An examination of the Pennsylvania 4Sight Benchmark Assessments as predictors of Pennsylvania System of School Assessment performance.

An Examination of the Pennsylvania 4Sight Benchmark Assessments as Predictors of Pennsylvania System of School Assessment Performance

by

Christina K. Lutz-Doemling

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Doctor of Education

in

Educational Leadership

Lehigh University

September 14, 2007

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

ProQuest LLC
789 E. Eisenhower Parkway
PO Box 1346
Ann Arbor, MI 48106-1346

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

September 14, 2007
Date

George P. White, Ed.D.
Dissertation Director

September 14, 2007
Accepted Date

Committee Members:

Jill Sperandio, Ph.D.

H. Lynn Columba, Ed.D.

GwenCarol Holmes, Ed.D.

ACKNOWLEDGMENTS

I am thrilled to finally complete the very long, challenging and rewarding journey as a Lehigh University doctoral student. It is a pleasure to express my sincere gratitude to the individuals who have supported, encouraged, and guided me throughout my pursuit of the Doctor of Education degree.

First, I would like to thank Dr. George White, Committee Chairman, for his genuine enthusiasm and support during this process. Dr. White's positive attitude and ongoing encouragement motivated me to persevere through the many obstacles that I encountered as I completed my research. Additionally, I would like to recognize committee members Dr. H. Lynn Columba, Dr. Jill Sperandio, and Dr. GwenCarol Holmes for their guidance and feedback. I am fortunate to have worked with such a knowledgeable and caring doctoral committee.

I would also like to thank Dr. Lorie Davis, my former colleague at Whitehall-Coplay School District, for encouraging me to enroll in the Educational Leadership Doctoral Program at Lehigh University early in my educational career. A special thank you goes to Robert Spengler and Kathy Kotran, my colleagues at Catasauqua Area School District, for their continued support.

A sincere thank you goes to Dr. Bridget O'Connell, a fellow doctoral student. From the moment we met in the summer Curriculum Management workshop, I knew that we would develop a lasting friendship. The weekends I devoted to dissertation writing at the Lehigh University Library were much more enjoyable when she was there to provide feedback, crack jokes, and enjoy coffee.

Most importantly, I would like to thank my family for their unwavering support and continued motivation as I completed my degree requirements. The love and support of my husband, Drew, daughter, Cassandra, parents, Ron and Audrey, and brothers, Ron, Jon, and Jeff, kept me going especially at times when I questioned my commitment to writing my dissertation. Thank you for helping me see the light at the end of the tunnel.

TABLE OF CONTENTS

CHAPTER I

Overview

CHAPTER II

Review of the Literature

CHAPTER IV

Results

CHAPTER V

Summary, Interpretations, and Recommendations

Abstract

The primary purpose of this study was to determine whether there was a significant relationship between sixth grade students' 4Sight Benchmark Assessment scores and their sixth grade PSSA scaled scores in both reading and mathematics. The secondary purpose of the study was to determine whether participation in the 4Sight Benchmark Assessments led to greater improvement on PSSA scores from fifth to sixth grade.

The sample for this study consisted of a total of six school districts. The three experimental school districts were match paired with three control school districts according to district enrollment, percentage of low-income families, and school locale code.

The Pearson correlation between students' sixth grade 4Sight Reading Assessment scores and their sixth grade Reading PSSA scores was determined to be positive and significant ($r = .77$, $p < .0005$). Additionally, the correlation between students' sixth grade 4Sight Mathematics Assessment scores and their sixth grade Mathematics PSSA scores was also determined to be positive and significant ($r = .77$, $p < .0005$). In relation to the study's secondary purpose, the experimental group of students who participated in the 4Sight Benchmark Assessments showed greater improvement on both the Reading and Mathematics PSSA than did the students in the control group who did not participate in the 4Sight Benchmark testing.

Several factors may explain why the 4Sight Benchmark Assessments were associated with small to moderate effect sizes. First, the 4Sight Benchmark Assessments were designed to be one component of a comprehensive school improvement process, not

as an intervention strategy in isolation. Additionally, 2005-2006 marked the first year of 4Sight Benchmark Assessment implementation in the experimental districts. The low to moderate effect sizes determined as a result of this study may be explained by the early stages of implementation of the 4Sight Assessment initiative. During the first year of implementation, it is likely that administrators and teachers spent the majority of their time familiarizing themselves with assessment administration and scoring practices and little time discussing how to utilize the performance results to appropriately modify instruction and positively influence student achievement.

CHAPTER I

Introduction

> Our nation, which has prevailed in conflict after conflict over several centuries, now faces a stark and sudden choice: adapt or perish. I'm not referring to the war against terrorism but to a war of skills – one that America is at a risk of losing to India, China, and other emerging economies. And we're not at risk of losing it on factory floors or lab benches. It's happening every day, all across the country, in our public schools. Unless we transform those schools and do it now . . . it will soon be too late. (Gerstner, 2004, p. A18)

The development of human capital in America is critical to ensuring the future success of high school graduates in the highly competitive global economy of the 21st Century (Hershberg, 2005). Presently, Americans are at risk of losing their jobs due to globalization. Hershberg (2005) reports that China and India consistently produce more college graduates and more engineers per year than America, and only two out of every ten American students who begin high school graduate from college with bachelor degrees. United States Secretary of Education Margaret Spellings recognized the need for school reform and recently remarked that globalization and America's diminishing competitiveness have resulted in the need to raise achievement levels and close the achievement gap (Spellings, 2005).

No Child Left Behind

In response to the human capital development threat, the United States federal government expanded its role in public education by passing with bipartisan consensus the No Child Left Behind Act (2002). By the year 2014, No Child Left Behind (NCLB)

requires that all students attending public schools will demonstrate proficiency in reading and mathematics. As stated by President George W. Bush (2003):

> The time for excuse-making has come to an end. With the No Child Left Behind Act, we have committed the Nation to higher standards for every single public school . . . Accountability for results is no longer just a hope of parents. Accountability for results is the law of the land. (United States Department of Education, 2003, ¶ 1)

At the state level, NCLB requires the development and implementation of accountability plans that include the administration of summative assessments aligned to state academic standards to illustrate student achievement in mathematics and reading (United States Department of Education, 2003). In the year 2005-2006, NCLB mandated an increased emphasis on state-level assessment. Prior to 2005-2006, school districts administered state mathematics and reading assessments in three grade levels: one elementary grade level, one middle level grade, and one high school grade level. In 2005-2006, districts were required to assess students in every grade from three through eight as well as in one high school grade level. Further increases in testing will occur in 2007-2008 when NCLB requires the implementation of state science assessments at the elementary, middle, and high school levels.

No Child Left Behind further requires local level accountability in that school districts and their students in all demographic subgroups must demonstrate adequate yearly progress toward the goal of 100% proficiency on state assessments in reading and mathematics by the year 2014. Failure to demonstrate adequate yearly progress on state assessments may result in consequences at the district or school level. Districts and/or

schools failing to meet proficiency targets may be designated in need of improvement or corrective action. Specific consequences may include the development of an improvement plan, school choice options for parents, supplemental educational services for students, significant changes in curriculum, leadership or professional development, and significant changes in school governance such as reconstitution, privatization, or chartering (NCLB, 2002). Therefore, state assessments have become a high-stakes adventure for schools and districts.

Statement of the Problem

No Child Left Behind (2002) triggered an increased emphasis on state assessment and accountability for results. State assessment systems vary greatly in their design, content, and stakes; however, all state systems of assessment have the same goal of improving instruction and positively impacting student learning (NCLB, 2002). There is ongoing debate as to the degree to which state assessment systems effectively contribute to the goal of increasing student achievement.

Herman (2005) recognized the results of numerous qualitative and quantitative state studies that suggest there are three benefits to annual state assessment. The identified benefits include focused instruction, teacher modeling, and improved test scores. More specifically, Herman reported that research has shown that state assessments have stimulated classroom teachers to deliver instruction aligned to content standards and model the content and format of state assessments within the classroom. Additionally, there is evidence to support that student achievement on state assessments has increased over the years (Florida Department of Education, 2005; Pennsylvania Department of Education, 2005; Texas Education Agency, 2005; Herman & Perry, 2002).

In contrast, a recent study of student achievement since the implementation of NCLB (Hall, 2005), illustrated that although students in elementary grade levels have demonstrated progress toward mathematics and reading proficiency, middle school and high school results are lagging behind. In addition, there continues to be evidence of achievement gaps when examining the performance of various demographic subgroups of students (Hall, 2005). Critics of summative state assessments argue that end-of-year assessments do not provide the timely, diagnostic data necessary to modify classroom instruction and improve student achievement (Herman & Baker, 2005; Linn, 2001). Consequently, teacher efforts to utilize summative assessment data to modify and adapt the instructional program to target specific student needs on a day-to-day basis are often challenging and sometimes unsuccessful. Researchers (Shepherd, 2000; Assessment Reform Group, 1999; Black & William, 1998) have asserted that although annual accountability testing can be used to evaluate instruction, once-per-year testing does not provide the detailed information needed to genuinely increase student learning that can be provided through formative assessment.

Linn, Baker, and Dunbar (1991) suggest that ongoing, standards-based student performance data must be provided to educators in a timely fashion if student performance data is to be used effectively to inform teaching and learning. A synthesis of formative assessment research (Black & William, 1998) identified support for formative assessment as an effective method to provide teachers and students with the required ongoing performance data that is necessary to successfully modify instruction and increase achievement. Black and William report that teachers benefit from the ongoing student performance data provided through formative assessment by making adjustments

to classroom instruction and modifying students' learning goals. Similarly, McTighe and O'Connor (2005) suggest formative assessments "provide specific feedback to teachers and students for the purpose of guiding teaching to improve learning" (p. 12). Research supports teacher use of corrective, timely, and criterion-referenced feedback to increase student learning (Marzano, Pickering & Pollock, 2001) and strengthen the links between classroom curriculum, instruction, and assessment (Guskey, 2005).

Need for the Study

Because summative state assessments "rarely provide the ongoing information that schools need to guide instructional programs and address the learning problems of students who might otherwise be left behind" (Herman & Baker, 2005, p. 48), assessment experts have recognized the importance of classroom formative assessment. Bernhardt (2004) emphasized the necessity of utilizing a combination of both summative and formative assessment, "...in a perfect world, schools would use both formative and summative assessments to ensure that all students are learning. If only summative assessment data are studied, however, solutions for improving the scores can come out half-baked" (p. 1). The implementation of standards-based formative assessments can provide increased opportunities for data collection. Educators can analyze the data and address instructional problems on an ongoing basis before the end of the year when it is too late to do anything about failure with the exception of holding children back. Hershberg (2005) opined:

> The mediocre, high-stakes standardized tests found in the large majority of states need to be replaced with a new integrated assessment system that would provide not only a high-quality "summative" exam at year's end focused on the

development of higher-order thinking skills, but "formative" assessments throughout the school year designed to give teachers regular feedback in the form of suggested pedagogical interventions to support improved instruction for this year's students. (p.4)

School districts are increasingly using formative assessments to periodically provide teachers with student performance data, gauge student progress toward grade level academic standards, and guide instructional modifications (Olson, 2005). These formative assessments, often labeled "benchmark assessments" are intended to predict student performance on high-stakes state assessments. By their design, benchmark assessments provide timely data to classroom teachers and their students, and shift the focus of the assessment information from policymakers and governmental officials to the learner and classroom teacher.

While assessment experts suggest that utilizing benchmark assessments to monitor student learning and modify instruction may assist districts in their preparation for state summative assessments (Reeves, Slavin & Wiggins, 2005 as cited in Olson, 2005), only limited research has been conducted regarding the utilization of benchmark assessments as predictors of student performance on state assessments (Success for All Foundation 2005-2006; The School Board of Broward County, Florida, 2005; Morris, 2004; Lauman 2001).

Purpose of the Study

The Success for All Foundation recently created benchmark assessments in reading and mathematics that are aligned to several states' summative assessments. These benchmark assessments were developed to provide states with a formative system of

evaluation of student, classroom, grade level, and school progress toward academic standards. Additionally, the benchmark assessments are touted to serve as a predictive measure of each student's performance on the state assessment (Success for All Foundation, 2005).

This Pennsylvania study seeks to affirm or refute the limited 4Sight Benchmark Assessment research (Success for All Foundation, 2005-2006) concerning the use of the Pennsylvania benchmark assessments as predictive measures of student performance on the Pennsylvania System of School Assessment (PSSA). The primary purpose of this study is to determine the effectiveness of utilizing the Pennsylvania 4Sight Benchmark Assessments to predict student performance on the Reading and Mathematics PSSA. First, the study will determine if there is a significant relationship between student performance data on the Reading and Mathematics 4Sight Benchmark Assessments and the Reading and Mathematics PSSA. Second, the study will determine if there is any significant difference between the change on Reading and Mathematics PSSA performance means for the students who participated in the 4Sight Reading and Mathematics Benchmark Assessments as compared with students that did not participate in the 4Sight Benchmark testing. These determinations will be valuable to local and state officials as they make decisions to begin, continue, or discontinue the use of the Pennsylvania Benchmark Assessments within school districts across the Commonwealth.

Research Questions

The following research questions guide this study:

1. Is there a significant relationship between the 4Sight Reading Benchmark Assessment Scores and PSSA Reading scaled scores at the sixth grade level?

2. Is there a significant relationship between the 4Sight Mathematics Benchmark Assessment Scores and PSSA Mathematics scaled scores at the sixth grade level?
3. Is there any significant difference in the change on Reading PSSA performance means from fifth to sixth grade for the students who participated in the 4Sight Reading Benchmark Assessments as compared with a matched group that did not participate in the 4Sight Reading testing?
4. Is there any significant difference in the change on Mathematics PSSA performance means from fifth to sixth grade for the students who participated in the 4Sight Mathematics Benchmark Assessments as compared with a matched group that did not participate in the 4Sight Mathematics testing?

Methodology

The research study will utilize a quasi-experimental design. More specifically, a convenience sample of three school districts implementing the 4Sight Benchmark Assessment testing at the sixth grade level will be match paired with a convenience sample of three other school districts not utilizing the 4Sight Benchmark Assessments. The districts will be match paired according to the following demographic characteristics: district enrollment, percentage of low income families by local education agency, and school locale code.

To determine whether a significant relationship exists between 4Sight Benchmark Assessment Reading scores and Reading PSSA scaled scores at the sixth grade level, sixth grade students' 2006 4Sight Benchmark Assessment Reading scores will be correlated with their 2006 Reading PSSA scaled scores in the three school districts utilizing the 4Sight Benchmark Assessments. Similarly, sixth grade students' 2006 4Sight

Benchmark Assessment Mathematics scores will be correlated with their 2006 Mathematics PSSA scaled scores to determine whether a significant relationship exists between 4Sight Benchmark Assessment Mathematics scores and Mathematics PSSA scaled scores in the same three school districts.

To determine whether there is any significant difference between change on the Reading and Mathematics PSSA performance means for the sixth grade students who participated in the 4Sight Reading Benchmark Assessments as compared with a matched group that did not participate in the 4Sight Reading testing, the students' fifth grade (2005) and sixth grade (2006) Reading and Mathematics PSSA scaled scores will be transformed to *T* scores. A repeated measures analysis of variance (ANOVA) will then be conducted to determine whether there is a significant difference in the change on Reading and Mathematics PSSA performance means for the experimental group versus the control group.

Definitions of Terms

Adequate Yearly Progress – An individual state's measure of yearly progress toward achieving state academic standards. "Adequate Yearly Progress" is the minimum level of improvement that states, school districts and schools must achieve each year (United States Department of Education, 2004).

Benchmark Assessments – Periodically administered tests "designed primarily to predict students' performance on end-of-the-year state exams. They measure the same set of knowledge and skills at several points during the school year to see if students are making progress and to provide and early warning of potential problems" (Slavin, 2005 as cited in Olson, 2005).

Criterion-Referenced Assessment – "An assessment where an individual's performance is compared to a specific learning objective or performance standard and not to the performance of other students. Criterion-referenced assessment tells us how well students are performing on specific goals or standards rather that just telling how their performance compares to a norm group of students nationally or locally. In criterion-referenced assessments, it is possible that none, or all, of the examinees will reach a particular goal or performance standard" (CRESST Assessment Glossary, 1999).

Formative Assessment – "The gathering of information about student learning-during the progression of a course or program and usually repeatedly-to improve the learning of those students" (Leskes, 2002).

Norm-Referenced Assessment – "An assessment where student performance or performances are compared to a larger group. Usually the larger group or "norm group" is a national sample representing a wide and diverse cross-section of students. Students, schools, districts, and even states are compared or rank-ordered in relation to the norm group. The purpose of a norm-referenced assessment is usually to sort students and not to measure achievement towards some criterion of performance" (CRESST Assessment Glossary, 1999).

Summative Assessment – "The gathering of information at the conclusion of a course, program, or undergraduate career to improve learning or to meet accountability demands" (Leskes, 2002).

