are teachers implementing and using the 4Sight Benchmark Assessments results in their classrooms?*

In order to answer this question, I interviewed all math teachers (n=7) in this rural school to determine their personal views of benchmark assessments and how they used the results in their classrooms. The questions asked during the interviews were categorized into six sections. Section one consisted of questions that determined the teachers' background so I could determine how many years each teacher had taught math, how many years they taught math in this school, and to help determine which principal(s) supervised them. As stated earlier in this chapter, veteran teachers, those who have taught for over ten years, accounted for three out of seven, while those who have tenured, yet are considered fairly new to the educational field, account for two out of seven. The remaining two teachers taught less than three years and are considered non-tenured. One

teacher was present under the supervision of Principal A, B, and C, four teachers taught under the supervision of Principal B and C, and the remaining two teachers taught only under the supervision of Principal C. Determining the change in leadership during this study was important since research indicates that the support and active involvement of the principal can greatly impact the teachers' approach to preparing students for high-stakes tests (Black & Wiliam, 2005, Duffy, 2007, Halverson, Prichett, & Watson, 2007, Kaplan & Owings, 2001, Wiliam, dng,). During this study, it was discovered that Principal A supervised the school during the benchmark year (2003-2004). Principal B started supervising the teachers during the 2004-2005 school year and emphasized the importance of using student data, like the previous year's PSSA results, to drive instruction. 4Sight Benchmark Assessments were implemented school-wide by Principal B during the 2005-2006 school year and teachers were trained and provided time to collaborate with other professionals. Teachers used the data to inform students and parents of academic strengths and weaknesses and revised the curricula to align more naturally with the content standards. Principal C, even though was able to see the benefits of 4Sight data, concentrated on becoming acclimated to the school as the new principal during the 2007-2008 school year and did not stress the expectation that all teachers should continue to use the data to guide their instruction. During this school year, the teachers were aware of the existence of the data, but did not feel compelled to retrieve, analyze, or use the data in their classrooms.

The second category consisted of questions to help derive the teachers' perception on the professional development activities that were provided by the school district. All of the teachers interviewed (seven out of seven) felt they were adequately trained to

analyze and use 4Sight Benchmark Assessment data to modify their classroom instruction. Most of the teachers (six out of seven) shared that they were trained multiple times to retrieve and analyze 4Sight data. One teacher thought the multiple years of 4Sight training was something expected from all school districts. This teacher did not see the repeated or ongoing training as anything unusual. All teachers (seven out of seven) agreed they continued to receive ongoing training to analyze data.

The third category consisted of questions to determine the confidence level of each teacher when working with 4Sight Benchmark data. Again, all math teachers felt their background in mathematics assisted in analyzing data and felt confident to share the results with students and parents. The teachers perceived themselves having more experience dealing with student data, felt more comfortable analyzing the 4Sight data, and easily used the results effectively in their classrooms. All teachers shared that some of their colleagues who taught other content areas were not accustomed to working with the enormous amount of student data and were overwhelmed with analyzing, as well as modifying their instruction based on the results. This apprehension toward analyzing and using student data did not effect the math department as it reportedly effected other departments.

The fourth category involving questions related to teachers' use of 4Sight in their classroom tried to determine exactly how they used this information in their classrooms. All teachers (seven out of seven) felt their use of 4Sight data helped improve their students' achievement, however, only three out of the seven teachers interviewed shared very detailed examples of classroom activities they used in the classroom that were developed based from the results received from the 4Sight data. All of the teachers used

an activity that was developed to assist the students to become more aware of their achievement progress. The activity helped the teachers and students use the data in a more formative manner when the students analyzed their 4Sight data and collaborated with their teachers to determine appropriate goals. In terms of how frequent these activities were used in their classrooms, two out of seven teachers shared that they provided activities based on the 4Sight data on an on-going, almost daily basis. One teacher provided activities once per week, another teacher provided activities once every two weeks, one teacher provided activities once during each grading period, one teacher provided activities when it was appropriate and appeared in the curriculum, and one teacher provided activities during the mid-point of the second grading period when they analyzed the second set of scores and increased the frequency to everyday during the third grading period up to the administration of the PSSAs. When using other activities that were developed by other teachers into their classrooms, three out of the seven teachers stated they did not incorporate other activities, while the remaining four teachers stated they did so frequently. All of the teachers (seven out of seven) felt empowered to make informed decisions that drove their instruction based on 4Sight data yet seven out of seven teachers felt benchmarking their students did not impact their overall view and use of assessments. They unanimously felt their background in math made them more acceptable to using student data to drive instruction. Continuing with how the teachers use 4Sight data by looking at how frequent they meet with other professionals to analyze data, four out of seven teachers stated they met monthly with other professionals including administrators and counselors. However three teachers reported they met

monthly to discuss the 4Sight results, three teachers stated they met once per grading period to discuss the results, and one teacher stated they only met as needed.

The fifth category addressed the teachers' perception of the effectiveness of 4Sight Benchmark Assessments. All of the teachers (seven out of seven) felt benchmarking impacted their classroom instruction and it directly impacted student achievement. The teachers' opinion (seven out of seven) were consistent that the 4Sight Benchmark Assessments is a good tool for teachers and with more time to use this tool will increase their ability to implement it in different ways within their classrooms.

The final category consisted of a question that asked the teachers would they recommend the continued use of 4Sight Benchmark Assessments. Overwhelmingly, all teachers (seven out of seven) felt the benchmark assessments should be continued. However, six out of seven teachers felt the tests should be given less frequently. Four teachers felt three times should be adequate in order to provide sufficient data, while two teachers felt twice a year should be sufficient. Only one teacher felt the tests should be given as they were in the past, four to five times per year at the end of each quarter. The teachers who stated the frequency of the administration of the tests should be reduced shared that their students were complaining, showed signs of test fatigue, and appeared to not take the 4Sight Benchmark Assessments or the PSSAs seriously. (Table 4.13)

Table 4.13: Teachers' Interview Results

Category:	**Professional Development**
Results:	7 out of 7 teachers felt they were adequately trained. 6 out of 7 teachers reported being trained multiple times. 7 out of 7 agreed they have ongoing training.
Quotes:	"…while I was student teaching I became very familiar with it there…so when I came here it was kind of like I just thought this is what schools do…" (Teacher D)

Category:	Confidence Levels
Results:	7 out of 7 felt confident (due to their math backgrounds) 7 out of 7 felt confident to share results with students and parents.
Quotes:	"…on a scale of high, medium, or low…I would say medium. However, I am advancing as more data becomes available." (Teacher C) "…Absolutely. I spoke one on one with each of my students in all levels I teach, twice so far this year. So, each student, I spoke to twice with their data in front of them." (Teacher E)

Category:	Teachers' Use
Results:	7 out of 7 felt their use of 4Sight help improve student achievement. 3 out of 7 shared detailed examples of activities they used in their classroom. 7 out of 7 used a tool to help the teachers & student use the data in a more formative manner. 2 out of 7 provided activities on an on-going, almost daily basis. 1 out of 7 provided activities at least once per week. 1 out of 7 provided activities at least once every two weeks. 1 out of 7 provided activities at least once during each grading period. 1 out of 7 provided activities at least when it fits naturally in the curriculum. 1 out of 7 provided activities at mid-point of 2nd grading period and increased frequency up to the PSSA testing week. 7 out of 7 teachers felt empowered to make informed decisions based on student data.
Quotes:	"Yes, I think so. For those students who took it seriously when taking the tests I thought it helped them to plan and also for me, their weaknesses and strengths… to see what needed to be adjusted and then where I needed to go. If they already knew something so strongly then I could just review that with them and not hit it hard." (Teacher A) "Once we reviewed all of their 4Sight data, I had them look at what their strengths were, what their weaknesses were, where they can improve and the majority of them said 'data analysis and probability.' I went through those concepts in the book and I hit them hard until the PSSAs." (Teacher G)

Category:	Teachers' Perception of Effectiveness
Results:	7 out of 7 teachers felt benchmarking impacted classroom instruction and directly impacted student achievement.