4Sight Assessments – One-hour benchmark assessments developed by the Success for All Foundation that are aligned to state assessments and designed to produce

subskill reports and overall scores predictive of students' scores on state assessments (Success for All Foundation, 2005).

Limitations of the Study

The following statements are offered as possible limitations to the Pennsylvania benchmark assessment study:

1. The unit of analysis for this study is a quasi-experimental, matched pair, convenience sample in that participating schools were solicited from a particular geographic area in Pennsylvania, which may not be representative of the general population of students. The results of this study may not be generalizeable to other students or to other states.
2. The study is limited in scope in that the researcher only examined the reading and mathematics results of sixth grade students. The results of this study will not be generalizeable to other grade levels.
3. Other factors within the schools chosen for this study such as curriculum, instruction, and professional development may have influenced student performance results.
4. The number of 4Sight assessments administered in each district varies and may influence the results of the study.
5. Researcher bias may have influenced the study—the researcher is employed as the Director of Curriculum and Assessment in one of the participating districts. Additionally, within this district, the researcher is responsible for the implementation of the 4Sight Benchmark Assessment initiative.

6. The data presented in this study resulted from a one year research study. A longitudinal research study may provide more detailed information pertaining to the effectiveness of the 4Sight Benchmark Assessments.

7. No urban school districts were included in this research. While the findings of this study seem significant, one should be cautious about generalizing the results to urban schools.

CHAPTER II

Review of the Literature

Empirical studies examining the general use of summative and formative student assessment in educational environments are more extensive than studies specific to benchmark assessments as predictors of student performance on state assessments. This review will provide a brief analysis of the major findings among the empirical studies of educational assessment. An analysis of the more significant studies pertaining to educational assessments as predictors of performance on state assessments with a focus on predictors of performance on the Pennsylvania System of School Assessment (PSSA) will follow. Additionally, the final section of the review of the literature will include a critical analysis of the Pennsylvania 4Sight Benchmark Assessment Technical Report (2005-2006), which warrants special consideration due to its relevance to the present study.

Assessment

Empirical, educational assessment studies may be classified as research pertaining to "assessment of learning" or "assessment for learning." Stiggins (2002) defines assessments of learning as one-time-per year summative assessments typically utilized to provide the evidence of achievement necessary for public accountability reporting. Because of the purposes for which governmental officials use the student performance data from state assessments, Stiggins considers high-stakes state assessments to be assessments of learning (2002). On the other hand, Stiggins describes assessment for learning as teacher use of low-stakes, formative assessments within the classroom to provide students with ongoing information about their own learning. Similarly, Heubert

and Hauser (1999) have identified fundamental differences between "low-stakes" and "high-stakes" assessments:

> A low-stakes test has no significant, tangible, or direct consequences attached to the results, with information alone assumed to be a sufficient incentive for people to act . . . actions based on tests results will improve educational quality and student achievement. In contrast, high-stakes policies assume that information alone is insufficient to motivate educators to teach well and students to perform to high standards. Hence, it is assumed, the promise of rewards or the threat of sanctions is needed to ensure change. Rewards in the form of financial bonuses may be allocated to schools or teachers; sanctions may be imposed through external oversight or takeover by higher-level authorities. (pp. 35-36)

This review will begin with a synthesis of the empirical assessment research concerning the positive and negative teacher and student effects of both high-stakes, assessments of learning and low-stakes, assessments for learning.

Assessments of Learning

There are arguments both supporting and opposing state implementation of high-stakes, state assessments of learning. Supporting arguments suggest that high-stakes tests: focus classroom instruction, increase teacher accountability, increase student accountability, increase student motivation, and increase student self-efficacy (Amrein & Berliner, 2002; Clarke, et. al., 2003; Manhattan Institute for Policy Research, 2003; National School Boards Association, 2004a, 2004b; New York State Education Department, 2004; Pedulla, et. al., 2003). Critics of high-stakes testing typically argue that high-stakes tests: narrow the curriculum, exclude curricular topics that are not

included on high stakes tests, result in test preparation rather than increased learning, affect students' attitudes and motivation negatively, increase drop-out rates, and widen the achievement gap (Clarke, et. al., 2003; Horn, 2003; Kober, 2002; National School Boards Association, 2004, 2004b; New York State Education Department, 2004; Pedulla, et. al., 2003; Yeh, 2005). The following sections will present the empirical data and resulting conclusions from the studies that investigated the high-stakes testing effects on teachers and students.

Teacher Effects. Although it is frequently reported that high-stakes tests negatively affect teachers because they result in narrowing the curriculum and emphasizing test preparation, there is little empirical evidence to support these claims (Cizek, 2001; Cimbricz, 2002; New York State Education Department, 2004; Yeh, 2005). Research pertaining to teacher perceptions of high-stakes testing provides a basis to begin to determine the effects of high-stakes testing on teachers (Berger & Elson, 1996; Clarke, et. al., 2003; Corbett & Wilson, 1990, 1991; Herman & Golan, n.d.; Mabry, Poole, Redmond, & Schultz, 2003; Pedulla, et. al., 2003; Rottenberg & Smith, 1990; Shephard & Dougherty, 1991).

Langenfeld, Thurlow, and Scott (1997) conducted a literature review on the effects of high-stakes testing on the curriculum. Their review included six state studies reporting teacher perceptions of high-stakes testing effects on the curriculum (Berger & Elson, 1996; Corbett & Wilson, 1990, 1991; Herman & Golan, n.d.; Rottenberg & Smith, 1990; Shephard & Dougherty, 1991). From their review, Lagenfeld, Thurlow, and Scott concluded that high-stakes testing resulted in teachers delivering a modified curriculum and emphasizing test preparation activities. However, teachers disagreed as to whether

the curricular modifications were positive or negative. Teachers considered curricular changes resulting in a more focused curriculum including relevant content and skills to be positive. On the other hand, they considered changes resulting in a narrowed curriculum excluding important non-tested content and skills to be negative.

In a more recent review of the literature on state-mandated testing and teacher beliefs and practices, Cimbricz (2002) reported that the majority of the existing literature concerning high-stakes testing consisted of theory and rhetoric. In her review, she excluded non-empirical works and focused on the "handful" of eligible studies that met qualitative and quantitative research standards. Cimbricz's review piggybacked upon the results of Langenfeld, Thurlow, and Scott (1997) and suggested that high-stakes testing both positively and negatively influenced teacher beliefs and practices. Furthermore, she recognized various influential factors, such as teachers' beliefs, knowledge, status, and grade-level taught, as well as district-level and building-level expectations, in determining the effects of high-stakes testing on teachers. Cimbricz also noted that the majority of the research included in her review consisted of teacher interviews and surveys resulting in data illustrating subjective teacher perceptions of high-stakes testing. However, very little research included objective data collected through classroom observations. For this reason, Cimbricz concluded that high-stakes tests consistently influence teacher beliefs and practice; however, the relationship is not clearly positive or negative and requires further observational research to determine the extent to which confounding variables influence teacher beliefs and practice.

A recent multi-state study sampling a total of 360 educators from the states of Kansas (low-stakes), Michigan (mid-stakes), and Massachusetts (high-stakes) sought to

determine the effects of state assessments with varying stakes on classroom practice (Clarke, et. al., 2003). The study resulted in similar findings to those in the aforementioned literature reviews. Teachers identified both positive and negative effects on curriculum, instruction, and assessment. More specifically, Massachusetts educators preparing for the state's high-stakes exam identified the most positive and negative effects on classroom practice. Teachers in the low-stakes assessment state of Kansas reported the fewest test-related effects on curriculum, instruction, and assessment. Because the study included states with varying stakes attached to their state assessments, Clarke et. al. (2003) reported a noteworthy inference from the data: "...as the stakes increase, so too do the consequences for classroom practice, making it imperative that the test is aligned with the standards and is a valid and reliable measure of student learning" (p. 9).

In a second study of teacher perceptions regarding state testing programs with varying stakes, Pedulla et. al. (2003) provided a national view of teacher reported state testing effects on classroom practice. Consistently, teachers in states mandating high-stakes assessments reported increased narrowing of the curriculum, increased alignment of practices with the state assessment, increased classroom time devoted to covering tested content versus non-tested content.

Numerous studies have illustrated positive and negative teacher reported effects due to the implementation of high-stakes state assessments. Teacher perceptions provide necessary, but not sufficient information to determine the comprehensive effects of high-stakes testing. Researchers must also study the student effects correlated with high-stakes

assessment to present a more thorough examination of the wide-ranging effects of state assessment initiatives.

Student Effects. Beginning in 2006, No Child Left Behind (2002) required increased testing of public school students. Consequently, all students in grades three through eight and grade eleven must participate in state assessments. As the number of students being tested across the nation increases, research concerning the effects of high-stakes state testing on students becomes more critical (Horn, 2003). To provide insight into this body of research, the following section will present the results of empirical studies examining the effects of high-stakes testing on student attitudes and motivation, dropout rate and achievement.

Numerous studies examining high-stakes testing effects on students are based on teacher and administrator perceptions (New York State Education Department, 2004). Smith and Rottenberg (1991) interviewed 19 teachers in two low-income, ethnically diverse schools and discovered that teachers perceived that assessments, both high and low-stakes, negatively affected younger students. More specifically, the teachers reported their beliefs that students suffered from increased stress, fatigue, and illness as a result of testing. In Kentucky, one-third of teachers surveyed reported that student morale decreased after the implementation of the statewide assessment system (Koretz et. al., 1996). Similarly, almost two-thirds of North Carolina teachers and administrators surveyed by the state teacher association reported students experienced increased stress due to the implementation of high-stakes testing (Hargrove, Jones, & Jones, 2000).

In a more recent, multi-state study, teachers interviewed in Massachusetts, Michigan, and Kansas reported "more negative than positive test-related effects on

students, such as test-related stress, unfairness to special populations, and too much testing" (Clarke et. al., 2003, p. 11). The greatest number of negative effects on students as a result of testing was reported by the teachers in the high-stakes testing state, Massachusetts. The least number of negative effects on students were reported by teachers in the low-stakes testing state, Kansas. Teachers typically report negative perceptions of the effects of high-stakes testing on students.

Although teachers and administrators report high-stakes testing negatively affects students, empirical evidence gathered directly from students in Arkansas suggests otherwise. The Arkansas study of 283 fifth grade students combined data from student questionnaires with both low-stakes norm-referenced test scores and the state's high-stakes examination scores (Mulvenon et. al., 2001). The results of the research illustrated that students experience very few to no negative effects as a result of testing. However, further research must be conducted to determine whether the results of the study conducted with fifth grade students in one school district in Arkansas can be generalized to a larger population of students (National School Boards Association, 2004).

In addition to examining the effects of high stakes testing on teacher and student attitudes and motivation, the Center on Education Policy (2003) organized a task force to review the limited and inconclusive research concerning the effects of high stakes exit exams on student dropout rates. Although the resulting review of the literature identified studies that correlated the implementation of high-stakes assessments with increased dropout rates (Amrein & Berliner, 2002; Jacob, 2001; Warren & Edwards, 2003) as well as studies that found no relationship between the exams and dropout rates (Carnoy & Loeb, 2003; Davenport et. al., 2002), the task force concluded "that there is moderately

suggestive evidence, to date, of exit exams causing more students to drop out of school" (Center on Education Policy, 2003, p. 4).

There are several limitations of the existing research examining the effects of high-stakes exit exams on dropout rates (Center on Education Policy, 2003). First, dropout calculations differ across studies. One common dropout formula does not exist for use within schools or research studies. Therefore, discrepancies result when researchers calculate dropouts using different formulas. Second, there are concerns with current sources of dropout data such as the National Educational Longitudinal Study (NELS) and the Common Core of Data (CCD). These data sources may contain data not applicable to today's current assessment system in addition to data that does not accurately track students over time. Finally, historical information regarding changes in testing policies over time is often difficult to access from state departments of education. Historical information documenting changes in the policies related to high-stakes is an important consideration when conducting dropout data analysis.

The effect of high-stakes testing on student achievement is of critical importance because one of the main goals of NCLB is to increase academic achievement for all students. "Evaluating the thesis that high-stakes tests have had an effect on achievement is a pivotal one since this is one of the critical "gap closing" assumptions on which the standards and high-stakes testing movement is based" (New York State Education Department, 2004, p. 4). One method to determine whether the implementation of high-stakes state assessments has resulted in increased student achievement is to compare student achievement scores on a low-stakes assessment before and after the implementation of the high-stakes, state assessment. If student achievement on the low-

stakes assessment increases after the implementation of a high-stakes assessment policy, the data suggests a positive correlation between the implementation of high-stakes assessments and student achievement. In contrast, if student achievement on the low-stakes assessment remains the same or decreases, the utilization of high-stakes assessments to increase student achievement would require further research to determine the effectiveness of this practice (Amrein & Berliner, 2002)

Several studies investigate the relationship between high-stakes assessment and student achievement on low-stakes assessments (Amrein & Berliner, 2002; Bishop, 2001; Braun, 2004; Koretz, Linn, Dunbar, & Shepard, 1991; Raymond & Hanushek, 2003; Rosenshine, 2003); however it is unclear whether the implementation of high-stakes testing is positively correlated with increased student achievement.

One single district study (Koretz et al., 1991) determined that third grade student performance on one urban district's high-stakes, commercial achievement test did not generalize well to other forms of assessment. More specifically, the study's findings suggested that the 840 students' third grade math scores did not generalize to the other parallel or alternate forms of assessment, whereas reading scores generalized to a small degree. Consequently, the researchers concluded "to a substantial degree, teachers in this district must be focusing on content that is specific to the particular test used for accountability, rather than trying to improve achievement in the broader sense that we would all desire" (p. 16).

Similarly, the results of a more recent study (Amrein & Berliner, 2002) illustrated that state assessment results did not transfer to scores on national tests and therefore concluded that state implementation of a high-stakes testing program is not an effective

method to increase student achievement. Amrein and Berliner (2002) first investigated whether student academic achievement increased after the implementation of high-stakes testing. The multi-state study focused on the 27 states with the highest stakes written into their state assessment policies. Amrein and Berliner examined the high-stakes states National Assessment of Educational Progress (NAEP) scores in eighth grade mathematics, fourth grade mathematics, and fourth grade reading over a four year period and compared these scores to the average NAEP scores for all states that took the national assessment during the four year period. Amrein and Berliner concluded that there was no consistent evidence to support that implementation of high-stakes testing policies results in increased student achievement.

On the contrary, critics (Raymond & Hanushek, 2003; Rosenshine, 2003) of Amrein and Berliner's (2002) research methodology and conclusions conducted studies correlating high-stakes assessments with increased student achievement. The main criticism of Amrein and Berliner's research methodology is the lack of a comparison group to which they could compare the results from the high-stakes testing states. Raymond and Hanushek (2003) explain "the truly fatal flaw" of Amrein and Berliner's methodology:

> If one wants to assess the effect of high-stakes testing, the obvious comparison is between states that adopted accountability systems and those that did not. Amrein and Berliner's decision instead to compare the gains in high-stakes states with the national average violates a basic principle of social-science research. The national gain on NAEP incorporates any gains in high-stakes states, so Amrein and Berliner's strategy is akin to a medical trial where the treatment group receives

> the full dose of a medication while the control group receives a half-dose. It would not be surprising to find that the full dose was not dramatically more effective. The real question is whether the full dose is more effective than no medication at all. (¶ 10)

Raymond and Hanushek re-examined Amrein and Berliner's data in a systemic manner in an attempt to address their concerns with the original study's methodology. In Raymond's and Hanushek's multi-state, longitudinal study of fourth and eighth grade students, the data suggested that the states that made greater gains on the NAEP in mathematics from 1996 to 2000 were the states with accountability systems that reward or sanction schools for their academic performance. High-stakes testing states outperformed states without accountability systems. Likewise, Rosenshine (2003) compared NAEP mathematics and reading scores in high-stakes testing states to scores in states without high-stakes tests. The data suggested that the students were able to transfer their learning in mathematics and reading from the state assessment to the NAEP. Consequently, Raymond, Hanushek and Rosenshine concluded that the implementation of high-stakes testing programs resulted in increased student achievement as measured by a national assessment. Moreover, the most recent high-stakes assessment research (Braun, 2004; Swanson, 2006) continues to support the link between the implementation of high-stakes assessment programs, standards-based education policies, and increased student achievement as measured by the NAEP.

Braun (2004) conducted an extended reanalysis of Amrein and Berliner's (2002) data designed to improve upon the Amrein and Berliner study in several ways. First, Braun's reanalysis examined data not only from high-stakes testing states, but also states

without high-stakes assessment policies. Next, Braun improved the replication study by including a more comprehensive measure of states' educational reform efforts and an improved interpretation of state gain statistics. Finally, Braun examined the data for grades 4 and 8 during a more comparable timeframe, 1992-2000. When utilizing NAEP scores to determine relative gains over time at grades 4 and 8, Braun's conclusions differed from Amrein and Berliner (2002). Braun concluded that high-stakes testing states demonstrated increased academic achievement over time when compared with low-stakes testing states.

Similarly, in a very recent study of the fifty states, Swanson (2006) corroborated Braun's findings. Swanson declared:

> We find strong evidence that implementing a solid program of standards based education policies has been associated with significant gains in mathematics achievement over the past decade, as measured by the NAEP. Positive, but less dramatic results are also found for achievement in reading. (p.1)

Swanson first assigned a yearly policy implementation score to each of the fifty states over a nine-year period of time from 1997-2006. He then examined states' NAEP mathematics and reading achievement patterns over a period of time. Regression analyses yielded "a consistently positive relationship between achievement gains and the implementation of standards-based policies related to academic-content standards, aligned assessments, and accountability measures." (p. 9)

Although the empirical evidence suggests that high-stakes testing increases student achievement (New York State Department of Education, 2004), Tapper (1997)

recommends the "rational, effective, and judicious use" of summative assessments. In a high-stakes assessment report, Tapper suggests:

> If we seek diagnostic information about our systems in order to improve them, then clearly we ought to embrace formative, diagnostic methods and approaches... Formative evaluation is an integral part of effective instructional programs and program development; it is more than a scientific measurement of achievement. (p.14)

Assessments for Learning

Studies have shown that formative assessments for learning coupled with frequent student feedback are most effective in increasing student achievement (Black & Wiliam, 1998). In their comprehensive synthesis of 250 formative assessment research publications comprised mainly of studies considered to be ecologically valid that were published during or after 1988, Black and Wiliam concluded that properly implemented formative assessment in real teaching situations within the classroom results in student learning gains. Additionally, they noted that formative assessment strategies assisted with closing the achievement gap. More specifically, low achieving and learning disabled students demonstrated the greatest achievement gains as a result of the implementation of formative assessment (Fuchs, 1986).

Black and Wiliam's (1998) literature review also illustrated that formative assessment translates into increased learning "in all educational settings: content areas, knowledge and skill types, and levels of education" (Sadler, 1998, ¶ 2). Additionally, their research compilation indicated the ineffectiveness of academic grades when compared with feedback containing specific comments. Furthermore, Black and Wiliam

recognized that the quality of feedback is of critical importance and directly related to the effectiveness of formative assessment and resulting learning gains.

However, Black and Wiliam's (1998) review of the literature concerning formative assessment is limited in scope and discussion (Black, 2000; Sadler, 1998; Sebatane, 1998). Black identifies the preparation of the review as a limitation in that:

> Citation analyses based on the two reviews and the use of key-words both turned out to be inadequate tools in the literature search...The only effective procedure was to turn the pages of 76 of the most likely journals making rapid judgments, first from the titles and then from the abstracts about likely relevance. (¶ 7)

Due to the poor results of the citation analyses and key-word literature search tools, Black suggests that researchers are failing to link their research to previous studies examining related topics that may have resulted in significant findings. Consequently, the literature review may not be all-inclusive of the studies examining the effectiveness of classroom formative assessment.

To further improve the scope of their review, Sebatane (1998) suggests Black and Wiliam should have also considered the influence of high-stakes assessment in teaching and learning. Although Black and Wiliam recognize that accountability and high-stakes assessment procedures will affect teacher assessment practices within the classroom, they did not consider empirical studies examining high-stakes situations. Consequently, to exclude the consideration of the effects of high-stakes assessment from the literature review may result in a skewed perception of the effects of formative assessment on student learning.

Additionally, Sadler (1998) criticizes Black and Wiliam (1998) for their lack of discussion regarding studies in which the results illustrated no effects or negative effects as a result of formative assessment within the classroom setting. Sadler questions, "How can good formative feedback ever backfire in the learning process and inhibit learning?" He suggests that the answer concerns how teachers provide feedback to students. More specifically, Sadler recognizes the type of feedback, frequency of feedback, and context of feedback as fundamental considerations when considering research illustrating negative learning outcomes associated with formative assessment.

In contrast to Black and Wiliam's (1998) findings, two recent studies were unable to illustrate the link between formative assessment coupled with feedback and increased academic performance (Van Evera, 2004; Waddell, 2004). Van Evera's research showed that written formative assessment feedback on homework and class-work assignments in the middle school science classroom resulted in a 0.7 effect size for low achievers and a 0.4 effect size for middle achievers. In contrast, high achievers were negatively affected by the formative feedback intervention. The high achievers experienced a -0.7 effect size as a result of the feedback. In Waddell's study of the effects of written feedback within formative assessment on fourth grade students, he concluded that the data was unable to demonstrate a relationship between feedback effectiveness and academic performance. Even though Van Evera and Wadell were unable to link feedback interventions to increased student achievement, the students in their studies reported increased self-efficacy as a result of the formative assessment feedback interventions in both studies.

Although Black and Wiliam (1998) assert that formative assessment in its broadest definition is a promising strategy to improve student achievement, further

research must be conducted to examine specific formative assessment strategies and their effectiveness in producing student learning gains. Specifically, researchers must explore and compare the effectiveness of formative assessment strategies such as teacher observation, classroom discussion, homework, and testing.

Predictors of State Assessment Performance

In the present era of accountability for student achievement, the results of research investigating the relationship between student performance on low-stakes assessments and achievement on high-stakes state assessments is of great interest to school officials and classroom teachers. Much of the research concerning predictors of state assessment performance pertains to norm-referenced, standardized test scores as predictors of student scores on summative state exams (Aguilera, 1994; Burson, 2001; Maher 2003; Pierce, 2005; Thacker, Dickinson & Koger, 2004). However, research exploring benchmark assessment performance as a predictor of student performance on state assessments has emerged recently (The School Board of Broward County Florida, 2005; Archer, 2005; Success for All Foundation, 2005-2006).

Norm-referenced test scores. Empirical studies in Arizona and California have illustrated that standardized tests can be utilized to predict student performance on state assessments (Aguilera, 1994; Pierce, 2005). These studies reported that students who performed well on standardized tests also performed well on the state assessments.

Based on a sample of 206 students who were randomly selected from an original pool of 40,000 students, Aguilera (1994) concluded that sixth grade student subtest scores on the Iowa Test of Basic Skills could be utilized to predict student performance on the eighth grade Arizona Student Assessment. The results of a California study (Pierce,

2005) also support the use of standardized test scores as predictors of state assessment performance. Pierce (2005) found a modest positive relationship between the seventh grade California Standards Test in conjunction with the California Standards Test for Algebra to the California High School Exit Exam. However, further analysis suggested that the seventh grade California Standards Test alone would serve as a better predictor of performance on the exit exam.

Although limited research suggests that norm-referenced test scores are predictive of state assessment performance, further research concerning a variety of norm-referenced assessments must be conducted in an increased number of states as well as at differing grade levels to determine if the research findings to date are applicable to larger populations of students.

Benchmark assessments. Sirotnik and Kimball (1999) have suggested that accountability systems must provide the resources and opportunities to monitor and support formative, classroom assessment by teachers that is aligned with a challenging, standards-based curriculum. Teacher use of benchmark assessments aligned with the curriculum and state assessments is critical to ensure that students demonstrate increased academic achievement on the end-of-year tests (Shanahan, Hyde, & Manrique, 2005). Sirotnik and Kimball stated:

> It is clear that assessment, teaching, and learning must be coordinated. . . Not to coordinate the kind of assessment (and teaching and learning) that goes on in the classroom with what goes on in state-level assessment and accountability systems seems foolish, at best. (¶ 37)

Summative state assessment results provided several months after the test was administered do little to assist teachers and students in their efforts to improve student achievement. In contrast, benchmark assessments administered on a more frequent basis, provide school officials and students with timely information related to the curricular and instructional practices taking place within the school and provide teachers the opportunity to provide support and remediation where necessary. Using the benchmark assessment data, teachers can modify their curriculum and pedagogical practices prior to the administration of the state assessments (Stiggins & Chappuis, 2005).

School district implementation of benchmark assessments to provide diagnostic student data and predictive student data is occurring more frequently (Olson, 2005). According to an Education Week national survey of 813 school superintendents, 68% of respondents indicated they administer benchmark assessments periodically during the school year. Twenty-seven percent of the superintendents indicated that they implemented the benchmark assessment strategy within the last three years (Archer, 2005).

In Florida, Broward County Public Schools recently implemented benchmark assessments as a strategy to provide diagnostic student data as well as data that will serve as a predictor of student Florida Comprehensive Assessment Test (FCAT) scores. In 2005, the Broward County School Board authorized a research study examining students' benchmark assessment scores and their relationship to FCAT scores (The School Board of Broward County, Florida, 2005). The results of the study, which examined the scores of over 17,700 students, found the predictive validity of the math benchmark assessments to be slightly higher than the predictive validity of the reading benchmark assessments.

However, both correlations were high: mean mathematics correlation = .84 and mean reading correlation = .80. Consequently, the study concluded that proficiency predictions as a result of benchmark assessment performance in mathematics and reading were accurate in identifying students that were in danger of failing the FCAT. Additionally, the study results suggested that the school faculty use the predictive information provided by the benchmark assessments to appropriately target instruction.

Recently, the Norfolk Virginia School District also implemented benchmark assessments at all grade levels. School officials attribute the reading and math score gains as well as the progress in closing the achievement gap to the increased emphasis on data-driven decision making as a result of the implementation of the benchmark assessments (Olson, 2005). In contrast, some teachers struggled with the implementation of benchmark assessments within their districts (Muir, 2001; Olson, 2005). More specifically, these teachers reported that the administration of benchmark assessments has resulted in a narrowing of the curriculum, an increased focus on test preparation, and a lack of instructional time due to test administration.

At this time it is difficult to determine accurately the positive and negative effects of the implementation of benchmark assessments. Because the implementation of benchmark assessments aligned to state assessments has recently emerged as a strategy to increase student achievement and improve state assessment scores, the current benchmark assessment research is limited in sample and experimental design. Additional research must be conducted at the state and district level to provide an increased understanding of the relationship between benchmark assessments and state assessments.

Predictors of Pennsylvania System of School Assessment Performance

Empirical research investigating non-demographic variables as predictors of student performance on the Pennsylvania System of State Assessment (PSSA) can be classified into three categories. The first category consists of studies that investigate curricular variables as predictors of performance on the PSSA (Lehmier, 2001; Rex, 2003; Shields, 2000; Walkovic 2003). The second category is comprised of three studies that examine standardized test scores as predictors of PSSA performance (Burson, 2001; Maher 2003; Thacker, Dickinson & Koger, 2004). The final category concerning benchmark assessments as predictors of PSSA performance includes one empirical study, which is most relevant to the present study (Success for All Foundation, 2005-2006).

Curricular variables. Studies investigating curricular variables as predictors of PSSA performance have consistently shown moderate to high correlations between curricular variables and PSSA performance. One study (Lehmier, 2001) found Computer Curriculum Corporation (CCC) math concepts and skills current average levels to be the strongest predictor of fifth grade PSSA math scores when compared with demographic variables and other CCC variables. Another study (Shields, 2000) found writing strategies to be predictors of ninth grade PSSA writing performance. More specifically, Shields (2000) found the greatest predictive value comes from teaching students about the types of writing assessed on the writing PSSA. Two studies (Rex, 2003; Walkovic, 2003) investigated the relationship between reading strategies and PSSA performance. Walkovic (2003) found that student reported use of comprehension strategies, such as slowing reading rate and using pictures and other illustrations prior to reading a passage, resulted in the strongest eighth grade reading PSSA correlations. Similarly Rex (2003)

found a correlation between reading strategies and PSSA reading performance at the eleventh grade level. In eleventh grade, the greatest predictive value resulted from the reading strategy in which students re-read parts of a reading selection.

Although several studies suggest a variety of math, writing, and reading curricular variables are linked to PSSA performance, the sample of Lehmier's (2001) study and the methodologies of Shields (2000), Rex (2003), and Walkovic (2003) studies limit their research contributions. For example, Lehmier's (2001) research was limited to examining the performance of fifth grade students in one south central Pennsylvania school district. Consequently, the results have limited generalizeability to other school district populations. Shields (2000), Rex (2003), and Walkovic (2003) examine writing and reading strategies and their correlations to mean school PSSA scores rather than individual student PSSA scores. Research correlating writing and reading strategies to individual student's PSSA scores would provide a more accurate correlation between the curricular strategies and the PSSA scores.

Norm-referenced test scores. Several recent studies investigated the correlation between student performance on norm-referenced, standardized tests and mathematics and reading PSSA scores (Burson, 2001; Maher, 2003; Thacker, Dickinson & Koger, 2004). These studies support the link between norm-referenced test score performance and PSSA performance.

Maher (2003) explored the relationship between students' seventh grade Terra Nova scores and eighth grade PSSA scores in one rural school district located in Northeastern Pennsylvania. Student score samples were collected for three consecutive years for a total of 499 student scores. The results indicated significant correlations

between the Terra Nova reading and PSSA reading scores as well as the Terra Nova math and PSSA math scores. The math correlation of .81 was slightly stronger than the reading correlation of .76.

Similarly, Thacker, Dickinson and Koger (2004) examined Terra Nova and PSSA correlations in 4 out of the 25 most populous school districts in Pennsylvania. They found the mathematics correlations to be stronger than the reading correlations in all four districts. More specifically, the correlation between PSSA mathematics and Terra Nova mathematics ranged from .69 to .83 for grade five. The grade eight mathematics correlations ranged from .61 to .83. Terra Nova reading and language scores were considered comparable to the PSSA reading scores. In all cases, the reading and language correlation coefficients were very similar. The grade five correlations ranged from .59 to .76 whereas the grade eight correlations ranged from .69 to .71.

Researchers have also investigated the correlation between PSSA scores and other commonly administered norm-referenced assessments such as the Cognitive Abilities Test, California Achievement Test-5, and the Stanford Achievement Test-9 (Burson, 2001; Thacker, Dickinson, & Koger 2004). The results of these studies illustrated that all comparison tests were highly correlated with the PSSA. These findings support Sicoly's (2002) observation that students who perform well academically typically perform well on all assessments in all subject areas.

The results of multiple Pennsylvania studies illustrate significant correlations between a variety of norm-referenced assessments and the PSSA. Consequently, the data suggests that school officials may utilize students' norm-referenced assessment scores as predictors of future PSSA performance. However, Thacker, Dickinson, and Koger (2004)

recognized that the PSSA assesses student ability in relation to the Pennsylvania Academic Standards and Assessment Anchors whereas the norm-referenced assessments do not assess Pennsylvania-specific content. They state:

> The extent to which another test measures content that is different from Pennsylvania's will limit the strength of the correlation between the two assessments. . . We expect student scores on the various tests to be related. We are left looking for what Hoffman (1998) refers to as "Goldilocks" correlations. Correlations between PSSA and other assessments should fall in this "not-too-high and not-too-low range." (p. 2)

Although norm-referenced standardized tests are helpful in predicting PSSA performance, they are not aligned to the content of the Pennsylvania Standards and Anchors. Periodic assessments designed to measure the content in the Pennsylvania Academic Standards and Assessment Anchors can serve as a more effective gauge of students' progress toward standards-based proficiency. Additionally, more frequent feedback as a result of regularly administered benchmark assessments can provide teachers with timely data that can positively impact classroom curriculum and instruction (Reeves, 2002).

The Success for All Foundation recently developed the Pennsylvania 4Sight Benchmark Assessments to provide teachers with the ongoing data necessary to modify classroom instruction and target students' needs. During the 2005-2006 school year, the 4Sight Benchmark Assessments were administered in 750 schools across more than 310 school districts. The Success for All Foundation conducted limited research concerning the correlation of student scores on two forms of the 4Sight Reading Benchmark

Assessments with PSSA scaled scores (Success for All Foundation, 2005-2006). The resulting reading correlations were .89 and .87 at the third grade level, .88 and .86 at the fifth grade level, and .84 and .83 at the eighth grade level. The 4Sight Mathematics Benchmark Assessment research is more limited. Success for All determined correlations for only one form of the mathematics benchmark assessments at each grade level. The resulting mathematics correlations were .76 at the third grade level, .82 at the fifth grade level, and .84 at the eighth grade level (J. B. Anderson, personal communication, August 3, 2006). Additional research exploring the relationship between students' performance on the benchmark assessments and performance on the PSSA in an increased number of school districts is necessary to strengthen the body of research concerning the 4Sight Assessments as predictors of PSSA performance.

Pennsylvania 4Sight Benchmark Assessments. The 4Sight Reading Benchmark Technical Report for Pennsylvania (2005-2006), the final study in this review, warrants special consideration due to its relevance to the present study. In 2005, the Success for All Foundation published the second version of the 4Sight Reading Benchmark Assessment Technical Report. This report provided information concerning the Pennsylvania 4Sight Reading Benchmark Assessment test development including: benchmark design, cultural sensitivity, benchmark blueprints, form assembly, review process, field testing, pre-pilot testing, pilot testing, final form development, and scoring. Additionally, the report described the processes used to establish test validity and reliability.