	7 out of 7 felt 4Sight is a good tool and with more time will increase their ability to implement the data in different ways.
Quotes:	"It seems as time goes on the whole process has become more routine as opposed to the beginning when it was fresh and new and we could really do big things with it. Now, it is becoming more routine. Students don't seem to be taking them as seriously. You really have to walk around and tap the corners of their desks and refocus them and encourage them. The more routine it becomes for them, the data is not going to be as exact." (Teacher A) "Yes, 4Sight definitely impacts student achievement if the teacher uses the data." (Teacher F)

Category:	**Teachers' recommendation to continue the use of 4Sight**
Results:	7 out of 7 teachers felt 4Sight should be continued. 6 out of 7 felt the tests should be given less frequently. 4 out of 7 felt the tests should be given 3 times per year. 2 out of 7 felt the tests should be given 2 times per year. 1 out of 7 felt the tests should be given as they were in the past.
Quotes:	"Yes and no. Yes, less often. It is a good practice for the PSSA tests. It looks like it, although, in all honesty, the students have seen enough PSSAs by that time that they don't need the practice format. No, because the teachers are receiving the same data over and over and it does take time away from instruction." (Teacher A) "Yes. Not as much. Probably like three times, the beginning, the middle, and somewhere towards the end. Some of the concepts on the PSSAs are not on the 4Sights, so then we put too much emphasis on the 4Sights and sometimes we overlook or forget the concepts that will be tested on the PSSAs." (Teacher B)

Based on this information, the teachers appeared to consistently implement and use 4Sight Benchmark Assessments results in their classrooms. Even though the use of 4Sight Benchmark data is fairly new, the teachers are attempting to make effective use of the data. The teachers felt that as more training is provided, specifically to help assist them in developing appropriate activities, they will become more proficient in using this tool.

CHAPTER 5

SUMMARY, INTERPRETATIONS, AND RECOMMENDATIONS

Summary / Interpretation

The purpose of this research was to determine the effectiveness of the use of benchmark assessments (4Sight) as a tool to increase student achievement on the standardized tests (PSSAs) and to determine the extent to which teachers actually use 4Sight data to modify classroom instruction. Ultimately, the question needs to be answered whether it is really worth the instructional time to incorporate benchmark assessments and whether benchmark assessments increase student achievement.

Research Question 1: *What is the nature of the relationship between the use of benchmark assessments (4Sight Benchmark Assessments) and student achievement on the state standardized tests (PSSA) for a rural school district?*

In order to determine the effectiveness of a particular tool, it is always wise to determine what was possibly occurring with the data prior to the implementation of the tool. I was fortunate to have two years worth of student data where the teachers relied on instructional strategies to increase student achievement. The average scale scores increased (1.85) during the 2003-2004 to 2004-2005 school years. Looking at the mean scale scores during the 2004-2005 to 2005-2006 school years, the data showed an increase of 30.80. This appears to be the impact that all educators would hope for but the following year, 2005-2006 to 2006-2007 school years, the mean scale scores decreased

by 25.29. The most intriguing part of this analysis showed another increase during the 2006-2007 to 2007-2008 school years by 44.16. This increase is intriguing since Principal C admitted to administering the 4Sight Benchmark Assessments but due to the needed focus on transitioning as the new building principal, did not require the teachers to use the data. The teachers I interviewed agreed that they knew the scores were available, but did not make using the data to modify their classroom instruction a priority. The most important part of this analysis is that even though the average scale scores may have fluctuated during any given year, the standard deviation decreased each year, indicating the students' PSSA scores are more tightly distributed around the mean. This supports that student achievement has consistently become more homogeneous. By examining the mean scale scores over time, the scores are significantly different across the five years.

Using the Games-Howell post hoc test, it was revealed that there was not a significant increase in the mean scaled scores when comparing 2003-2004 scores with the 2004-2005 scores. When using 2003-2004 data as the baseline and comparing subsequent years of data, there were only moderate significant difference between the baseline year and the 2005-2006 student data. There were higher significant differences when comparing the baseline year to the 2007-2008 student data.

In addition to understanding the behaviors of the mean scaled scores over time, it was important to also track the achievement of the students by comparing their seventh grade math scores to their eighth grade math scores. Cohort #1 students were seventh graders during the 2005-2006 school year and eighth graders during the 2006-2007 school year. Cohort #2 students were seventh graders during the 2006-2007 school year

and eighth graders during the 2007-2008 school year. The data showed statistical correlation between the eighth grade 4Sight scores and the eighth grade PSSA scores for both cohort #1 and cohort #2. There was also statistical correlation between the seventh grade PSSA data and the eighth grade PSSA data for both cohorts. In conclusion it appears the correlation of eighth grade 4Sight and the eighth grade PSSA had similar moderate effect for both cohort #1 (.517) and cohort #2 (.476). In addition, the correlation of seventh grade PSSA and the eighth grade PSSA had moderate effect for both cohorts, cohort #1 (.443) and cohort #2 (.531). This moderate effect was surprising since the reports from Success For All advertised a stronger correlation and thereby suggested the schools would benefit greatly by using 4Sight instead of just using the previous PSSA scores (Success For All, 2004, 2007). There was a slight increase in correlation between the use of 4Sight (.517) in comparison to just using the seventh grade PSSA results (.443) for cohort #1. When looking at the results for cohort #2, the use of 4Sight (.476) had less effect than using the seventh grade PSSA (.531). However, the Pearson Correlation analysis did represent a much stronger correlation between the seventh grade 4Sight and the seventh grade PSSA for both cohort #1 (.871) and cohort #2 (.901). These correlation results more closely reflected the results that were reported by Success For All in 2008 where the math correlation ranged from .86 to .91 (Success For All, 2008, p.18). Even though this showed a significant correlation between the 4Sight and PSSAs, I anticipated a stronger correlation between using 4Sight Benchmark Assessments for each group of students than just relying on the previous year's student data on the PSSA.