The Success for All Foundation developed the Pennsylvania 4Sight Assessments to serve as a "formative evaluation of each student's progress and as a predictive measure

of each student's performance on the state-wide assessments" (Success for All Foundation, 2005-2006, p.4). The low-stakes, benchmark assessments provide student performance data that can be of use to both teachers and administrators. Teachers may utilize the results of the one-hour, benchmark assessments to inform their classroom instruction, whereas school administrators may analyze the benchmark assessment results to determine appropriate professional development.

The 4Sight Reading Benchmark Technical Report for Pennsylvania (2005-2006) is the only study that examines Pennsylvania-specific formative assessments aligned to the PSSA as predictors of state assessment performance. Although the sample population in this study was limited to only four school districts in Pennsylvania, the results illustrate a significant positive correlation between the students' 4Sight benchmark assessment scores and their PSSA scaled scores in the four participating districts.

A noteworthy limitation of the 4Sight Reading Benchmark Technical Report for Pennsylvania pertains to the pre-pilot phase of the research. The 4Sight assessment data gathered in the pre-pilot study was not on grade level. More specifically, students who had already taken the third, fifth, or eighth grade PSSA and were now promoted to fourth, sixth, or ninth grade were asked to take the 4Sight assessment designed for students in grade three, five, or eight. It is possible that the resulting pre-pilot study benchmark assessment scores did not accurately reflect performance of typical third, fifth, or eighth grade students. However, following the pre-pilot study, the researchers conducted a pilot study in which the reading benchmark assessments were administered on grade level. The resulting reading correlations determined from the on-grade level pilot study were

published in the 4Sight Reading Benchmark Technical Report for Pennsylvania (2005-2006).

Additionally, the Success for All technical report only provides information concerning the development and characteristics of the Pennsylvania 4Sight Reading Benchmark Assessments for grades three, five, and eight. There is no published technical report for the Pennsylvania 4Sight Mathematics Benchmark Assessments. Moreover, the math correlations were determined as a result of a pilot study where the 4Sight Mathematics Benchmark Assessments were administered to students off grade level thus raising concern as to whether the correlations accurately reflect the mathematics performance of students in the third, fifth, and eighth grades.

The present study will address the three identified limitations of the Pennsylvania 4Sight Assessment research. First, the study will administer the 4Sight assessments on grade level. Sixth grade students will complete sixth grade benchmark assessments. Second, the sixth grade students participating in the study will complete both the reading and the mathematics benchmark assessments. The student performance data from both the reading and the mathematics 4Sight Benchmark Assessments will be correlated with the students' PSSA scaled scores at the sixth grade level. Additionally, the present study will add to the current 4Sight Benchmark Assessment research by determining whether there is any significant difference in Reading and Mathematics PSSA performance means for the students in districts using the 4Sight Benchmark Assessments versus the students in districts not utilizing the 4Sight Benchmark Assessments.

Summary

The future of America depends upon the quality of the educational programs provided to students across the nation. The main objectives of the No Child Left Behind legislation (2002) were to raise educational standards and increase accountability within the nation's public schools. The national increased emphasis on high-stakes testing to determine students' achievement of state content standards has resulted in an increased focus on the results of research related to educational assessment. As recent research indicates, the implementation of high-stakes testing programs appears to result in increased student achievement (Braun, 2004; Raymond & Hanushek, 2003; Swanson, 2006; Rosenshine, 2003); however, educational experts suggest that summative assessment must be coupled with formative assessment strategies to attain greater increases in student achievement (Black & William, 1998; Reeves, 2002; Stiggins & Chappuis, 2005).

The administration of ongoing benchmark assessments is one formative assessment strategy that is gaining popularity within the nation's public schools (Olson, 2005). Stiggins and Chappuis (2005) identified two main reasons for the recent emergence of benchmark assessments as a school improvement strategy. First, end-of-year state assessments are unlikely to affect daily instruction because states provide student performance results to teachers months after the administration of the high-stakes exam. Second, the data from the summative state assessments frequently lacks the individual student diagnostic information that assists teachers in determining students' mastery of specific learning goals rather than overall proficiency in relation to content standards.

In contrast, benchmark assessments can be administered several times during the school year thus providing multiple opportunities to provide teachers with timely student performance data in the form of diagnostic reports as well as overall proficiency reports. Throughout the year, teachers can utilize these reports to modify classroom instruction to target students' specific skill weaknesses. Furthermore, teachers can make informed curricular and instructional adjustments to increase student learning prior to the administration of the state summative assessment.

Although research supports formative assessment as a classroom strategy to increase student achievement and close the achievement gap (Black & William, 1998; Fuchs, 1986; Van Evera, 2004), the effectiveness of the administration of benchmark assessments as a strategy to increase student achievement and predict state assessment performance needs to be further explored. The examination of an empirical study comparing the student performance data in three school districts implementing state-specific benchmark assessments versus three match paired school districts not utilizing the state-specific benchmark assessments will enhance the current research reflecting formative assessment and effective school improvement strategies in today's high-stakes testing environment.

CHAPTER III

Methodology

The primary purpose of this study was to determine the effectiveness of utilizing the Pennsylvania 4Sight Benchmark Assessments to predict student performance on the Reading and Mathematics PSSA. First, the study determined if there was a significant relationship between sixth grade student performance data on the Reading and Mathematics 4Sight Benchmark Assessments and the Reading and Mathematics PSSA. Second, the study determined if there was any significant difference between sixth grade student performance on the Reading and Mathematics PSSA for the students who participated in the 4Sight Reading and Mathematics Benchmark Assessments as compared with students that did not participate in the 4Sight Benchmark testing. The following research questions guided this study:

1. Is there a significant relationship between the 4Sight Reading Benchmark Assessment scores and PSSA Reading scaled scores at the sixth grade level?
2. Is there a significant relationship between the 4Sight Mathematics Benchmark Assessment scores and PSSA Mathematics scaled scores at the sixth grade level?
3. Is there any significant difference in the change on Reading PSSA performance means for the students who participated in the 4Sight Reading Benchmark Assessments as compared with a matched group that did not participate in the 4Sight Reading testing?
4. Is there any significant difference in the change on Mathematics PSSA performance means for the students who participated in the 4Sight Mathematics

Benchmark Assessments as compared with a matched group that did not participate in the 4Sight Mathematics testing?

Design

The research study utilized a quasi-experimental design. To determine whether a significant relationship existed between Pennsylvania 4Sight Benchmark Assessment Reading scores and Reading PSSA scaled scores, sixth grade students' 2006 4Sight Benchmark Assessment Reading scores were correlated with their 2006 Reading PSSA scaled scores in a convenience sample of three Pennsylvania school districts utilizing the 4Sight Benchmark Assessments. Similarly, sixth grade students' 2006 4Sight Benchmark Assessment Mathematics scores were correlated with their 2006 Mathematics PSSA scaled scores to determine whether a significant relationship existed between 4Sight Benchmark Assessment Mathematics scores and Mathematics PSSA scaled scores in the same three school districts.

To determine whether there was any significant difference between student performance on the Reading PSSA for the sixth grade students who participated in the 4Sight Reading Benchmark Assessments as compared with a matched group that did not participate in the 4Sight Reading testing, the students' fifth grade (2005) and sixth grade (2006) Reading and Mathematics PSSA scaled scores were transformed to *T* scores. A repeated measures analysis of variance (ANOVA) was then conducted to determine whether there was a significant difference in the change on PSSA performance means for the experimental group versus the control group.

Sample

Representative of three southeastern counties, a convenience sample of three Pennsylvania school districts participating in the 4Sight Benchmark Assessment initiative at the sixth grade level were selected for the study. The districts varied in their district enrollment, percentage of low-income families by local education agency, and school locale code classifications (Appendix A). As per the 2004-2005 Public School Enrollments reported by the Pennsylvania Department of Education (PDE), school district A's total enrollment was 2083 students. As identified in the PDE 2004-2005 Percentage Enrollment of Low Income Families by Local Education Agency (LEA) document, low-income families made up 5.8% of the district's population. Additionally, according to the PDE 2003-2004 Ubancentric and Metrocentric School Locale Codes, the school was classified as a rural school district. School district B's total enrollment was 1999 students. Low-income families made up 12.7% of district B's population and the district was classified as an urban-fringe school district. District C was comprised of a total of 1743 students. Low-income families made up 25.8% of the district's population. District C was classified as an urban fringe school district.

The convenience sample was match paired with three other Pennsylvania school districts not utilizing the 4Sight Benchmark Assessments. The districts were match paired according to the following demographic characteristics: district enrollment, percentage of low-income families by local education agency, and school locale code. To determine the districts eligible to be match paired with the convenience sample, the researcher contacted a representative at the Pennsylvania Training and Technical Assistance Network (PaTTAN) and requested a list of districts and schools that did not participate in

the 4Sight testing at the sixth grade level during the 2005-2006 school year. The researcher then identified three schools that matched the demographic characteristics of the convenience sample. The districts were match paired using the 2004-2005 Public School Enrollments, 2004-2005 Percentage Enrollment of Low Income Families by LEA, and 2003-2004 Urbancentric and Metrocentric School Locale Codes located on the Pennsylvania Department of Education website. A district was considered a sufficient match if it met the following criteria: a) the match district's enrollment was plus or minus 375 students of the convenience sample district's total enrollment, b) the match district's percentage of low income families was plus or minus 1% of the convenience sample district's percentage, and c) the match district's school locale code was an exact match or near match to the convenience sample district's school locale code.

Instrumentation

The researcher utilized two instruments to collect student achievement data for use in this study: the Pennsylvania 4Sight Benchmark Assessments and the 2005 and 2006 Reading and Mathematics components of the Pennsylvania System of School Assessment (PSSA). The three school districts participating in the Pennsylvania 4Sight Benchmark Assessments selected for the study provided the sixth grade students' 2006 Reading and Mathematics 4Sight Benchmark Assessment scores as well as their corresponding 2005 and 2006 Reading and Mathematics PSSA scaled scores. Additionally, the matched pair school districts' provided the researcher with their students' 2005 and 2006 Reading and Mathematics PSSA scaled scores.

Pennsylvania 4Sight Benchmark Assessments

Presently, the Success for All Foundation has published only the 4Sight Reading Benchmark Technical Report for Pennsylvania (2005). The Success for All Foundation has not released a 4Sight Mathematics Technical Report for Pennsylvania. For this reason, only the 4Sight Reading Benchmark Assessment development process will be reviewed in detail.

4Sight Item Development. The Success for All Foundation utilized a multi-step process to align the content and format of the Pennsylvania 4Sight Reading Benchmark Assessments to the PSSA. First, Success for All Foundation benchmark assessment developers researched the content of Pennsylvania's standards, assessment anchors and state assessment. Second, the developers reviewed the format of the state assessments including the lengths and types of texts, as well as the types of test items, such as multiple-choice and open-ended questions. Third, Success for All identified the weights of the items included on the PSSA. The test developers used the weights to determine the number of each type of item needed on the benchmark assessments to mirror the proportions of the various test items on the PSSA. Finally, the benchmark assessment developers used the PSSA content, format, and item weight research to map out 4Sight Reading Assessment blueprints aligned to the content and format of the state assessment.

Two forms of the reading assessment were developed directly from the assessment blueprint. To ensure cultural sensitivity, the Success for All Foundation test developers internally reviewed for race, ethnicity, and gender bias the newly created 4Sight assessment items. During the internal review, Success for All test developers analyzed the content of the benchmark assessment items to ensure compliance with the

California Standards for Evaluation of Instructional Materials with Respect to Social Content (1986).

The Success for All Foundation staff conducted a field test to provide opportunities for Success for All staff, pilot schools, district officials, and state education officials to provide feedback pertaining to the content, format and cultural sensitivity of the reading benchmark assessment forms. Next, Success for All revised the benchmark assessment drafts to reflect the changes as a result of the field testing. Success for All then submitted the benchmark assessments for editing, formatting, and printing. Finally, Success for All provided the printed tests to the schools selected to participate in the pre-pilot and pilot studies.

The Success for All Foundation conducted the Reading Benchmark Assessment pre-pilot and pilot studies in the same four districts. Students in grades four, six and nine in one rural school district, two suburban, and one urban district comprised the pre-pilot study sample population. In these districts, school officials administered the 60-minute reading benchmark assessments to the students. The assessments were not administered on grade level. To ensure consistency with the initial scoring procedure, the Success for All Foundation staff scored the completed assessments. The Success for All Foundation then obtained student state assessment scores from the previous year and matched each student's current benchmark assessment score to their previous year state assessment score. Matching the student scores from the benchmark assessment with the scores from the state assessment resulted in the development of grade level correlations that could be used to provide an estimate of student performance on the PSSA in the future.

As a follow up to the pre-pilot study, the Success for All Foundation conducted a pilot study that required the administration of the 4Sight Reading Benchmark Assessments to students in the same school districts that were selected to participate in the pre-pilot study; however, during this phase, school officials administered the benchmark assessments on grade level. School staff utilized scoring masks and scoring guides to score the completed assessments. The school districts provided the Success for All Foundation with the resulting student performance data from the 4Sight Benchmark Assessments as well as students' reading levels as determined from the Gates-MacGinitie Reading Test. Success for All staff matched the 4Sight Benchmark Assessment Scores with the reading levels to develop grade level correlations. These resulting correlations allow a school district to utilize 4Sight Benchmark Assessment scores to estimate students' reading levels.

After the completion of the pre-pilot and pilot studies, the Success for All Foundation used the original two forms of the benchmark assessments to create the final five forms of the benchmark assessments. The test developers created five forms of the assessment to provide school districts with multiple instruments to collect student performance data: baseline assessment, first quarter assessment, second quarter assessment, third quarter assessment, and fourth quarter assessment. The original two forms of the assessment were created directly from the 4Sight Assessment blueprint. The other three forms are scrambles of the original forms developed from the blueprint. More specifically, test forms #1 and #2 were created directly from the test blueprint. Test forms #3 and #5 are scrambled forms of test #1. Test form #4 is a scramble of test #2. Scrambled forms of the assessment contain the same multiple-choice test items as the

original form; however, the test items and answer choices appear in a different order. Additionally, the open-ended items are slightly varied on the scrambled test forms.

4Sight Validity and Reliability. Validity and reliability frequently characterize the quality of an educational assessment. McMillan (2000) defined instrument validity as:

> ...an overall evaluation of the extent to which theory and empirical evidence support interpretations that are implied in given uses of the scores...In other words, validity is a judgment of the appropriateness of a given measure for the specific inferences or decisions that result from the scores generated by the measure. (p. 132-133)

Validity of an instrument may be established by providing evidence related to test content, internal structure, and relations to other variables.

The Success for All Foundation Reading Benchmark Technical Report for Pennsylvania (2005) identified the procedures used to establish validity. The Success for All test developers established content validity through the PSSA and 4Sight Benchmark Assessment blueprint development process. The PSSA blueprints served as a guide for the construction of the benchmark assessments. The item stems from the released state assessment items were used to create the benchmark assessment items, resulting in formative, benchmark assessments that reflect the design of the summative PSSA. Additionally, Success for All established the criterion validity of the benchmark assessments by developing correlations of the pre-pilot and pilot study students' benchmark assessment scores with their scaled scores on the PSSA in grades three, five, and eight. The resulting correlations ranged from .83 to .89 (Appendix B). Although the 4Sight Mathematics Benchmark Assessment correlations were not published in a

technical manual, Success for All determined correlations for one form of the mathematics benchmark assessments at each grade level. The resulting mathematics correlations were .76 at the third grade level, .82 at the fifth grade level, and .84 at the eighth grade level (Appendix C; J. B. Anderson, personal communication, August 3, 2006).

The reliability of an educational assessment typically refers to the consistency of the results produced from multiple administrations of the same test (Thacker, Dickinson, & Koger, 2004). The Success for All test developers established two forms of reliability, alternate-form and inter-rater reliability, for the Pennsylvania 4Sight Reading Benchmark Assessments. The Pearson correlation procedure determined the inter-form reliability for the two original forms of the benchmark assessments in grades three, five, and eight. The Pearson correlation ranged from .793 to .865 indicating high reliability (Appendix D). Inter-rater reliability was established by providing benchmark assessment scorers with professional development concerning the use of the scoring masks and scoring guides to score the benchmark assessments. Furthermore, the Success for All Foundation required that two or more individuals score the open-ended questions to increase the reliability of the open-ended scores. Finally, benchmark assessment performance was also correlated with student reading levels to provide an additional measure of reliability.

Pennsylvania System of School Assessment

Prior to the implementation of the No Child Left Behind (2002) federal legislation, Pennsylvania adopted academic standards for Reading, Writing, Speaking and Listening, and Mathematics in 1999 (Pennsylvania State Board of Education, 1999). The state standards, which are part of the Chapter 4 Regulations on Academics Standards and

Assessment, provide developmentally appropriate benchmarks that designate what students should know and be able to do in grades 3, 5, 8, and 11. School district personnel utilize the academic standards as a guide when designing district curriculum, classroom instruction, and local assessments.

The Pennsylvania System of School Assessment (PSSA) is an annual criterion-referenced assessment based upon the Reading, Writing, Speaking and Listening, and Mathematics academic standards. Beginning in 2006, Pennsylvania required that all students in grades 3 through 8 and 11 participate in the Reading and Mathematics PSSA to comply with the annual assessment requirements specified by No Child Left Behind. Additionally, Pennsylvania requires that school districts administer the Writing PSSA to students in grades 5, 8, and 11.

PSSA Item Development. Both the Reading and Mathematics PSSA consist of multiple-choice and open-ended tasks. The majority of each assessment is comprised of multiple-choice items worth one point each, which measure a broad range of the students' content knowledge. A smaller portion of the assessment consists of open-ended tasks where students must apply problem-solving skills and provide a written explanation of the process utilized to solve the problem. On the 2005 PSSA, students receive a score of 0-3 points for each reading open-ended task whereas they receive a score of 0-4 points for each mathematics open-ended task.

In 2005, the PSSA test content blueprint was revised to reflect the content and item distribution specified in the Reading and Mathematics Assessment Anchors (Pennsylvania Department of Education, 2004). The Pennsylvania Department of Education created the Assessment Anchors as a tool to improve the alignment of

curriculum, instruction and assessment practices in the state of Pennsylvania. The Pennsylvania Department of Education released the 2004 Reading and Mathematics Assessment Anchors to "anchor" the state assessment system and school district curriculum and instruction and to communicate the assessment limits to stakeholders.

The Data Recognition Corporation (DRC) in collaboration with the National Center for Improvement of Educational Assessment (NCIEA) created the 2005 Reading and Mathematics PSSA operational layout for grades 5, 8, and 11. Twenty grade level forms of the PSSA consisting of both common and matrix items were created. All forms of the assessment contain common items to be completed by all students within each grade level. Matrix items, however, are typically unique to one form of the assessment and serve to expand the test item pool, field test new items, and link items from the previous year's assessment. Student PSSA scores are derived from only the common items. In 2005, the maximum numbers of fifth, eighth and eleventh grade operational points were 52 points in reading and 66 points in mathematics (Appendix E). Reading and Mathematics PSSA scores were reported in two reading categories (Comprehension and Reading Skills and Interpretation and Analysis of Fiction and Nonfiction Text) and five mathematics categories (Numbers and Operations, Measurement, Geometry, Algebraic Concepts, and Data Analysis and Probability) that aligned to the Pennsylvania Reading and Mathematics Assessment Anchors.

The Data Recognition Corporation and WestEd selected college graduates with teaching experience and knowledge and expertise in reading or mathematics to become PSSA item writers. The Data Recognition Corporation and WestEd provided the selected writers with individual PSSA item writing training concerning the development of

multiple-choice items and open-ended tasks. When commencing the test item development process, the item writers and reviewers of the 2005 Reading and Mathematics PSSA referenced and considered a variety of resources. More specifically, they considered the "Pennsylvania Assessment Anchors and Eligible Content, grade-level appropriateness (reading/interest level, etc.), Depth of Knowledge, cognitive level, item/task level of complexity, estimated difficulty level, relevancy of context, rationale for distractors, style, accuracy, and correct terminology" (Data Recognition Corporation, 2005, December). Reference materials included: *The Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), *Principles of Universal Design* (Thompson, Johnstone, & Thurlow, 2002), DRC's *Bias, Fairness, and Sensitivity Guidelines,* and *Principles, Guidelines* and *Procedures for Developing Fair Assessment Systems* (Pennsylvania Assessment Through Themes, 1999, September).

Prior to field testing, content committee experts, educators and reading and mathematics supervisors from Pennsylvania school districts, thoroughly reviewed and evaluated the multiple-choice items and open-ended tasks. The Data Recognition Corporation and WestEd personnel facilitated the assessment item review process and required the content experts to review the Reading and Mathematics test items for: quality, content, anchor alignment, content limits, grade level appropriateness, difficulty level, depth of knowledge, appropriate source of challenge, correct answer, quality of distractors, graphics, appropriate language demand, and freedom from bias (Data Recognition Corporation, 2005, December). Applying the aforementioned criteria, the content expert committee assigned a status of "approved, accepted with revision, move to another assessment anchor or grade, or revise/rewrite" to each assessment item. After the

content expert review, a Bias and Sensitivity Committee reviewed and evaluated the assessment items with regard to bias and sensitivity issues. The Data Recognition Corporation then prepared the acceptable items for field testing.

The Data Recognition Corporation applied conventional item analysis methods to the field test multiple-choice and open-ended items and statistically analyzed the results. As stated in the PSSA Technical Report (2005):

> With any psychometric model, an item analysis is a search for unexpected results. In general, more capable students are expected to respond correctly to easy items and less capable students are expected to respond incorrectly to difficult items. If either of these situations does not occur, the item would be reviewed by DRC test development staff and committees of Pennsylvania educators to determine the nature of the problem and the characteristics of the students affected. The primary way of detecting such conditions is through the point-biserial correlation coefficient for dichotomous (MC) items and the item-total correlation for polytomous (OE) items. In each case the statistic will be positive if the total test mean score is higher for the students who respond correctly to the MC items (or attain a higher OE item score) and negative when the reverse is true. (p. 27)

Flagged items resulting from the item analysis process are subjected to further review by committees of Pennsylvania educators.

Finally, the statisticians applied the Mantel-Haenszel procedure to the acceptable assessment items to determine whether differential item functioning was present. Differential item functioning results when an assessment is administered to test participants of equal ability but the results indicate differing success rates with regard to

correctly answering test items. Test items may be described as being potentially gender or ethnically biased if the differing success rates are associated with various gender or ethnic groups. Items identified as having strong evidence of a problem with regard to test bias were reviewed and revised or discarded.

The items selected for the operational 2005 PSSA in Reading and Mathematics emerged from the Spring 2004 field test and were subjected to numerous reviews including reviews by content area and curriculum specialists, representatives form the National Center for Educational Outcomes, the Bias and Sensitivity Review Committee, Pennsylvania educators, Pennsylvania Department of Education personnel, PDE subject-area teacher advisory committee, and PDE subject-area teacher committees. After the rigorous review process, items deemed "acceptable" for the 2005 PSSA met three criteria. An acceptable item was considered to be "appropriately aligned with its designated Assessment Anchor and sub-classifications, acceptable in terms of bias/sensitivity issues, including differential item functioning (for gender and race), and free of major psychometric flaws, including a special review of flagged items" (Data Recognition Corporation, 2005, December, p. 31).

PSSA Validity and Reliability. Content validity, convergent validity and discriminant validity studies (Achieve, Inc, 2005; Data Recognition Corporation, 2005, December; Thacker, Dickinson, & Koger, 2004) provide the evidence to support the validity of the PSSA. As described in the PSSA Technical Report (2005), assessment developers utilized a test blueprint aligned to the Pennsylvania Academic Standards and Assessment Anchors to construct the Mathematics and Reading PSSA. Moreover, the test creators developed assessment items through the previously described multi-step item

development process. The use of the test blueprint and the implementation of a rigorous test item development process provide evidence of the content validity of the PSSA. Additionally, in an independent study examining the alignment between PSSA test items and the Pennsylvania Assessment Anchors and Eligible Content, Achieve Incorporated (2005) concluded "Pennsylvania's tests in reading and mathematics at grades 3, 5, 8, and 11 are strongly aligned to the Assessment Anchors and Eligible Content statements" (p. 7). The results of the Achieve Inc. study provide additional evidence to support the content validity of the PSSA.

In a 2004 study of an earlier version of the PSSA, Thacker, Dickinson, and Koger provided evidence to support the convergent and discriminant validity of the assessment. In this study which examined the correlation between the results of the PSSA and five different achievement tests (CTBS/Terra Nova, CAT-5, SAT-9, Northwest Evaluation Association's achievement tests, and New Standards Reference Exam) in seven school districts, the researchers concluded "all comparison tests were highly correlated with the PSSA, even among dissimilar subjects (for example, PSSA reading and Terra Nova science)" (p. 160). The highest same subject correlations were illustrated in mathematics and ranged from .7 to .9, whereas reading correlations ranged from .6 to .8. Thacker, Dickinson, and Koger (2004) articulated:

> The extent to which another test measures content that is different from Pennsylvania's will limit the strength of the correlation between the two assessments. In fact, if the correlations were very high, it would raise questions as to whether the two assessments were measuring anything different at all, and consequently whether both are necessary. (p. 2)

For this reason, Thacker, Dickinson, and Koger (2004) concluded that the PSSA is highly correlated with the alternate measures of achievement and provided evidence to support the convergent validity of the PSSA.

Thacker, Dickinson, and Koger (2004) also examined the discriminant validity of the PSSA by analyzing the performance of various subgroups on the PSSA and comparison achievement tests "to determine if the PSSA exhibited any differential impact based on gender, ethnicity, English proficiency, or socioeconomic status" (p. 160). Although differences in the scoring of the demographic subgroups emerged, the differences in scoring were similar to national scoring patterns. Consequently, researchers concluded that there was no evidence to support PSSA bias toward any demographic subgroup.

The Data Recognition Corporation provided additional evidence to support the construct validity of the PSSA. Using Pearson's Correlation Coefficient, DRC computed the PSSA correlations by reporting category. The results illustrated "the expected pattern and magnitude of correlations" (Data Recognition Corporation, 2005, December, p. 101). The reading total correlated higher with the reading reporting categories than with the mathematics total or mathematics reporting categories and vice versa.

The Data Recognition Corporation provided PSSA reliability information by computing the Cronbach Alpha reliability indices for the total student population and for students in each demographic subgroup and applying an extension of the Huynh (1976) statistical model to estimate the proportion of students who would be reclassified in the same performance levels on two equivalent administrations of the PSSA. The Cronbach Alpha overall reliability estimates for grade 5 reading and mathematics were .91 and .92,

respectively (Appendix F). Grade 8 overall reliability estimates were .91 for reading and .93 for mathematics (Appendix G). Grade 11 overall reliability estimates were .92 and .94 for reading and mathematics, respectively (Appendix H). Additionally, the Huynh decision indices depict "the proportion of students who are consistently classified in the same achievement levels on two equivalent administrations of the test" (Data Recognition Corporation, 2005, December, p. 103). The highest overall decision indices were observed at grade 11 in both reading and mathematics (Appendix I).

Data Collection

The district-level administrator responsible for assessment at each of the three school districts making up the convenience sample and each of the match pair school districts received an E-mail cover letter and informed consent form explaining the purpose of the study and inviting the district to participate in the study (Appendix J). The initial E-mail included the following: a) a cover letter explaining the purpose of the study b) a description of the research, c) information concerning the requirements and procedures for participation in the study d) an assurance of confidentiality, e) an explanation of the risks and benefits of the study, f) an explanation of the right to withdraw, and g) the researcher's contact information.

Convenience sample school districts choosing to participate in the study proceeded with the 2005-2006 sixth grade 4Sight data extraction into Microsoft Excel by following the instructions included in the cover letter. Participating districts logged into the 4Sight Member Center using the district username and password. They then selected the school that contained the districts' sixth grade data. District officials clicked "Admin" on the left side of the screen and then "Data Extracts." They then chose "EX-3524

Student Test Results" and selected the following report parameters: School Yr. 2005-2006, 4Sight, Reading, Homeroom Teacher, and Grade 6. After indicating the parameters of the report, district officials clicked "Create Extract." To ensure confidentiality, the districts deleted student names, but retained the student identification numbers in the file. They saved the file and repeated the same steps to complete the extraction of the Grade 6 4Sight Mathematics scores. District officials sent via E-mail both 4Sight data files to the researcher.

Both the convenience sample and match pair districts acquired and prepared for submission the Microsoft Excel files containing 2005 fifth grade Reading and Mathematics PSSA scores, and 2006 sixth grade Reading and Mathematics PSSA scores by following the instructions included in the cover letter. District officials logged into the DRC Web Reporting System using the district user ID and password. They clicked on "Reports" and then "2005 Grade 5 Math Reading District Data File." The district officials deleted all student information except for the columns containing the District Student Identification Numbers, Reading PSSA Scaled Scores, and Mathematics PSSA Scaled Scores and then saved the Excel file. The participating districts repeated the steps above to complete the extraction of the "2006 Grade 6 Math Reading District Data File." The district officials sent via E-mail both PSSA data files to the researcher.

Data Analysis

Data analysis revealed the relationship between 4Sight Benchmark Assessment scores and PSSA scaled scores in the convenience sample consisting of three school districts. The Statistical Package for the Social Sciences (SPSS) was utilized to conduct the data analysis portion of this study. First, SPSS determined the means and standard

deviations of the Reading and Mathematics 4Sight Benchmark Assessment and PSSA scaled scores. Next, the Pearson correlation coefficient was calculated. The significance of the Pearson correlation was tested at the .0005 level.

After determining the significance of the correlation between the Reading and Mathematics 4Sight Benchmark Assessment scores and PSSA scaled scores for the convenience sample, it was necessary to transform the PSSA scaled scores into standard scores. The standard score calculations were necessary because the fifth and sixth grade versions of the PSSA are developmentally different. For this reason, the 2005 reading and mathematics scaled scores were converted to *T* scores based on the mean and standard deviation of the fifth graders in the combined experimental and control group. Similarly, the 2006 reading and mathematics scaled scores were converted to *T* scores based on the mean and standard deviation of the sixth graders in the combined experimental and control group. Consequently, the 2005 and 2006 PSSA scaled scores were made comparable by using the same reference scale. A repeated measures analysis of variance (ANOVA) was then conducted to determine whether there was a significant difference in the change on PSSA performance means for the experimental group versus the control group in both reading and mathematics. Finally, effect sizes of the group differences for the Reading and Mathematics PSSA were calculated.

CHAPTER IV

Results

The primary purpose of this study was to determine the effectiveness of utilizing the Pennsylvania 4Sight Benchmark Assessments as a predictor of student performance on the Reading and Mathematics PSSA. More specifically, the study examined whether there was a significant relationship between sixth grade student performance data on the Reading and Mathematics 4Sight Benchmark Assessments and the Reading and Mathematics PSSA. Additionally, the study determined whether there was any significant difference between sixth grade student performance on the Reading and Mathematics PSSA for the students who participated in the 4Sight Reading and Mathematics Benchmark Assessments as compared with a matched group of students that did not participate in the 4Sight Benchmark testing. The following sections of this chapter provide the answers to the four research questions that guided this study:

1. Is there a significant relationship between the 4Sight Reading Benchmark Assessment scores and PSSA Reading scaled scores at the sixth grade level?
2. Is there a significant relationship between the 4Sight Mathematics Benchmark Assessment scores and PSSA Mathematics scaled scores at the sixth grade level?
3. Is there any significant difference in the change on Reading PSSA performance means for the students who participated in the 4Sight Reading Benchmark Assessments as compared with a matched group that did not participate in the 4Sight Reading testing?
4. Is there any significant difference in the change on Mathematics PSSA performance means for the students who participated in the 4Sight Mathematics

Benchmark Assessments as compared with a matched group that did not participate in the 4Sight Mathematics testing?

Descriptive Statistics

Descriptive statistics for the Reading 4Sight Benchmark Assessment for the students in the experimental group showed a mean score of 22.58 and a standard deviation of 4.29 (n = 438). The corresponding descriptive statistics for the Math 4Sight Benchmark Assessments revealed a mean score of 22.88 and a standard deviation of 6.68 (n = 438).

Table 1 shows the PSSA mean scaled scores for math and reading for the control group and the experimental group at fifth and sixth grade.

Table 1

Means and Standard Deviations on PSSA Reading and Math for Fifth and Sixth Grade

	Reading	Math	Reading	Math
Group	*M(SD)*	*M(SD)*	*M(SD)*	*M(SD)*
	Fifth Grade		Sixth Grade	
Control[a]	1397(204)	1471(199)	1363(172)	1421(200)
Experimental[b]	1357(211)	1418(186)	1365(169)	1427(195)

[a]n = 413. [b]n = 438.

An examination of Table 1 reveals that the scaled scores for the control group declined from fifth to sixth grade. In contrast, the scaled scores for the experimental group improved from fifth to sixth grade. It is important to note that the sixth grade versions of the Reading and Mathematics PSSA are developmentally different from the fifth grade versions of the Reading and Mathematics PSSA in that they are based upon the more

challenging concepts and skills identified in the sixth grade Reading and Mathematics Assessment Anchors. For this reason, it would not be appropriate to conclude that the lower sixth grade scaled scores for the control group mean the students knew less in sixth grade than in fifth grade. Instead, a more appropriate interpretation is that the control group's performance on the sixth grade Reading and Mathematics PSSA was lower relative to what they were expected to know at the sixth grade level. On the other hand, the experimental group's performance on the sixth grade Reading and Mathematics PSSA was higher relative to what they were expected to know at the sixth grade level.