Finally, the use of multiple regressions was used to determine if 4Sight Benchmark Assessments could be used to significantly predict the PSSA scaled scores. For cohort #1, when used alone, 4Sight scores accounted for 26.8% of the variance in the eighth grade PSSA scores. If used in addition to the seventh grade PSSA scores, the total amount of variance accounted for by both predictors increased to 35.3%. The 4Sight Benchmark Assessments contributed more information with having a standardized Beta coefficient of .418 when compared to the seventh grade PSSA, which had a standardized Beta coefficient of .308. Similar results were revealed when looking at cohort #2 student data. When used alone, 4Sight scores accounted for 22.7% of the variance in the eighth grade PSSA scores. If used in addition to the seventh grade PSSA scores, the total amount of variance accounted for by both predictors increased to 35.8%. The seventh grade PSSA contributed more information with having a standardized Beta coefficient of .401 when compared to the 4Sight Benchmark Assessments, which had a standardized Beta coefficient of .304.

The most profound result occurred when using multiple regression again to predict the seventh grade PSSA by using the seventh grade 4Sight Benchmark Assessments as a predictor, it was discovered the correlation coefficient resulted in 75.9% of variance in the seventh grade PSSA scores as accounted for by knowing the students 4Sight scores for cohort #1 and 81.2% of variance for cohort #2.

Overall, there are significant statistical results that support a connection between the use of 4Sight Benchmark Assessments and student achievement on the PSSA for this rural school district. Even though the results for seventh grade were much different from eighth grade, the 4Sight Benchmark Assessments were a much stronger predictor of

PSSA math scores in seventh grade compared to the eighth grade math scores. However, by using both PSSA scaled scores and 4Sight Benchmark Assessments scaled scores to assist in addressing math instruction, the teachers have two tools that may assist them in determining the assessed levels of their students and for making more informed decisions concerning classroom instruction.

Research Question 2: *To what extent are teachers implementing and using the 4Sight Benchmark Assessments results in their classrooms?*

Generally, teachers will tend to use tools like data analysis in their classroom if they feel comfortable, confident, and have a sense of reassurance that the tool can directly impact student achievement in a positive manner. All of the math teachers interviewed (n=7) for this study felt the professional development training they were provided by the school district and their principals was adequate to enable them to analyze and use the 4Sight Benchmark Assessment results to modify their classroom instruction. Even though the teachers stated they felt comfortable, based on their content area, using the data to drive instruction, they all felt ongoing training would be beneficial to them. As it is important for teachers to have ongoing training in working with student data through analysis and implementation, it is also equally important for the school administrators to also be a supporter and active participant in the endeavor. Leslie Kaplan and William Owings determined in their research that "principals and assistant principals who provide ongoing professional development in varied formats to assist novice and marginal teachers learn and practice these effective pedagogical strategies can also increase the

prevalence of these behaviors in their schools" (Kaplan & Owings, 2001, p. 18). "When teachers lack confidence in the…program, they want a principal who can help them understand the new expectations and either reassure them that their instructional skills are up to the challenge or respectfully introduce them to the instructional practices that will help their students to be successful on these important measures" (p. 22). Once the teachers develop a sense of assurance when working with student data, interpreting the results, and developing classroom activities that address the weak skills, teachers tend to teach to each student's learning needs.

The teachers interviewed felt their backgrounds in mathematics provided a level of comfort to analyze data and all of the teachers (100%) felt their use of 4Sight data helped improve their students' achievement. 4Sight Benchmark Assessments provided the teachers with current data that reflected the strengths and weaknesses of each student which helped the teachers start a dialogue with other professionals and students in order to address these areas. During the interviews, the teachers shared their increased attention to analyzing student data and using this information to develop instructional activities that addressed the focus skill. Even though only three out of the seven teachers interviewed shared very detailed examples of classroom activities they developed based on the 4Sight results and used in their classrooms, when asked if they used other activities developed by other teachers, three teachers stated they did not incorporate other activities, while four teachers stated they did so frequently. In the past the teachers addressed student academic achievement from teacher-derived assessments. The teachers felt the use of 4Sight assisted them more quickly and accurately to identify specific areas through the use of item analysis as well as, allowed them to develop activities that

addressed specific skills. One activity frequently used by teachers was a self-assessment activity that required the students to analyze and interpret their test scores. The activity required the students to determine their growth and to define factors that may have contributed to the growth. For areas where there weren't growth, again the students had to calculate how much their scores declined and what factors contributed to the decline. The teachers felt it was very important to conference with each student to help facilitate a plan to increase student achievement. The teachers shared that even though they had always referred to student data to help inform them of student progress, the use of 4Sight encouraged them to make a difference through their classroom instruction and improve student achievement. It also encouraged them to actively involve the students rather than just report results as it was done in the past. "The students were very concerned when they saw a decline in their scores, but they were also very happy when they could recognize growth. This really encouraged the students to continue trying to do better" (Teacher G).

All of the teachers (100%) felt benchmarking impacted their classroom instruction and it directly impacted student achievement. The teachers' opinion (100%) were consistent that the 4Sight Benchmark Assessments is a good tool for teachers and with more time to use this tool it will increase their ability to implement it in different ways within their classrooms. When asked would they recommend the continued use of 4Sight Benchmark Assessments, overwhelmingly, all teachers (100%) felt the benchmark assessments should be continued. However, six out of seven teachers felt the tests should be given less frequently. Four teachers felt three times should be adequate in order to provide sufficient data, while two teachers felt twice a year should be sufficient.

"[4Sight] is a good practice for the PSSA tests. It looks like it, although, in all honesty, the students have seen enough PSSAs by that time that they don't need the practice format…teachers are receiving the same data over and over and it does take time away from instruction" (Teacher A). Only one teacher felt the tests should be given as they were in the past, four to five times per year at the end of each quarter. The teachers who agreed that the frequency of the administration of the tests should be reduced shared that their students were complaining, showing signs of test fatigue, and appeared to not take the 4Sight Benchmark Assessments or the PSSAs seriously even though teachers consistently derived ways to use the assessment results in a more formative manner. Students don't seem to be taking them as seriously. You really have to walk around and tap the corners of their desks and refocus them and encourage them. The more routine it becomes for them, the data is not going to be as exact" (Teacher A).

Recommendations

This study suggests there are significant statistical results that support a connection between the use of 4Sight Benchmark Assessments and student achievement on the PSSA for this rural school district. In addition the teachers are in favor of continuing to use this tool to help them make informed decisions in their classrooms. This study alone cannot generalize the results this school district experienced with other school districts that use this tool. However, this study combined with other similar studies can start adding to the results that could one day confirm or refute that 4Sight Benchmark Assessments can increase student achievement on PSSAs.

In addition to replicating this study for other schools, another finding that was repeated during the teachers' interviews was that they discovered for three years of administering 4Sight that students were consistently weak in the same area, geometry and data analysis & probability. Even though the curriculum was aligned to state standards and revised in order to address these weak areas after the initial discovery in 2005, these areas continue to be weak for both seventh and eighth grade students. This brings to question if the weak areas are due to the specific design of the 4Sight or are there still problems within the curriculum. Based on this trend, it is recommended that a more thorough analysis be conducted to analyze the subcategories of both the 4Sight Benchmark Assessments and the PSSA to determine if these findings are not merely the result from the test format or whether the students are demonstrating weak skills in these subcategories due to the math curriculum.

Another recommendation for future studies is a closer analysis of the types of classroom activities the teachers are developing. During the interviews, teachers shared that they wondered if their activities really addressed specific skills with the needed rigor that would result in increasing student proficiency. The teachers felt most of the professional development focused on helping them analyze the data and failed to help them develop rigorous classroom activities that addressed a specific skill. Two out of the seven teachers resorted to using the Pennsylvania Department of Education website along with other sites to find sample activities that are linked to the state standards and eligible content. All of the teachers felt there should be something more offered to teachers to help them quickly develop activities to address specific content areas or maybe a warehouse of previously developed activities by other teachers who use 4Sight that could

easily be implemented into their classrooms. Even though the teachers tried to develop new activities that addressed their students' needs, they did agree there should be more offered to encourage teachers and school districts to continue the use of 4Sight. They were concerned that once their colleagues who taught other disciplines become comfortable analyzing student data, they would face frustration while developing rigorous classroom activities that may lead to resorting back to past practice of not using various student data to drive instruction.

It appears that even though there are increasing studies that focus on the analysis of 4Sight Benchmark Assessments, it is too early to determine the long term value of this tool on student achievement. As the teachers in this study indicated, it is more beneficial to continue helping teachers develop their analytical skills and train teachers how to transform student data to something they can interpret and use to modify their instruction according to the needs of their students than to resort back to past practice that resulted in many children being left behind and not experiencing success in our educational system. Yet the ultimate questions have yet to be answered: Is it worth the loss of instructional time to administer these benchmark assessments? Are we gaining more information that could increase student achievement by administering benchmark assessments or are we doing a disservice to our children by forcing their world to be dominated by repeated assessments and continuous evaluations?

REFERENCES

Act 299 of 1963. PA Law. Section 299.1.

Aladjem, D. K. & Borman, K. M. (2006). *Summary of findings from the National Longitudinal Evaluation of Comprehensive School Reform.* Paper presented at the annual meeting of the American Education Research Association, San Francisco, CA.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). Standards for teachers competence in educational assessment of students. *Educational Measurement: Issues and Practice, 9(4),* 30-32.

Archibald, S. (2006). Narrowing in on educational resources that do affect student achievement. *Peabody Journal of Education*, 81(4), 23-42.

Arter, J. A. (dng). Assessment for learning: Classroom assessment to improve student achievement and well-being. (ERIC Document Number: ED480068) Assessment for Learning.

Black, P. & Wiliam, D. (1988, October). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80(2),* 139-149.

Black, P. & Wiliam, D. (2005). Changing teaching through formative assessment: Research and practice. The King's-Medway-Oxfordshire Formative Assessment Project. *English Literature Review*. 223-240.

Bloom, B. (1984). The search for methods of group instruction as effective as one-to-one tutoring. *Educational Leadership 41(8),* 4-17.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning.* New York, NY: McGraw-Hill.

Bracey, G. W. (2000). *Thinking About Tests and Testing: A Short Primer in "Assessment Literacy."* Washington, D.C.: American Youth Policy Forum in cooperation with the National Conference of State Legislatures. Online: www.aypf.org/publication/braceyrep.pdf.

Brandt, R. (1998). *Assessing student learning: New rules, new realities*. Arlington, VA: Educational Research Service.

Brookhart, S. M. (2001). *The standards and classroom assessment research.* Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, Dallas, Texas. (ERIC Document Number ED 451-189)

Bushweller, K. (1997, September). Teaching to the test. *American School Board Journal*. 184(9), 20-25. (ERIC Document Reproduction Service No. EJ548992) Retrieved November 1, 2008, from ERIC database.

Canner, J. (1992, September). Regaining the public trust: A review of school testing programs, practices. *NASSP Bulletin*. 6-15.

Chappius, S. & Chappius, J. (2007). The best value in formative assessment. *Educational Leadership, 65(4),* 14-18.

Crooks, T. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58(4),* 438-481.

Diamond, J. B. (2007). *Where the rubber meets the road: Rethinking the connection between high-stakes testing policy and classroom instruction.* Sociology of Education. 80(4), 285-313.

Data Recognition Corporation. (2006). *Technical Report for the Pennsylvania System of School Assessment 2005 Reading and Mathematics.* Retrieved December 25, 2008 from http://www.pde.state.pa.us/a_and_t/lib/a_and_t/2005_Performance_Levels__Validation__Technical_Report.pdf

Data Recognition Corporation. (2007a). *Technical Report for the Pennsylvania System of School Assessment 2006 Reading and Mathematics Grades 4, 6, and 7.* Retrieved December 25, 2008 from http://www.pde.state.pa.us/a_and_t/lib/a_and_t/2006_ReadingMathGr4_6_7_Tech_Report.pdf

Data Recognition Corporation. (2007b). *Technical Report for the Pennsylvania System of School Assessment 2006 Reading and Mathematics Grades 5, 8, and 11.* Retrieved December 25, 2008 from http://www.pde.state.pa.us/a_and_t/lib/a_and_t/2006_ReadingMathGr5_8_11_Tech_Report.pdf

Data Recognition Corporation. (2008). *Technical Report for the Pennsylvania System of School Assessment 2007 Reading and Mathematics Grades 3, 4, 5, 6, 7, 8, and 11.* Retrieved December 25, 2008 from http://www.pde.state.pa.us/a_and_t/lib/a_and_t/2007_PSSA_Reading_&_Mathematics_Tech_Report.pdf

Datnow, A., Park, V. & Wohlsletter, P. (2007). Achieving with data: How high performing school systems assess data to improve instruction for elementary students. Los Angeles: Center on Educational Governance, Rossier School of Education, University of Southern California.

Duffy, G. G. (2007). Thriving in a high-stakes testing environment. *Journal of Curriculum and Instruction*. 1(1), 7-13.

Duke, N. K., & Ritchhart, R. (1997, October). Standardized test preparation. *Instructor*. 107(3), 89-92, 119.

Executive Summary: The No Child Left Behind Act of 2001. (2002). Retrieved from http://www.ed.gov/nclb/overview/intro/execsumm.pdf/

Fashola, O. S. & Slavin, R. E. (1997). Effective and replicable programs for students placed at risk in elementary and middle schools. Johns Hopkins University.

Gibbs, G. & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education, 1,* 3-30.

Goals 2000: Educate America Act of 1994. Public Law No. 103-227. (1994). Washington, DC: U.S. Government Printing Office.

Gulek, C. (2003). Preparing for high-stakes testing. *Theory into Practice*, 42(1), 42-50.

Guskey, T. R. (2007). The rest of the story. *Educational Leadership, 65(4),* 28-35.

Halverson, R., Prichett, R. B., & Watson, J. G. (2007). Formative feedback systems on the new instructional leadership. Madison: University of Wisconsin – Wisconsin Center for Education Research, School of Education, University of Wisconsin – Madison. Retrieved from http://www.wcer.wisc.edu/publications/workingPapers/index.php.

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

Heubert, J. P., & Hauser, R. M. (1999). *High Stakes: Testing for Tracking, Promotion, and Graduation*. Washington, D.C.: National Academy Press.

Kaplan, L. S., & Owings, W. A. (2001). How principals can help teachers with high stakes testing: One survey's findings with national implications. *NASSP Bulletin*, 85(622) 15-23.