Primary Purpose

The primary purpose of the study was to determine whether there was a significant relationship between sixth grade students' 4Sight Benchmark Assessment scores and their sixth grade PSSA scaled scores in both reading and mathematics.

The Pearson correlation between students' sixth grade 4Sight Reading Assessment scores and their sixth grade Reading PSSA scaled scores was positive and significant ($r = .77$, $p < .0005$). Thus, there was a strong positive correlation between sixth grade student performance on the 4Sight Reading Benchmark Assessment and the Reading PSSA.

The correlation between sixth grade 4Sight Math Assessment scores and Math PSSA scaled scores was also positive and significant ($r = .77$, $p < .0005$). Consequently, the researcher concluded there was a strong positive correlation between sixth grade student performance on the 4Sight Mathematics Benchmark Assessments and the Mathematics PSSA.

Secondary Purpose

The secondary purpose of the study was to determine whether participation in 4Sight Benchmark Assessments led to greater improvement on PSSA scores from fifth to sixth grade. The null hypothesis was that there would be no difference in PSSA change for the experimental group versus the control group. The alternative hypothesis was that the experimental group would show more improvement than the control group on PSSA scores. For the purpose of testing this hypothesis, PSSA scaled scores were normalized within the data set. Fifth grade students' scaled scores were transformed to *T* scores based on the mean and standard deviation of the fifth graders in the combined experimental and control group. Similarly, sixth grade students' scaled scores were transformed to *T* scores based on the mean and standard deviation of the sixth graders in the combined experimental and control group. The standard score calculations were necessary because the fifth grade versions of the Reading and Mathematics PSSA are developmentally different than the sixth grade versions of the Reading and Mathematics PSSA. Table 2 presents the *T* score means and standard deviations for the dependent variables, Reading and Math PSSA, in fifth and sixth grade for the control and experimental groups.

Table 2

Means and Standard Deviations on T Scores for PSSA Reading and Math for Fifth and Sixth Grade

	Reading	Math	Reading	Math
Group	*M(SD)*	*M(SD)*	*M(SD)*	*M(SD)*
	Fifth Grade		Sixth Grade	
Control[a]	50.99(9.77)	51.41(10.25)	49.94(10.14)	49.84(10.15)

Experimental[b]	49.07(10.14)	48.67(9.58)	50.06(9.97)	50.15(9.87)

[a]$n = 413$. [b]$n = 438$.

The table shows a decline in mean Reading PSSA performance from 50.99 in fifth grade to 49.94 in sixth grade for students in the control group. However, students in the experimental group demonstrated improved performance on the Reading PSSA from 49.07 in fifth grade to 50.06 in sixth grade. A similar pattern was observed in math. The control group mean declined from 51.41 in fifth grade to 49.84 in sixth grade. And, the experimental group mean increased from 48.67 in fifth grade to 50.15 in sixth grade.

The researcher used repeated measures analysis of variance (ANOVA) to determine whether there was a significant difference in the change on PSSA performance means for the experimental group versus the control group.

For Reading, the value of Pillai's Trace for the joint effect of PSSA change and group was .020. This was associated with a significant value for F ($F[1, 849] = 17.59$, $p <$ .0005). The interpretation of this result was that the change in means for the experimental and control groups were significantly different at the .0005 level. In other words, the experimental group, i.e. those students who participated in the Reading 4Sight Benchmark Assessments, showed greater improvement on the Reading PSSA than did the students in the control group who did not participate in the Reading 4Sight testing.

The effect size of the group difference for the Reading PSSA was calculated by finding the difference in mean change for the experimental and control groups and dividing the mean change by the pooled standard deviation to determine the size of the statistically significant difference observed between the groups. The resulting effect size

measure, Cohen's *d*, for the Reading PSSA was 0.20. According to Cohen and Cohen (1983), this would be considered a small effect size.

A corresponding repeated measures ANOVA was conducted to test whether changes in PSSA Math scores differed for the experimental group versus the control group. The interaction between group and Math score change was associated with a value for Pillai's trace of .051. This value corresponded to a significant F statistic ($F[1, 849] = 45.57$, $p < .0005$). This result was interpreted to mean that the experimental group, who participated in the Mathematics 4Sight Benchmark Assessments, showed greater improvement on the Mathematics PSSA than did the students in the control group who did not participate in the Mathematics 4Sight testing. The resulting effect size, Cohen's *d*, for the Mathematics PSSA was 0.31. According to Cohen and Cohen (1993), this would be considered a moderate effect size.

Summary

The data analyses indicated that a significant relationship existed between student performance on the 4Sight Benchmark Assessments and the PSSA. More specifically, a strong positive correlation existed between sixth grade students' performance on the Reading 4Sight Benchmark Assessment and the Reading PSSA. Additionally, a strong positive correlation existed between sixth grade students' performance on the Mathematics 4Sight Benchmark Assessment and the Mathematics PSSA. In relation to the study's secondary purpose, the experimental group of students who participated in the Reading and Mathematics 4Sight Benchmark Assessments showed significantly greater improvements than did the control group of students on both the Reading and

Mathematics PSSA. However, the effect size for the Reading was small, whereas the effect size for Math was moderate in size.

CHAPTER V

Summary, Interpretations and Recommendations

Summary of Method

The primary purpose of the quasi-experimental study was to determine whether there is a significant relationship between sixth grade students' 4Sight Benchmark Assessment scores and their sixth grade PSSA scaled scores in both reading and mathematics. The secondary purpose of the study was to determine whether participation in 4Sight Benchmark Assessments led to greater improvement on PSSA scores from fifth to sixth grade. The convenience sample consisted of a total of six school districts. Three control school districts not participating in the 4Sight benchmark assessments were match paired according to school size, percentage of low-income families, and school locale code classification with three school districts participating in the 4Sight benchmark assessments.

Both the experimental and control districts provided the researcher with Excel files containing the 2004-2005 fifth grade Reading and Mathematics PSSA scaled scores and 2005-2006 sixth grade Reading and Mathematics PSSA scaled scores. Additionally, the three experimental school districts provided the researcher with Excel files comprised of the 2005-2006 4Sight Reading and Mathematics Benchmark Assessment Scores for their sixth grade students. Incomplete student records were deleted from the Excel files and excluded from the study thus resulting in a total of 413 student records in the control group and 438 student records in the experimental group.

The results of this study illustrate that there is a strong correlation between student performance on both the 4Sight Reading and Mathematics Benchmark

Assessments and the Reading and Mathematics PSSA. Additionally, students who participated in the 4Sight Reading and Mathematics Benchmark testing demonstrated greater improvement on the Reading and Mathematics PSSA than did students who did not participate in the 4Sight Benchmark testing.

Interpretations

Primary Purpose

In relation to the study's primary purpose, there is a strong positive correlation between sixth grade student performance on the 4Sight Reading Benchmark Assessments and the Reading PSSA. Similarly, there is a strong positive correlation between sixth grade student performance on the 4Sight Mathematics Benchmark Assessments and the Mathematics PSSA.

The correlations between the 4Sight Reading and Mathematics Benchmark Assessments and Reading and Mathematics PSSA determined as a result of this study were .77 for sixth grade reading and .77 for sixth grade mathematics. The reading correlation of .77 is lower than the reading correlations of .89 and .87 for third grade, .88 and .86 for fifth grade, and .84 and .83 for 8th grade determined for two forms of the reading benchmark assessment by the Success For All Foundation (2005-2206). The mathematics correlation of .77 is slightly higher than the third grade mathematics correlation of .76 and lower than the fifth and eighth grade mathematics correlations, .82 and .84 respectively, determined by the Success for All Foundation (2005).

In educational measurement, the identified strength of correlation coefficients differs depending on the situation. Regarding the relative strengths of correlation coefficients Popham and Sirotnik (1991) elucidated:

> For example, in educational measurement, test constructors often must demonstrate the equivalence of two different forms of the same test. A group of students may be required to complete both forms of the test, and then their scores on the two measures are correlated. If the two test forms are to be considered equivalent, a high positive correlation between scores on the two forms, somewhere in the neighborhood of 0.90, would be expected. Yet in other educational situations, for example, relating academic achievement to a predictor test of achievement such as an aptitude test, an *r* of between 0.40 and 0.50 often is considered satisfactory. An *r* of 0.70 in such a situation would be exceptional indeed. (p.72)

Based on Popham 's and Sirotnik's (1991) correlation strength differentiation, it is appropriate to describe the reading correlation of .77 and mathematics correlation of .77 determined in this study as strong positive correlations.

Secondary Purpose

In relation to the study's secondary purpose, the experimental group of students who participated in the 4Sight Reading Benchmark Assessments showed greater improvement on the Reading PSSA than did the students in the control group who did not participate in the Reading 4Sight testing. Similarly, the experimental group of students who participated in the 4Sight Mathematics Benchmark Assessments showed greater improvement on the Mathematics PSSA than did the students in the control group who did not participate in the Mathematics 4Sight testing.

The magnitude of the positive effect of the 4Sight Benchmark Assessments in this study can be quantified in terms of effect sizes. The 4Sight Benchmark Assessments had

a small effect size of 0.20 for reading and a moderate effect size of 0.31 for math. In educational research, effect sizes of 0.20 are considered to be a minimum for significance whereas effect sizes that are greater than 0.50 are considered very strong (Slavin, 2003). Although the 4Sight Benchmark Assessment effect sizes determined as a result of this study meet the criteria for minimum significance, several factors may explain why the implementation of the 4Sight Reading and Mathematics Benchmark Assessments were associated with small to moderate effect sizes (Cohen & Cohen, 1983).

First, the Success For All Foundation does not consider the administration of the 4Sight Benchmark Assessments to be an effective intervention strategy in isolation (G. Holmes, personal communication, May 30, 2007). Instead, the Success For All Foundation recommends the administration of the benchmark assessments as one part of a more comprehensive school improvement process. The executive vice president and general manager of the K-12 Services Division of the Princeton Review agrees, "...testing kids without taking meaningful action is pointless and tiring. Coming in and working with assessments, and then following up with professional development with staff members, really resonates and shows marked improvement in student instruction" (Delisio, 2004, ¶ 4).

Educational research concerning assessment supports the belief that the administration of benchmark assessments and the review of the resulting data are key school improvement practices (Rand Corporation, 2006); however, other factors such as "data quality and access, data disaggregation, the role of collaborative inquiry in understanding data, and leadership structures that support data use" (Center for Comprehensive School Reform and Improvement, 2006, p.1) are critical to using data to

positively affect student achievement. Benchmark assessment data must be accurate and reliable and must be provided to teachers in a timely fashion. Additionally, the data must be disaggregated to provide specific information concerning the performance of student subgroups. Furthermore, teachers' schedules must permit time for collaborative discussion regarding student performance on the benchmark assessments to identify school, grade, subgroup, and individual student strengths and weaknesses.

Most importantly, however, school administrators and teachers must not only *understand* the student performance data produced as a result of the benchmark assessment administrations, but they must also, *act* on the information provided in the performance reports. Analyzing data and acting on data are two different steps in the school improvement process:

> Taking action is often more challenging and might require more creativity than analysis. Yet, to date, taking action generally receives less attention, particularly in the professional development provided to educators.
>
> School staff often lack not only the data analysis skills (e.g., knowledge of how to interpret test results), but also guidance in identifying solutions and next steps in addressing diagnosed problems. (Marsh, Pane, & Hamilton, 2006, p. 10)

In the present study, the experimental school districts were required to participate in data analysis and data action professional development sessions provided by the Success For All Foundation in collaboration with the Pennsylvania Training and Technical Assistance Network (PaTTAN). However, the number of teachers and administrators attending these sessions varied by district. Although the present study did not determine to what extent data analysis and data action steps occurred within each

experimental district as a result of the professional development sessions, it is likely that the extent to which data analysis and data action took place influenced the results of the study.

Additionally, it is important to recall that the 4Sight Benchmark Assessment initiative commenced in 2005-2006. The implementation of any new educational initiative involves the change process (Fullan, 1993) and the Concerns Based Adoption Model (1984, 1987). "Change is a journey, not a blueprint" (Fullan, 1993, p. 24). Additionally, the primary concerns of the individuals involved in the change process vary during the various phases of implementation. Hord, Rutherford, Huling-Austin, and Hall (1987) have identified three main stages of concerns. In the beginning stages of the implementation of a new educational initiative, the individuals involved primarily exhibit a concern for self. In a school setting, teachers in this first stage are mostly concerned with how the new initiative will affect them. The individuals then progress to the second stage identified as concern for task. During this stage, teachers are focused on time, materials, and resources. In the final stage, the individuals become focused on concern for impact. It is during this third and final stage when teachers begin to expand their concerns to their students. Furthermore, they begin to analyze the effectiveness of the initiative and also collaborate with other teachers.

Keeping the change process and the concerns of teachers and administrators involved with new educational initiatives in mind, it is important to reconsider the results of this study involving the administration of the 4Sight Benchmark Assessments. The low to moderate effect sizes determined as a result of this study may be explained by the early stages of the implementation of the 4Sight Benchmark Assessment initiative. Because

2005-2006 marked the first year of the 4Sight Reading and Mathematics Benchmark Assessment implementation in the experimental districts, administrators and teachers may have spent the majority of their time familiarizing themselves with the benchmark assessment administration and scoring process and reading student performance reports and little time discussing how to utilize the performance data to appropriately modify instruction and positively influence student achievement. It would be interesting to follow up with the experimental school districts in five to seven years near the end of the change process. During the final stages of implementation, teachers and administrators will be more likely to focus on a concern for impact. Consequently, the use of data to modify classroom instruction and district curricula may result in an increase in the effect sizes for the 4Sight benchmark assessments.

The results of this Pennsylvania study provide evidence that the 4Sight Reading and Mathematics Benchmark Assessments are strongly correlated to the Reading and Mathematics PSSA. Additionally, the results of the present study provide evidence that students who participated in the 4Sight Reading and Mathematics assessments showed greater improvement on the Reading and Mathematics PSSA. However, the results of this single study should be viewed with caution. Additional research concerning assessment administration, performance reporting, data analysis, and data action should be conducted to provide additional evidence regarding the effectiveness of the 4Sight Benchmark Assessments.

Recommendations

The number of Pennsylvania schools and districts using the 4Sight benchmark assessments has increased from 750 schools in 310 districts in 2005-2006 to 1,200

schools in 340 districts in 2006-2007 (Pennsylvania Training and Technical Assistance Network, 2007). Further research is required at the regional, state, and national level to validate the results of this Pennsylvania study. A replication study examining a different region of Pennsylvania or the entire state of Pennsylvania could confirm or refute the results of this study with regard to the effectiveness of the 4Sight Benchmark Assessments in predicting and increasing student PSSA performance. Additionally, because the 4Sight benchmark assessments are utilized in numerous states to predict students' performance on the state mandated assessments, a national study comparing the correlations and effect sizes of the 4Sight Reading and Mathematics Benchmark Assessments between states would increase the body of research concerning benchmark assessments and provide a more comprehensive perspective as to the effectiveness of the 4Sight Benchmark Assessments. Additionally, the results of a large-scale, national study can be used to inform curricular modifications and teaching methods.

Second, it is recommended that researchers examining future PSSA performance contact Pennsylvania Department of Education officials to collect PSSA scaled scores in reading and mathematics rather than contact individual school districts. Working with individual school districts to gain access to student PSSA data was often unsuccessful and time consuming. Despite the assurance of confidentiality provided by the researcher, several school districts originally identified for inclusion in the study declined participation due to confidentiality concerns. However, now that Pennsylvania Secure Identification Numbers (PA Secure IDs) have been assigned to all students in Pennsylvania, researchers can contact the Pennsylvania Department of Education to access school and district level PSSA data without any identifying student information.

A third recommendation concerning the design of the study would be to control for additional threats to internal validity. Because there was no possibility for random assignment in the present study, the researcher utilized a quasi-experimental design which controlled for the percentage of low income students, enrollment, and school locale code classification. More specifically, to strengthen inferences about causality, it is recommended that future studies match control and experimental districts not only using percentage of low income students, enrollment, and school locale classification, but also using level of faculty motivation, professional development, and/or administrative integrity.

Future studies could include more sophisticated levels of data analysis. A multi-factor Analysis of Variance (MANOVA), rather than the repeated measures analysis of variance (ANOVA) may provide additional insight as to whether participation in the 4Sight Reading and Mathematics Benchmark Assessments is positively related to increases in student achievement on other forms of standardized assessment.

Also, since 2007-2008 will mark the third year of 4Sight Benchmark Assessment implementation for the first cohort of school districts, a longitudinal study examining and comparing PSSA performance in these school districts is warranted to determine whether achievement increases directly with years of implementation. Similarly, a longitudinal study of the school districts included in this study which compares the benchmark assessment effect sizes calculated in later stages of implementation to the effect sizes calculated in the first year of implementation would assist in determining whether the 4Sight Benchmark Assessment effect sizes increase directly with years of implementation.