Kilian, L. J. (1992). A school district perspective on appropriate test preparation practices: A reaction to Popham's proposals. *Educational Measurement: Issues and Practices*, 11(4), 12-15, 26.

Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. In J. L. Herman & E. H. Haertel (eds.), *Uses and Misuses of Data in Accountability Testing.* Yearbook of the National Society for the Study of Education, 104 vol. 2,

99-118.

Lai, E. R., Waltman, K. (2008). Test preparation: Examining teacher perceptions and practices. *Educational Measurement, Issues and Practices*, 27(2), 28-45.

Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment that keeps learning on track minute-by-minute, day-by-day. *Educational Leadership, 63(3),* 19-24.

Linn, R. L. (2000, March). Assessments and accountability. *Educational Researcher*, 29(2), 4-16.

Linn, R. L. (2006). Validity of influences from test-based educational accountability systems. *Journal of Personnel Evaluation in Education*, 19, 5-15.

Lutz-Doemling, C. K. (2007). An examination of the Pennsylvania 4sight benchmark assessments as predictors of Pennsylvania system of school assessment performance. (Doctoral Dissertation, Lehigh University, 2007). *ProQuest LLC.* (UMI No. 3314497)

Marzano, R. J. (2006). Classroom *assessment & grading that work.* Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J., Pickering, D. J., & McTighe, J. (1993). *Assessing student outcomes: Performance assessment using the dimensions of learning model.* Alexandria, VA: Association for Supervision and Curriculum Development.

Mele-McCarthy, J. A. (2007). NCLB assessment for accountability: Good teaching or teaching to the test? *Perspective on Language and Literacy*, 33 (1), 11-16.

Mintrop, H., & Trujillo, T. (2005). Corrective action in low performing schools: Lessons for NCLB implementation from first-generation accountability systems. University of South Florida, College of Education: *Education Policy Analysis Archives, 13(48)*.

Miyasaka, J. (2000). A framework for evaluating the validity of test preparation practices. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA. (ERIC Document Number: ED454256) Retrieved December 19, 2008 from ERIC database.

Muir, M. (2001). When stakes are high. *Northwest Educational Magazine.* Retrieved November 25, 2008 from http://www.nwrel.org/nwedu/2001fall/stakes.html

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* Washington, D.C.: U.S. Government

Printing Office.

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? University of South Florida, College of Education: *Education Policy Analysis Archives, 14(1).*

No Child Left Behind Act of 2001. Public Law 107-110. (2002). Retrieved from http://www.ed.gov/legislation/ESEA02/

No Child Left Behind: What No Child Left Behind Means for Pennsylvania. (2003). Retrieved from http://www.pde.state.pa.us/nclb/cwp/

No Child Left Behind: Overview of No Child Left Behind. (2003). Retrieved from http://www.pde.state.pa.us/nclb/cwp/

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

Olson, L. (2005b). Not all teachers keen on periodic tests. *Education Week*, *25(13),* 13.

O'Neill, P., & Johnson, C. (2007). *No Child Left Behind: Compliance Manual (2nd ed.)* Horsham, Pennsylvania: LPR Publications.

Pennsylvania State Board of Education. (1999). Chapter 4, academic standards and assessment. Harrisburg, PA: Pennsylvania State Board of Education. Retrieved December 22, 2008 from http://www.pacode.com/secure/data/022/chapter4/s4.51.html.

Pogrow, S. (1994, January 1). Helping students who "just don't understand." *Educational Leadership*, 52(3), 62-66.

Pogrow, S. (2000a). The unsubstantiated 'success' of success for all: Implications for policy, practice, and the soul of our profession. *Phi Delta Kappan*, 81(8), 596 600.

Pogrow, S. (2000b). Success for all does not produce success for students. *Phi Delta Kappan*, 82(1), 67-80.

Pogrow, S. (2002, February). Success for all is a failure. *Phi Delta Kappan*, 83(6), 463.

Popham, W. J. (2001). *The truth about testing: An educator's call to action.* Alexandria, VA: Association for Supervision and Curriculum Development.

Popham, W. J. (2003). *Test better, teach better: The instructional role of assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Popham, W. J. (2008). *Transformative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Potteiger, C. A. (2008). Correlation study of the Pennsylvania system of school mathematics assessment with the 4sight mathematics assessment. (Doctoral Dissertation, Widener University, 2008). *ProQuest LLC.* (UMI No. 3313320)

Protheroe, N., Educational Research Service, A., & National Association of Elementary School Principals, A. (2001, January 1). *Meeting the Challenges of High-Stakes Testing. Essentials for Principals [TM]*. (ERIC Document Reproduction Service No. ED459522) Retrieved June 21, 2008, from ERIC database.

Reeves, D. B. (2008). Reframing *teacher leadership: To improve your school.* Alexandria, VA: Association for Supervision and Curriculum Development.

Rex, L. A., Nelson, M. C. (2004). How teachers' professional identities position high stakes test preparation in their classrooms. *Teachers College Record*, 106(6), 1288-1331.

Ruiz-Primo, M. A., & Furtak, E. M. (2006). Informal formative assessment and scientific inquiry: Exploring teachers' practices and student learning. *Educational Assessment*, 11(3&4), 205-235.

Schaffer, M., Burry-Stock, J.A., Cho, G., Boney, T., & Hamilton, G. (2000). What do kids think when their teachers grade? Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Shapiro, E. S., Solari, E., & Petscher, Y. (2008). Use of a measure of reading comprehension to enhance prediction on the state high stakes assessment. *Learning and Individual Differences*, 18, 316-328.

Shepard, L. A. (1989). Inflated test score gains: Is it old norms or teaching the test? *Center for Research on Evaluation, Standards, and Student Testing.* Los Angeles, CA: University of California.

Sinclair, A. L., Thacker, A. A., & HumRRO. (2005). *Relationships Among Pennsylvania System of School Assessment (PSSA) Scores, University Proficiency Exam Scores, and College Course Grades in English and Math.* Retrieved December 26, 2008 from http://www.pde.state.pa.us/stateboard_ed/lib/stateboard_ed/HumRROPSSAUnivProficiencyTestReport.pdf

Slavin, R. E. (dng). A model of effective instruction. John Hopkins University, Center for Research on the Education of Students Placed at Risk.

Slavin, R. E. (2001, September). Response: The facts about comprehensive school reform. *Educational Leadership*, 84-85.

Slavin, R. E. (2002, February). Mounting evidence supports the achievement effects of success for all. *Phi Delta Kappan*, 469-471, 480.

Slavin, R. E. (2008, January/February). Perspectives on evidence-based research in education: What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 37(1), 5-14.

Slavin, R. E., & Madden, N. A. (dng). Reducing the gap: Success for all and the achievement of African American students, *The Journal of Negro Education*, 75(3), 389-400.

Slavin, R. E., Madden, N. A., John Hopkins University, & University of York. (2008). Understanding bias due to measures inherent to treatments in systematic reviews in education. Paper presented at the annual meetings of the Society for Research on Effective Education, Crystal City, Virginia, March 3-4, 2008.

Smith, M. L. (1991). Meanings of test preparation. *American Educational Research Journal*, 28(3), 521-542.

Steadman, M. (1998). Using classroom assessment to change both teaching and learning. *New Directions for Teaching and Learning, 75,* 23-35.

Sternberg, R. J. (2007). Assessing what matters. *Educational Leadership, 65(4),* 20-26.

Stiggins, R. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan International*, 83(10), 758-765. Retrieved from http://www.pdkintl.org/kappan/k0206sti.htm

Stiggins, R., Arter, J., Chappius, J., & Chappius, S. (2006). *Classroom assessment for student learning: Doing it right-using it well*. Portland, OR: Educational Testing Service.

Stiggins, R., Chappius, J. (2006). Using student-involved classroom assessment to close achievement gaps. *Theory Into Practice, 44(1),* 11-18.

Stigler, J. W. & Hiebert, J. (1997). Understanding and improving classroom mathematics instruction: An overview of the TIMSS video study. *Phi Delta Kappan, 79(1),* 14-21.

Success For All Foundation (2004). Powerpoint Presentation: *Pennsylvania Benchmarks 4Sight Returning User: Benchmark Review and Member Center Update*. Retrieved from Member Center Website http://members.successforall.net.

Success For All Foundation (2007). 4Sight correlations remain strong: 4Sight re-correlated to 2007 PSSA. *Pennsylvania and 4Sight.*

Success For All Foundation. (2007, November). 4Sight reading and math benchmarks 2006-07 technical report for Pennsylvania. Baltimore, Maryland.

Success For All Foundation. (2008). 4Sight reading and math benchmarks 2007-2008 technical report for Pennsylvania. Baltimore, Maryland.

Tankersley, K. (2007). Tests *that teach: Using standardized tests to improve instruction.* Alexandria, VA: Association for Supervision and Curriculum Development.

Thacker, A. A., Dickinson, E. R. & Koger, M. E. (2004). *Relationships Among the Pennsylvania System of School Assessment (PSSA) and Other Commonly Administered Assessments.* Alexandria, VA: Human Resources Research Organization.

Tomlinson, C. A. (2000, September). Reconcilable differences? Standards-based teaching and differentiation. *Educational Leadership*, 6-11.

Tomlinson, C. A. (2007). Learning to love assessment. *Educational Leadership, 64(4),* 8-13.

U.S. Department of Education. (2007). Building *on results: A blueprint for strengthening the No Child Left Behind Act,* Washington, D.C.

Volante, L. (2006). Toward appropriate preparation for standardized achievement testing. *The Journal of Educational Thought*, 40(2), 129-144.

Wiliam, D. (dng). *Assessment and the regulation of learning*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Wiliam, D. (2006). Formative assessment: Getting the focus right. *Educational Assessment*, 11(3&4), 283-289.

Wiliam, D., Lee, C., Harrison, C., Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education, 11(1),* p.49-65.

Zhang, Z. (1996). *Teacher assessment competency: A Rasch model analysis.* Paper presented at the annual meeting of the American Educational Research Association, New York. (ERIC Document Number ED 400 322)

Appendix A: PSSA Math Test Plan per Operational Form

Table 1

2006 MATH TEST PLAN									
GRADE	**No. of Forms**	**No. of Core MC per Op. Form**	**No. of Matrix MC per Op. Form**	**No. of Embedded FT MC per Op. Form**	**No. of Core 4-pt OE per Op. Form**	**No. of Matrix OE per Op. Form**	**No. of Embedded FT OE per Op. Form**	**Total of No. of Items per Op. Form MC/OE**	**Total No. of Core Points per Op. Tests**
7	16	54	4	8	3	1	1	66/5	66
8	20	54	6	6	3	1	1	66/5	66
	MC = Multiple Choice Test Items OE = Open-ended Test Items FT = Field Tested Items Core = Test Items taken by all students Matrix = Test Items Assigned to Selected Forms								

Data Recognition Corporation. (2007a). *Technical Report for the Pennsylvania System of School Assessment 2006 Reading and Mathematics Grades 4, 6, and 7.* Retrieved December 25, 2008 from http://www.pde.state.pa.us/a_and_t/lib/a_and_t/2006_ReadingMathGr4_6_7_Tech_Report.pdf

Data Recognition Corporation. (2007b). *Technical Report for the Pennsylvania System of School Assessment 2006 Reading and Mathematics Grades 5, 8, and 11.* Retrieved December 25, 2008 from http://www.pde.state.pa.us/a_and_t/lib/a_and_t/2006_ReadingMathGr5_8_11_Tech_Report.pdf

Appendix B: PSSA Math Test Plan per Operational Form

Table 2

2007 MATH TEST PLAN									
GRADE	**No. of Forms**	**No. of Core MC per Op. Form**	**No. of Matrix MC per Op. Form**	**No. of Embedded FT MC per Op. Form**	**No. of Core 4-pt OE per Op. Form**	**No. of Matrix OE per Op. Form**	**No. of Embedded FT OE per Op. Form**	**Total of No. of Items per Op. Form MC/OE**	**Total No. of Core Points per Op. Tests**
7, 8	20	54	6	6	3	1	1	66/5	66
	MC = Multiple Choice Test Items OE = Open-ended Test Items FT = Field Tested Items Core = Test Items taken by all students Matrix = Test Items Assigned to Selected Forms								

Data Recognition Corporation. (2008). *Technical Report for the Pennsylvania System of School Assessment 2007 Reading and Mathematics Grades 3, 4, 5, 6, 7, 8, and 11.* Retrieved December 25, 2008 from http://www.pde.state.pa.us/a_and_t/lib/a_and_t/2007_PSSA_Reading_&_Mathematics_Tech_Report.pdf

Appendix C: Pennsylvania 4Sight Mathematics Benchmark Assessment Descriptive Statistics and Validity Correlation to the 2007 Mathematics PSSA *(Data was unavailable for grade 8 at the time of SFA report)*

Table 3

7TH GRADE MATH FORM 1				
	N	**MEAN**	**SD**	**R**
4SIGHT	120	17.03	7.14	0.71
PSSA	120	1311.89	163.46	
PSSA Math = 1034.280 + 16.298*(4Sight – 1)				

Table 4

PERFORMANCE LEVEL	**4SIGHT MATH TOTAL SCORE**	**PSSA MATH**
Below Basic	1 – 10	1034 – 1181
Basic	11 – 17	1197 – 1295
Proficient	18 – 27	1311 – 1458
Advanced	28 – 36	1474 - 1605

Table 5

7TH GRADE MATH FORM 1 CORRELATION					
4Sight Total Score	**PSSA**	**4Sight Total Score**	**PSSA**	**4Sight Total Score**	**PSSA**
1	905	13	1232	25	1560
2	932	14	1259	26	1587
3	959	15	1287	27	1614
4	986	16	1314	28	1642
5	1014	17	1341	29	1669
6	1041	18	1369	30	1696
7	1068	19	1396	31	1724
8	1096	20	1423	32	1751
9	1123	21	1451	33	1778
10	1150	22	1478	34	1805
11	1178	23	1505	35	1833
12	1205	24	1532	36	1860

Success For All Foundation. (2007). *4Sight reading and math benchmarks 2006-07 technical report for Pennsylvania*. Baltimore, Maryland.