Finally, the results of this study suggest that the administration of benchmark assessments may positively impact student achievement on state assessments; however, teachers must analyze the benchmark assessment data, make changes to classroom instruction, and provide feedback to students in order maximize the effectiveness of the benchmark assessment initiative within their classrooms (Black & William, 1998; Guskey, 2005; Linn, Baker, and Dunbar, 1991; McTighe & O'Connor, 2005; Stiggins & Chappuis, 2005). It is suggested that teachers using the 4Sight Reading and Mathematics Benchmark Assessments within their classrooms be surveyed to determine the extent to which they use the student performance data to determine skill strengths and weaknesses, modify lesson plans and pacing, and provide student feedback.

REFERENCES

Achieve, Inc. (2005). Measuring Up 2005: A Report on Assessment Anchors and Tests in Reading and Mathematics for Pennsylvania. Washington, DC: Achieve, Inc.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education). (1999). *Standards for Educational and Psychological Tests*. Washington, DC: American Educational Research Association.

Aguilera, R. V. (1994). Iowa test of basic skills as a predictor of the Arizona student assessment performance instruments. (Doctoral Dissertation, Northern Arizona University, 1995). *Dissertation Abstracts International,* 55(10), 3043. (UMI No. 9505454)

Amrein, A. L., & Berliner, D. C. (2002). High stakes testing, uncertainty and student learning. *Education Policy Analysis Archives*, 10(18). Retrieved March 10, 2006 from: http://epaa.asu.edu/epaa/v10n18/

Archer, J. (2005). Guiding hand. *Supplement to the September 14, 2005 issue of Education Week*, 25(3), S5-10. Retrieved March 26, 2006 from: http://www.edweek.org/media/wallace.pdf

Assessment Reform Group. (1999). *Assessment for learning: Beyond the black box.* Cambridge: University of Cambridge School of Education.

Berger, N. & Elson, H. H. (1996). *What happens when MCT's are used as an accountability device: Effects on teacher autonomy, cooperation and school mission.* Paper presented at the annual meeting of the American Educational Research Association, New York.

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

Bishop, J. H. (2001). A steeper, better road to graduation. *Education Next,* (4), 56-61.

Black, P. (2000). Research and the development of educational assessment. *Oxford Review of Education*, 26(3/4), 407-421.

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

Braun, H. (2004). Reconsidering the impact of high-stakes testing. *Education Policy Analysis Archives*, 12(1). Retrieved March 10, 2006 from: http://epaa.asu.edu/epaa/v12n1/

Burson, K.C. (2001). A correlation analysis of Pennsylvania System of School Assessment, cognitive abilities tests and report card grades for students in grades 3, 5, 8, and 11. (Doctoral Dissertation, Widener University, 2002). *Dissertation Abstracts International,* 62(12), 4003. (UMI No. 3036998)

Carnoy, M., & Loeb, S. (2003). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis,* 24(4), 305-331.

The Center for Comprehensive School Reform and Improvement. (2006, October). *Practices that support data use in urban high schools*. Washington, DC.

Center on Education Policy (2003). Effects of High School Exit Exams on Dropout Rates: Summary of a Panel Discussion Held on March 15, 2003. Washington, DC.

Cimbricz, S. (2002, January). State-mandated testing and teachers' beliefs and practice.

Education Policy Analysis Archives, 10(2). Retrieved: March 12, 2006 from: http://epaa.asu.edu/epaa/v10n2.html

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement, Issues and Practice*, 20(4), 19-28.

Clarke, M., Shore, A., Rhoades, K., Abrams, L., Miao, J. & Li, J. (2003, January). *Perceived effects of state-mandated programs on teaching and learning: Findings from interviews with educators in low-, medium-, and high-stakes states*. Boston, MA: National Board on Educational Testing and Public Policy. Retrieved March 6, 2006 from: http://www.bc.edu/research/nbetpp/statements/nbr1.pdf

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Corbett, H. D., & Wilson, B. (1990). Unintended and unwelcome: The local impact of state testing. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Corbett, H. D., & Wilson, B. L. (1991). Testing, reform, and rebellion. Norwood, NJ: Ablex.

CRESST Assessment Glossary. (1999). Retrieved on February 23, 2006 from: http://www.cse.ucla.edu/CRESST/pages/glossary.htm

Data Recognition Corporation. (2004). *Fairness in Testing: Training Manual for Issues of Bias, Fairness, and Sensitivity*. Maple Grove, MN: DRC.

Data Recognition Corporation. (2005, December). *Technical report for the Pennsylvania System of School Assessment 2005 Reading and Mathematics*. Retrieved August 15, 2006 from http://www.pde.state.pa.us/a_and_t/cwp/view.asp?a=108&Q=

108328&a_and_tNav=|6395|

Davenport, E. C., Davison, M. L., Kwak, N., Irish, M. L., & Chan, C. K. (2002). Minnesota high stakes high school graduation test and completion status for the class of 2000. Office of Educational Accountability, College of Education and Human Development, University of Minnesota.

Delisio, E. R. (2004). Making data work for your school. *Education World.* Retrieved July 16, 2007 from: http://www.education-world.com/a_admin/admin/admin377.shtml

Florida Department of Education. (2005). *Florida Comprehensive Assessment Test.* Retrieved February, 12, 2006 from http://fcat.fldoe.org/index.asp#reports

Fuchs, L. (1986). Effects of task-focused goals on low-achieving students with and without learning disabilities. *American Educational Research Journa*l, 34, 199-208.

Fullan, M. (1993). *Change forces: Probing the depths of educational reform.* Bristol, PA: The Falmer Press.

Gerstner. L. (2004, October 7). Bad school plus shackled principals equals outsourcing. *The Wall Street Journal*, p. A18.

Guskey, T. (2005) Mapping the road to proficiency. *Educational Leadership*, 63(3), 32-38.

Hall, D. (2005). *Stalled in secondary: A look at student achievement since the No Child Left Behind Act.* Washington, DC: The Education Trust.

Herman, J. (2005). *Making accountability work to improve student learning* (CSE Technical Report 649). Los Angeles, University of California: National Center for

Research on Evaluation, Standards, and Student Testing (CRESST).

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

Herman, J. L., & Golan, S. (Undated). Effects of standardized testing on teachers and learning-another look (CSE Technical Report 334). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CREEST), University of California.

Herman, J. L. & Perry, M. (2002, June). California Student Achievement: Multiple views of K-12 progress. Menlo Park, CA: Ed Source.

Hershberg, T. (2005, February). *Value-added assessment and systemic reform: a response to America's human capital development challenge.* Symposium conducted at the Aspen Institute's Congressional Institute, Cancun, Mexico.

Hord, S. M., Rutherford, W. L., Huling-Austin, L., & Hall, G. E. (1987). *Taking charge of change.* Alexandria, VA: Association for Supervision and Curriculum Development.

Horn, C. (2003). High-stakes testing and students: stopping or perpetuating a cycle of failure? *Theory Into Practice.* Retrieved March 5, 2006 from: http://www.findarticles.com/p/articles/mi_m0NQM/is_1_42/ai_99909368/print

Jacob, B. A. (2001). Getting tough? The impact of high school graduation exams. *Educational Evaluation and Policy Analysis,* 23(2).

King, M. D. (2003). The effects of formative assessment on student self-regulation, motivational beliefs, and achievement in elementary science. (Doctoral

Dissertation, George Mason University, 2003). *Dissertation Abstracts International,* 64(02), 403. UMI No. 3079342

Kober, N. (2002, June). *Teaching to the test: The good, the bad, and who's responsible. Test Talk for Leaders.* Washington, DC: Center on Education Policy. Retrieved March 10, 2006 from: http://www.cep-dc.org/testing/testtalkjune2002.pdf

Koretz, D., Barron, S., Mitchell, K. and Stecher, B. (1996). *Perceived effects of the Kentucky instructional results information system.* Retrieved March 15, 2006 from: http://www.rand.org/cgi-bin/Abstracts/e-getabbydoc.pl?MR-792-PCT/FF

Koretz, D., Linn R. L., Dunbar, S. B., & Shepard, L. A. (1991). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests.* Presented in R. L. Linn (Chair), Effects of High-Stakes Educational Testing on Instruction and Achievement symposium presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April 5, 1991

Lagenfeld, K. L., Thurlow, M. L., & Scott, D. L. (1996). High stakes testing for students: Unanswered questions and implications for students with disabilities (Synthesis Report No. 26. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved March 20, 2006, from:http://education.umn.edu /NCEO/OnlinePubs/Synthesis26.htm

Lauman, P. G. (2001). The ability of one district's second-grade literacy assessments and other variables to predict student proficiency on the third-grade reading Colorado Student Assessment Program (CSAP). (Doctoral Dissertation. University of

Northern Colorado, 2001). *Dissertation Abstracts International*, 62 (09), 3002. (UMI No. 3025099)

Lehmier, M. J. (2001). Predicting fifth grade mathematics scores on the Pennsylvania System of School Assessment using a computer integrated learning system. (Doctoral Dissertation, Temple University, 2001). *Dissertation Abstracts International,* 62(05), 1659. (UMI No. 3014458)

Leskes, A. (2002). Beyond confusion: an assessment glossary [electronic version]. *Peer Review,* 4(2/3). Retrieved February 23, 2006 from: http://www.aacu.org/peerreview/pr-sp02/pr-sp02reality.cfm

Linn, R. (2001, April). *The design and evaluation of educational assessment and accountability systems* (CSE Technical Report 539). Los Angeles, University of California: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15-21. (ERIC Document Reproduction Service No. EJ 436 999)

Maher, T. R. (2003). The predicted validity of the Pennsylvania System of School Assessment using the CTB McGraw-Hill Terra Nova test. (Doctoral Dissertation, Widener University, 2004). *Dissertation Abstracts International,* 64(09), 3150. (UMI No. 3103752)

Manhattan Institute for Policy Research. (2003, February). *Testing high stakes tests: Can we believe the results of accountability tests?* Manhattan, NY: Greene, J.P., Winters, M.A., & Forster, G.

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

Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2001). *Classroom instruction that works* Alexandria, VA: Association for Supervision and Curriculum Development.

McMillan, J. H. (2000). *Educational research: Fundamentals for the consumer.* New York: Longman.

McTighe, J. & O'Connor, K. Seven practices for effective learning. *Educational Leadership*, 63(3), 10-17.

Morris, B. (2004). Informing reading in an era of accountability: The relationship between formative assessments and high stakes testing. (Doctoral Dissertation, University of Virginia, 2004). *Dissertations Abstracts International*, 65 (04), 1302. (UMI No. 3131462).

Muir, M. (2001). When stakes are high. *Northwest Education Magazine*. Retrieved March 26, 2006 from: http://www.nwrel.org/nwedu/2001fall/stakes.html

Mulvenon, S. W., Connors, J. V., & Lenares, D. (2001). Impact of accountability and school testing on students: Is there evidence of anxiety? Paper presented at the Annual Meeting of the Mid-South Educational Research Association November 13-16, 2001. Little Rock, AK. (ERIC Document Reproduction Service No. 460155)

National School Boards Association. (2004).*Research review: Exploring the effects of high-stakes testing on instruction.* Retrieved March 2, 2006 from: http://www.nsba.org/site/sec_peac.asp?TRACKID=&CID=1242&DID=37830

National School Boards Association. (2004).*The effects of high-stakes testing on instruction: Key lessons learned.* Retrieved March 2, 2006 from: http://www.nsba.org/site/sec_peac.asp?TRACKID=&DID=37856&CID=1242

New York State Education Department. (2004, August). The impact of high-stakes exams on students and teachers. Retrieved March 5, 2006 from http://www.oms. nysed.gov/faru/TheImpactofHighStakesExams.htm

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

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

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

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., Miao, J. (2003, March). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers.* Boston, MA: National Board on Educational Testing and Public Policy. Retrieved March 6, 2006 from: http://www.bc.edu/research/nbetpp/statements/nbr2.pdf

Pennsylvania Assessment Through Themes (PATT), Pennsylvania Department of Education, Division of Evaluation and Reports, Division of Arts, Sciences, Communication and Mathematics. (1999, September). *Fairness policies, guidelines and procedures for developing fair assessments.*

Pennsylvania Department of Education. (2004, April). *Assessment Anchors and Eligible Content.* Harrisburg, PA: PDE. Retrieved August 15, 2006 from http://www.pde.state.pa.

Pennsylvania Department of Education. (2005). *PSSA Results*. Retrieved February 12, 2006 from: http://www.pde.state.pa.us/a_and_t/cwp/browse.asp?a=3&bc=0&c=27525&a_and_tNav=|633|&a_and_tNav=|

Pennsylvania State Board of Education. (1999, January). *Chapter 4. Academic Standards and Assessment*. Harrisburg, PA: Pennsylvania State Board of Education. Retrieved August 15, 2006 from http://www.pde.state.pa. Also available from http://www.pacode.com/secure/data/022/Chapter4/s4.51.html.

Pennsylvania Training and Technical Assistance Network. (2007). *Assessing to learn: PA benchmark initiative*. Retrieved July 10, 2007 from: http://www.pattan.k12.pa.us/teachlead/AssessingtoLearn.aspx

Pierce, C. A. (2005). An examination of student performance on California standards-based tests in mathematics as indicators for future performance on the mathematics portion of the California High School Exit Exam. (Doctoral Dissertation, Alliant International University-San Diego, 2006). *Dissertation Abstracts International,* 66(07), 2521. (UMI No. 3184405)

Popham, W. J., & Sirotnik, K. A. (1991). *Understanding Statistics in Education*. Itasca, IL: F. E. Peacock Publishers, Inc.

Raymond, M. E., & Hanushek, E. A. (2003). High stakes research. *Education Next*, 3(3), 48-55).

Reeves, D. (2002). Accountability based reforms should lead to better teaching and learning—period. *Harvard Education Letter,* March/April 2002. Retrieved March 25, 2006 from: http://www.edletter.org/past/issues/2002-ma/reeves.shtml

Rex, S. L. (2003). Reading strategies as a predictor of student scores on the Pennsylvania

System of School Assessment reading exam at the eleventh-grade level. (Doctoral Dissertation, Widener University, 2004). *Dissertation Abstracts International,* 64(12), 4409. (UMI No. 3116335)

Rosenshine, B. (2003). High-stakes testing: Another analysis. *Education Policy Analysis Archives*, 11(24). Retrieved Feb. 15, 2006 from: http://epaa.asu.edu/epaa/v11n24/

Rottenberg, C., & Smith, M. L. (1990). Unintended effects of external testing in elementary schools. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education*, 5(1), 77-85.

The School Board of Broward County, Florida. (2005, July). *The benchmark assessment test: Predicting FCAT proficiency* (Research Brief No. 99). Broward County, FL: Clement, R.

Sebatane, E. M. (1998). Assessment and classroom learning: A response to Black and Wiliam. *Assessment in Education*, 5(1), 123-131.

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

Shepherd, L. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4-14.

Shepherd, L. A., & Dougherty, K. (1991). Effects of high-stakes testing on instruction.

Paper presented at the annual meeting of the American Education Research Association and the National Council on Measurement in Education, Chicago.

Shields, R. N. (2000). Writing strategies as predictors of student scores on the Pennsylvania System of School Assessment writing test. (Doctoral Dissertation, Widener University, 2000). *Dissertation Abstracts International,* 61(08), 3043. (UMI No. 9980729)

Sicoly, F. (2002). What do school-level scores from large-scale assessments really measure? Educational Measurement: Issues and Practice, 21(4), 17-26.

Sirotnik, K. S., & Kimball, K. (1999). Standards for standards-based accountability systems. *Phi Delta Kappan*, 81(3), 209-214.

Slavin, R.E. (2003, February). A reader's guide to scientifically based research. Educational Leadership, 60(5), p. 12-16.

Sloane, F. C., & Kelly, A. E. (2003). Issues in high-stakes testing programs. *Theory Into Practice*, 42(1), 12-17. Retrieved March 10, 2006 from: http://muse.jhu.edui/demo/theory_into_practice/v042/42.1sloane.ht

Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement, Issues and Practice*, 10, 7-11.

Spellings, M. (2005). *Is America really serious about educating every child?* (Prepared remarks for Secretary Spellings at the Education Writers Association National Seminar, St. Petersburg, FL, May 6, 2005). Retrieved February 20, 2006 from: http://www.ed.gov/news/speeches/2005/05/05062005.html

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, 83(10), 758-765.

Stiggins, R. J., Arter. J., Chappuis, J., & Chappuis, S. (2004). *Classroom Assessment for student learning: Doing it right—using it well.* Portland, OR: Assessment Training Institute

Stiggins, R. J., & Chappuis, S. (2005). Putting testing in perspective: It's for learning. *Principal Leadership*, 6(2), 16-20.

Success for All Foundation. (2005-2006). *4Sight reading benchmark version 2005-2006, second edition* (Technical Report for Pennsylvania). Baltimore, MD.