Appendix D: Pennsylvania 4Sight Mathematics Benchmark Assessment Descriptive Statistics and Validity Correlation to the 2007 Mathematics PSSA *(Data was unavailable for grade 8 at the time of SFA report)*

Table 6

7TH GRADE MATH FORM 2				
	N	**MEAN**	**SD**	**R**
4SIGHT	320	17.67	7.30	0.92
PSSA	320	1359.56	216.88	

Table 7

7TH GRADE MATH FORM 2 CORRELATION					
4Sight Total Score	**PSSA**	**4Sight Total Score**	**PSSA**	**4Sight Total Score**	**PSSA**
1	905	13	1232	25	1560
2	932	14	1259	26	1587
3	959	15	1287	27	1614
4	986	16	1314	28	1642
5	1014	17	1341	29	1669
6	1041	18	1369	30	1696
7	1068	19	1396	31	1724
8	1096	20	1423	32	1751
9	1123	21	1451	33	1778
10	1150	22	1478	34	1805
11	1178	23	1505	35	1833
12	1205	24	1532	36	1860

Success For All Foundation. (2007). *4Sight reading and math benchmarks 2006-07 technical report for Pennsylvania*. Baltimore, Maryland.

Appendix E: Pennsylvania 4Sight Mathematics Benchmark Assessment Descriptive Statistics and Validity Correlation to the 2008 Mathematics PSSA

Table 8

7TH GRADE MATH				
	N	**MEAN**	**SD**	**R**
4SIGHT	3236	20.97	7.95	0.91
PSSA	3236	1354.91	213.18	

Table 9

7TH GRADE MATH CORRELATION					
4Sight Total Score	**PSSA**	**4Sight Total Score**	**PSSA**	**4Sight Total Score**	**PSSA**
1	819	13	1112	25	1405
2	843	14	1136	26	1429
3	868	15	1161	27	1453
4	892	16	1185	28	1478
5	916	17	1209	29	1502
6	941	18	1234	30	1527
7	965	19	1258	31	1551
8	990	20	1283	32	1575
9	1014	21	1307	33	1600
10	1038	22	1331	34	1624
11	1063	23	1356	35	1649
12	1087	24	1380	36	1673

Success For All Foundation. (2008). *4Sight reading and math benchmarks 2007-2008 technical report for Pennsylvania*. Baltimore, Maryland.

Appendix F: 2008 Pennsylvania 4Sight Mathematics Benchmark Assessment Descriptive Statistics and Validity Correlation to the Mathematics PSSA

Table 10

8TH GRADE MATH				
	N	**MEAN**	**SD**	**R**
4SIGHT	3041	18.78	7.21	0.90
PSSA	3041	1348.38	186.55	

Table 11

8TH GRADE MATH CORRELATION					
4Sight Total Score	**PSSA**	**4Sight Total Score**	**PSSA**	**4Sight Total Score**	**PSSA**
1	886	12	1143	23	1400
2	909	13	1166	24	1424
3	933	14	1190	25	1447
4	956	15	1213	26	1470
5	979	16	1237	27	1494
6	1003	17	1260	28	1517
7	1026	18	1283	29	1540
8	1050	19	1307	30	1564
9	1073	20	1330	31	1587
10	1096	21	1353	32	1611
11	1120	22	1377	33	1634

Success For All Foundation. (2008). *4Sight reading and math benchmarks 2007-2008 technical report for Pennsylvania*. Baltimore, Maryland.

Appendix G: 2007 Pennsylvania 4Sight Mathematics Benchmark Assessment Pearson Correlation Analysis – Reliability

Table 12

MATHEMATICS, FIRST EDITION		
GRADE	**AVERAGE CORRELATION**	**AVERAGE N**
3	.76	14,000
4	.74	14,000
5	.77	15,000
6	.80	15,000
7	.81	17,500
8	.81	17,500

Success For All Foundation. (2007). *4Sight reading and math benchmarks 2006-07 technical report for Pennsylvania*. Baltimore, Maryland.

Appendix H: 2008 Pennsylvania 4Sight Mathematics Benchmark Assessment Pearson Correlation Analysis – Reliability

Table 13

MATHEMATICS, THIRD EDITION		
GRADE	**AVERAGE CORRELATION**	**AVERAGE N**
3	.78	20,300
4	.77	20,400
5	.79	21,200
6	.84	22,100
7	.84	21,800
8	.83	21,100

Success For All Foundation. (2008). *4Sight reading and math benchmarks 2007-2008 technical report for Pennsylvania*. Baltimore, Maryland.

Appendix I: Fall 2006 Pennsylvania 4Sight Mathematics Benchmark Assessment Predictive Validity with 2007 PSSA Scores

Table 14

7TH GRADE MATH				
	N	**MEAN**	**SD**	**R**
4SIGHT	7596	15.48	6.62	0.87
PSSA	7596	1385.04	214.27	

Table 15

8TH GRADE MATH				
	N	**MEAN**	**SD**	**R**
4SIGHT	7274	14.51	6.32	0.87
PSSA	7274	1371.47	188.94	

Success For All Foundation. (2008). *4Sight reading and math benchmarks 2007-2008 technical report for Pennsylvania*. Baltimore, Maryland.

Appendix J: PSSA Descriptive Statistics and Reliability Using Cronbaugh's Alpha Reliability Indices

Table 16

7TH GRADE 2006 PSSA				
STRAND	**N**	**MEAN**	**SD**	**R**
Overall	143471	39.67	13.41	0.92
A) Numbers and Operations	143471	9.47	3.99	0.75
B) Measurement	143471	4.92	2.29	0.66
C) Geometry	143471	8.31	2.63	0.70
D) Algebra	143471	9.07	3.74	0.72
E) Data Analysis and Probability	143471	7.89	2.81	0.66

Data Recognition Corporation. (2007a). *Technical Report for the Pennsylvania System of School Assessment 2006 Reading and Mathematics Grades 4, 6, and 7.* Retrieved December 25, 2008 from http://www.pde.state.pa.us/a_and_t/lib/a_and_t/2006_ReadingMathGr4_6_7_Tech_Report.pdf

Appendix K: PSSA Descriptive Statistics and Reliability Using Cronbaugh's Alpha Reliability Indices

Table 17

8TH GRADE 2006 PSSA				
STRAND	**N**	**MEAN**	**SD**	**R**
Overall	145655	42.22	13.75	0.93
A) Numbers and Operations	145655	8.43	3.10	0.77
B) Measurement	145655	4.87	2.59	0.66
C) Geometry	145655	7.63	2.86	0.66
D) Algebra	145655	12.49	4.39	0.79
E) Data Analysis and Probability	145655	8.81	2.61	0.74

Data Recognition Corporation. (2007b). *Technical Report for the Pennsylvania System of School Assessment 2006 Reading and Mathematics Grades 5, 8, and 11.* Retrieved December 25, 2008 from http://www.pde.state.pa.us/a_and_t/lib/a_and_t/2006_ReadingMathGr5_8_11_Tech_Report.pdf

Appendix L: PSSA Descriptive Statistics and Reliability Using Cronbaugh's Alpha Reliability Indices

Table 18

7TH GRADE 2007 PSSA				
STRAND	**N**	**MEAN**	**SD**	**R**
Overall	140692	40.427	13.261	0.93
A) Numbers and Operations	140692	8.203	3.138	0.773
B) Measurement	140692	5.603	2.519	0.691
C) Geometry	140692	8.258	2.757	0.718
D) Algebra	140692	10.403	3.756	0.763
E) Data Analysis and Probability	140692	7.960	2.786	0.746

Table 19

8TH GRADE 2007 PSSA				
STRAND	**N**	**MEAN**	**SD**	**R**
Overall	143430	42.496	13.747	0.931
A) Numbers and Operations	143430	7.541	3.243	0.678
B) Measurement	143430	6.663	2.582	0.757
C) Geometry	143430	7.600	2.483	0.646
D) Algebra	143430	12.734	4.093	0.810
E) Data Analysis and Probability	143430	7.958	3.114	0.730

Data Recognition Corporation. (2008). *Technical Report for the Pennsylvania System of School Assessment 2007 Reading and Mathematics Grades 3, 4, 5, 6, 7, 8, and 11.* Retrieved December 25, 2008 from http://www.pde.state.pa.us/a_and_t/lib/a_and_t/2007_PSSA_Reading_&_Mathematics_Tech_Report.pdf

Appendix M: Proposed Teacher Interview Coding Rubric

Since coding of open ended responses rely on the answers provided by interviewees, the researcher will have a better sense of coding once all interviews are transcribed. Depending on the transcribed responses, this coding rubric may need to be modified and therefore is subject to change. A second person will be used to code the transcribed interview responses in order to gain inter-rater reliability.

QUESTION NUMBER	CODING
Q1	0. 0-3 YEARS (NOT TENURED) 1. 4-9 YEARS (TENURED, FAIRLY NEW) 2. 10+ YEARS (VETERAN)
Q2	0. 1-2 YEAR – UNDER SUPERVISION OF PRINCIPAL C 1. 3-4 YEARS – UNDER SUPERVISION OF PRINCIPAL B & C 2. 5+ YEARS – UNDER SUPERVISION OF PRINCIPALS A, B, & C
Q3	0. NO 1. YES
Q4	0. NO 1. YES
Q5	0. TRAINING NOT PROVIDED 1. MINIMALLY TRAINED 2. ADEQUATELY TRAINED
Q6	0. TRAINING NOT PROVIDED 1. MINIMALLY TRAINED 2. ADEQUATELY TRAINED
Q7	0. NO

	1. YES
Q8	0. NOT KNOWLEDGEABLE AT ALL 1. SOMEWHAT KNOWLEDGEABLE 2. VERY KNOWLEDGEABLE
Q9	0. NO CONFIDENCE 1. LITTLE CONFIDENCE 2. VERY CONFIDENT
Q10	0. NO 1. YES
Q11	0. NO APPARENT CORRELATION TO 4SIGHT 1. MINIMAL CORRELATION TO 4SIGHT 2. MODERATE CORRELATION TO 4SIGHT 3. APPARENT CORRELATION TO 4SIGHT
Q12	0. 0% 1. 25% 2. 50% 3. 75% 4. 100%
Q13	0. NO 1. YES
Q13a	0. 0% 1. 25% 2. 50% 3. 75% 4. 100%
Q14	0. NO 1. YES

Q14a	0. LITTLE IMPACT 1. MODERATE IMPACT 2. SIGNIFICANT IMPACT
Q14b	0. NO IMPACT
Q15	0. NO 1. YES
Q15a	0. LITTLE CHANGE 1. MODERATE CHANGE 2. SIGNIFICANT CHANGE
Q15b	0. NO CHANGE
Q16	0. NO 1. YES
Q16a	0. LITTLE EMPOWERMENT 1. MODERATE EMPOWERMENT 2. GREAT EMPOWERMENT
Q16b	0. NO EMPOWERMENT
Q17	0. DO NOT MEET 1. 1-3 MEETINGS 2. 4-6 MEETINGS 3. 7-9 MEETINGS 4. 10+ MEETINGS
Q17a	0. ADMINISTRATORS ONLY 1. TEACHERS ONLY 2. TEACHERS AND ADMINISTRATORS
Q18	0. PRINCIPAL-PROVIDED REPORTS 1. OTHER PROFESSIONAL-PROVIDED REPORTS 2. INTERVIEWEE-PROVIDED REPORTS

Q19	0. DO NOT MEET 1. 1-3 MEETINGS 2. 4-6 MEETINGS 3. 7-9 MEETINGS 4. 10+ MEETINGS
Q19a	0. ADMINISTRATORS ONLY 1. TEACHERS ONLY 2. TEACHERS AND ADMINISTRATORS
Q20	0. NO 1. YES
Q20a	0. REPORTS SENT HOME QUARTERLY, NO OTHER COMMUNICATIONS 1. DISCUSSED RESULTS WITH STUDENTS AND SENT HOME FOR PARENTS TO REVIEW 2. DISCUSSED AND COLLABORATED (QUARTERLY) WITH STUDENTS TO IMPROVE STUDENT ACHIEVEMENT 3. ON-GOING DISCUSSION AND COLLABORATION WITH STUDENTS TO IMPROVE STUDENT ACHIEVEMENT
Q21	0. NO 1. YES
Q21a	0. REPORTS SENT HOME QUARTERLY, NO OTHER COMMUNICATIONS 1. DISCUSSED RESULTS WITH PARENTS AND SENT HOME FOR PARENTS TO REVIEW 2. DISCUSSED AND COLLABORATED (QUARTERLY) WITH PARENTS TO IMPROVE STUDENT

	ACHIEVEMENT 3. ON-GOING DISCUSSION AND COLLABORATION WITH PARENTS TO IMPROVE STUDENT ACHIEVEMENT
Q22	0. NO IMPACT 1. LITTLE IMPACT 2. MODERATE IMPACT 3. SIGNIFICANT IMPACT
Q23	0. NO EFFECT 1. LITTLE EFFECT 2. MODERATE EFFECT 3. SIGNIFICANT EFFECT
Q24	0. NO EFFECT 1. LITTLE EFFECT 2. MODERATE EFFECT 3. SIGNIFICANT EFFECT
Q25	0. NO CHANGE 1. LITTLE CHANGE 2. MODERATE CHANGE 3. SIGNIFICANT CHANGE
Q26	0. NO 1. YES
Q26a	0. LITTLE IMPACT 1. MODERATE IMPACT 2. SIGNIFICANT IMPACT
Q26b	0. NO IMPACT
Q27	0. NO 1. YES

Q27a	0. DON'T CARE 1. WITH RESERVATIONS 2. WITHOUT RESERVATIONS 3. STRONGLY RECOMMEND
Q27b	0. DON'T CARE 1. WITH RESERVATIONS NOT TO CONTINUE USE 2. WITHOUT RESERVATIONS NOT TO CONTINUE USE 3. STRONGLY RECOMMEND NOT TO CONTINUE USE