Success for All Foundation. (2005). *Elementary and Middle School 4Sight Assessments.* Retrieved February 22, 2006 from: http://www.successforall.net/elementary/4sight.htm and http://www.successforall.net/middle/4sight.htm

Swanson, C. B. (2006). Making the connection: A decade of standards-based reform and achievement. Editorial Projects in Education Research. Retrieved March 10, 2006 from: http://www.edweek.org/media/ew/qc/2006/MakingtheConnection.pdf

Tapper, R. (1997). *The problem of high-stakes assessment in public education.* (Eric Document Reproduction Service No. ED430021)

Texas Education Agency. (2005). *2005 Accountability Rating System.* Retrieved February 12, 2006 from http://www.tea.state.tx.us/perfreport/account/2005/index.html

Thacker, A. A., Dickinson, E. R., & Koger, M. E. (2004). *Relationships among the Pennsylvania System of School Assessment (PSSA) and other commonly administered assessments.* Louisville, KY: Human Resources Research Organization.

Thompson, S., Johnstone, C. J. & Thurlow, M. L. (2002). *Universal Design Applied to Large Scale Assessments* (Synthesis Report 44), Minneapolis, MN: University of

Minnesota, National Center on Educational Outcomes.

United States Department of Education. (February 3, 2003). *Fiscal Year 2004 Budget Summary*. Retrieved February 20, 2006 from: http://www.ed.gov/about/overview/budget/budget04/ summary/edlite-section1.html

United States Department of Education. (June 10, 2003). President Bush, Secretary Paige Celebrate Approval of Every State Accountability Plan Under No Child Left Behind. *United States Department of Education Press Release*. Retrieved February 20, 2006 from the World Wide Web: http://www.ed.gov/news/pressreleases/2003/06/06102003.html

United States Department of Education. (July 2004). *Glossary of terms*. Retrieved February 15, 2006 from: http://www.ed.gov/nclb/index/az/glossary.html

Van Evera, W. C. (2004). Achievement and motivation in the middle school science classroom: the effects of formative assessment feedback. (Doctoral Dissertation, George Mason University, 2004). *Dissertation Abstracts International,* 64(10), 3637. (UMI No. 3110082)

Waddell, C. A. (2004). The effects of negotiated written feedback within formative assessment on fourth-grade students' motivation and goal orientations. (Doctoral Dissertation, University of Missouri-Saint Louis, 2005). *Dissertation Abstracts International,* 65(10), 3697. (UMI No. 3151330)

Walkovic, C. E. (2003). Reading strategies as predictors of school scores on the

Pennsylvania System of School Assessment reading test. (Doctoral Dissertation, Widener University, 2004). *Dissertation Abstracts International,* 64(12), 4409. (UMI No. 3116336)

Warren, J. R., & Edwards, M. R., (2003). The impact of high stakes graduation tests on high school diploma acquisition. Paper presented at the Fourth Annual Undergraduate Research Symposium at the University of Washington.

Yeh, S. S. (2005). Limiting the unintended consequences of high-stakes testing. *Education Policy Analysis Archives,* 13(43), 1-23.

Appendix A: Convenience Sample & Match Districts

Table 1

School District	District Enrollment	Percentage of Low Income Families	School Locale Code
District A	2083	5.8%	Rural
Match District A	2311	6.8%	Rural
District B	1999	12.7%	Rural
Match District B	1852	12.6%	Urban Fringe
District C	1743	25.8%	Urban Fringe
Match District C	2094	26.7%	Urban Fringe

Appendix B: Pennsylvania 4Sight Reading Benchmark Assessment Descriptive Statistics and Correlation to the 2005 Reading PSSA

Table 2

3rd Grade Form 1				
	N	**Mean**	**SD**	**R**
4Sight	718	20.41	5.92	0.89
PSSA	718	1262.39	194.62	

Table 3

3rd Grade Form 2				
	N	**Mean**	**SD**	**R**
4Sight	699	20.47	5.75	0.87
PSSA	699	1259.43	195.30	

Table 4

5th Grade Form 1				
	N	**Mean**	**SD**	**R**
4Sight	449	19.17	5.89	0.88
PSSA	449	1200.86	204.44	

Table 5

5th Grade Form 2				
	N	**Mean**	**SD**	**R**
4Sight	345	18.20	5.74	0.86
PSSA	345	1193.76	211.71	

Table 6

8th Grade Form 1				
	N	**Mean**	**SD**	**R**
4Sight	381	18.15	5.45	0.84
PSSA	381	1178.09	221.64	

Table 7

8th Grade Form 2				
	N	**Mean**	**SD**	**R**
4Sight	371	20.03	5.34	0.83
PSSA	371	1198.32	218.81	

Success for All Foundation. (2005-2006). *4Sight reading benchmark version 2005-2006, second edition* (Technical Report for Pennsylvania). Baltimore, MD.

Appendix C: Pennsylvania 4Sight Mathematics Benchmark Assessment Descriptive Statistics and Correlation to the 2005 Mathematics PSSA

Table 8

3rd Grade Form 1				
	N	**Mean**	**SD**	**R**
4Sight	96	33.20	5.53	0.76
PSSA	96	1430.05	150.79	

Table 9

5th Grade Form 1				
	N	**Mean**	**SD**	**R**
4Sight	99	24.29	5.54	0.82
PSSA	99	1290.43	129.35	

Table 10

8th Grade Form 1				
	N	**Mean**	**SD**	**R**
4Sight	102	19.38	5.92	0.84
PSSA	102	1513.37	174.54	

J.B. Anderson, personal communication, August 3, 2006.

Appendix D: Pennsylvania 4Sight Reading Benchmark Assessment Inter-form Reliability

Table 11

Benchmark Assessment	Pearson Correlation	N of Items
PA Form 3.1		30
PA Form 3.2	0.829	30
PA Form 5.1		30
PA Form 5.2	0.865	30
PA Form 8.1		30
PA Form 8.2	0.793	30

Success for All Foundation. (2005-2006). *4Sight reading benchmark version 2005-2006, second edition* (Technical Report for Pennsylvania). Baltimore, MD

Appendix E: 2005 Reading and Mathematics PSSA Test Plan per Operational Form

(16 Forms: Forms 1-16)

Table 12

Reading PSSA									
Grade	**Core MC**	**Matrix MC**	**Embedded FT MC**	**Core 3-pt. OE**	**Matrix OE**	**Embedded FT OE**	**Total Items MC/OE**	**Passages**	**Total Operational Points**
5	40	8	8	4	1	1	56/6	6-8	52
8	40	8	8	4	1	1	56/6	6-8	52
11	40	8	8	4	1	1	56/6	6-8	52

Table 13

Mathematics PSSA								
Grade	**Core MC**	**Matrix MC**	**Embedded FT MC**	**Core 3-pt. OE**	**Matrix OE**	**Embedded FT OE**	**Total Items MC/OE**	**Total Operational Points**
5	54	4	8	3	1	1	66/5	52
8	54	4	8	3	1	1	66/5	52
11	54	4	8	3	1	1	66/5	52

Data Recognition Corporation. (2005, December). *Technical report for the Pennsylvania System of School Assessment 2005 Reading and Mathematics.* Retrieved August 15, 2006 from http://www.pde.state.pa.us/a_and_t/cwp/view.asp?a=108&Q=108328&a_and_tNav=|6395|

Appendix F: 2005 Grade 5 PSSA Descriptive Statistics and Reliability

Table 14

Reading PSSA				
Strand	**N**	**Mean**	**SD**	**Reliability**
Overall	134142	35.87	9.52	0.91
A) Comprehension & Reading Skills	134412	26.03	6.99	0.88
B) Interpretation and Analysis of Fiction and Nonfiction	134412	9.84	2.96	0.73

Table 15

Mathematics PSSA				
Strand	**N**	**Mean**	**SD**	**Reliability**
Overall	134322	47.21	12.31	0.92
A) Numbers and Operations	134322	20.87	5.85	0.87
B) Measurement	134322	5.77	2.19	0.46
C) Geometry	134322	6.59	2.03	0.65
D) Algebra	134322	7.07	2.08	0.60
E) Data Analysis and Probability	134322	6.91	2.01	0.69

Data Recognition Corporation. (2005, December). *Technical report for the Pennsylvania System of School Assessment 2005 Reading and Mathematics.* Retrieved August 15, 2006 from http://www.pde.state.pa.us/a_and_t/cwp/view.asp?a=108&Q=

108328&a_and_tNav=|6395|

Appendix G: 2005 Grade 8 PSSA Descriptive Statistics and Reliability

Table 16

Reading PSSA				
Strand	**N**	**Mean**	**SD**	**Reliability**
Overall	145752	39.12	9.07	0.91
A) Comprehension & Reading Skills	145752	26.89	6.13	0.88
B) Interpretation and Analysis of Fiction and Nonfiction	145752	12.23	3.35	0.74

Table 17

Mathematics PSSA				
Strand	**N**	**Mean**	**SD**	**Reliability**
Overall	145999	43.97	13.69	0.93
A) Numbers and Operations	145999	8.38	3.07	0.67
B) Measurement	145999	6.35	2.57	0.73
C) Geometry	145999	8.40	2.73	0.74
D) Algebra	145999	13.00	4.26	0.80
E) Data Analysis and Probability	145999	7.84	2.78	0.70

Data Recognition Corporation. (2005, December). *Technical report for the Pennsylvania System of School Assessment 2005 Reading and Mathematics.* Retrieved August 15, 2006 from http://www.pde.state.pa.us/a_and_t/cwp/view.asp?a=108&Q=

108328&a_and_tNav=|6395|

Appendix H: 2005 Grade 11 PSSA Descriptive Statistics and Reliability

Table 18

Reading PSSA				
Strand	**N**	**Mean**	**SD**	**Reliability**
Overall	129693	38.75	9.51	0.92
A) Comprehension & Reading Skills	129693	21.56	5.14	0.88
B) Interpretation and Analysis of Fiction and Nonfiction	129693	17.19	4.82	0.83

Table 19

Mathematics PSSA				
Strand	**N**	**Mean**	**SD**	**Reliability**
Overall	129962	39.89	15.17	0.94
A) Numbers and Operations	129962	6.34	2.26	0.73
B) Measurement	129962	5.87	2.33	0.73
C) Geometry	129962	6.56	2.96	0.67
D) Algebra	129962	14.46	6.61	0.88
E) Data Analysis and Probability	129962	6.66	2.75	0.65

Data Recognition Corporation. (2005, December). *Technical report for the Pennsylvania System of School Assessment 2005 Reading and Mathematics.* Retrieved August 15, 2006 from http://www.pde.state.pa.us/a_and_t/cwp/view.asp?a=108&Q=

108328&a_and_tNav=|6395|

Appendix I: Huynh Decision Indices for All 2005 PSSA Performance Levels

Table 20

	All Performance Levels
Reading Grade 5	0.73667
Reading Grade 8	0.73936
Reading Grade 11	0.75620
Math Grade 5	0.67698
Math Grade 8	0.69066
Math Grade 11	0.72320

Data Recognition Corporation. (2005, December). *Technical report for the Pennsylvania System of School Assessment 2005 Reading and Mathematics.* Retrieved August 15, 2006 from http://www.pde.state.pa.us/a_and_t/cwp/view.asp?a=108&Q=108328&a_and_tNav=|6395|

Appendix J: E-mail Cover Letter and Informed Consent Form Sent to District-Level Administrators

Pennsylvania 4Sight Math & Reading Benchmark Assessments:
How Accurately do they Predict Reading & Mathematics PSSA Performance?

A Lehigh University Doctoral Dissertation by Christina Lutz-Doemling
Director of Curriculum & Assessment
Catasauqua Area School District

September 2006

Dear Administrator:

Your school district is presently participating in the Pennsylvania Reading and Mathematics 4Sight Benchmark Assessment Initiative. I am contacting you to invite your district to participate in a doctoral research study to determine the effectiveness of the Pennsylvania Reading and Mathematics 4Sight Benchmark Assessments as predictors of student performance on the Pennsylvania System of State Assessment (PSSA). Your school district's participation in this project will assist in providing answers to the following research questions:

Primary Research Questions:

1. Is there a significant relationship between the 2005-2006 4Sight Reading Benchmark Assessment Scores and PSSA Reading scaled scores at the sixth grade level?
2. Is there a significant relationship between the 2005-2006 4Sight Mathematics Benchmark Assessment Scores and PSSA Mathematics scaled scores at the sixth grade level?
3. Is there any significant difference between sixth grade student gain scores on the 2006 Reading PSSA for the students who participated in the 4Sight Reading Benchmark Assessments as compared with a matched group that did not participate in the 4Sight Reading testing?
4. Is there any significant difference between sixth grade student gain scores on the 2006 Mathematics PSSA for the students who participated in the 4Sight Mathematics Benchmark Assessments as compared with a matched group that did not participate in the 4Sight Mathematics testing?

I expect that the results of this research will be meaningful to you as you evaluate the effectiveness of the implementation of the Pennsylvania 4Sight Benchmark Assessment Initiative within your school district. Should you decide to participate in this study, you will be expected to provide the district's 2005-2006 sixth grade 4Sight Reading and Mathematics Benchmark Assessment results to the researcher. Additionally, your district will provide the researcher with the 2004-2005 Grade 5 PSSA Reading and Mathematics Scaled Scores and 2005-2006 Grade 6 PSSA Reading and Mathematics Scaled Scores.

I am hopeful that you will support my research by volunteering to participate in this study. Please be advised that at no point will you or your district be identified by name or by other indicators. The purpose of the study is not to identify the results of a particular district, but to determine the effectiveness of the 4Sight Benchmark Assessments as predictors of student performance on the PSSA.

Thank you for your time and consideration of this doctoral research study.

Sincerely,

Christina Lutz-Doemling
Director of Curriculum & Assessment, *Catasauqua Area School District*
Educational Leadership Doctoral Student, *Lehigh University*

INFORMED CONSENT

Description of the research: "Pennsylvania 4Sight Reading and Mathematics Benchmark Assessments: How Accurately do they Predict Reading and Mathematics PSSA Performance" is a doctoral research project that will primarily determine the relationship between the Pennsylvania 4Sight Reading and Mathematics Benchmark Assessment scores and Reading and Mathematics PSSA scaled scores.

The first phase of data collection for this study will involve gathering data pertaining to your 6th grade students' performance on the 2005-2006 Reading and Mathematics 4Sight Benchmark Assessments. The second phase of data collection will involve gathering 5th grade students' scaled scores on the 2005 Reading and Mathematics PSSA and 6th grade students' scaled scores on the 2006 Reading and Mathematics PSSA.

The data collected from this study will be utilized to determine the relationship between students' performance on the Reading and Mathematics 4Sight Benchmark Assessments and their performance on the Reading and Mathematics PSSA. It also will inform the design of future district and building level assessment plans. Research findings may be disseminated through publication in professional journals, presentations at professional meetings, and use in professional development curricula.

Participation in the Study: As a participant in this study, your district will provide the researcher with the 2005-2006 sixth grade 4Sight Reading and Mathematics Benchmark Assessment results. Following the instructions below, it will take approximately five minutes to export the 4Sight Benchmark Assessment results from the 4Sight online Member Center:

- Log into the 4Sight Member Center (https://members.successforall.net) using your district username and password
- Choose the school with your 6th Grade data
- Click "Admin" on the left side of the screen and then Click "Data Extracts"
- Choose EX-3524 Student Test Results
- Include the following data:
  School Yr. 2005-2006
  4Sight
  Reading
  Homeroom Teacher
  Grade 6
- Click "Create Extract"
- To ensure confidentiality, delete only the student names from the file. Please retain the student identification numbers in the file.
- Save the Excel file to your computer
- Repeat the steps above to complete the extraction of the Grade 6 4Sight Mathematics scores.
- Email both files to lutzdoemlingc@cattysd.org as Excel attachments

Additionally, as a participant in this study, your district will provide the researcher with the 2004-2005 Grade 5 PSSA Reading and Mathematics Scaled Scores and 2005-2006 Grade 6 PSSA Reading and Mathematics Scaled Scores. Following the instructions below, it will take approximately five minutes to export the PSSA results from the Data Recognition Center (DRC):

- Log into the DRC Web Reporting System (www.drc-web.com/reportdelivery) using your district user ID and password
- Click "Reports" at the top of the screen
- Click "2005 Grade 5 Math Reading District Data File"
- Delete all student information except for the columns containing the District Student Identification Numbers, Reading PSSA Scaled Scores, and Mathematics PSSA Scaled Scores
- Save the Excel file to your computer
- Repeat the steps above to complete the extraction of the "2006 Grade 6 Math Reading District Data File"
- Email both files to lutzdoemlingc@cattysd.org as Excel attachments.

Confidentiality: The data collected for use in this study will be kept strictly confidential, with individual student identities and the identities of the schools and districts remaining confidential. Only combined data will be reported and no indicators that would permit identification of you, your students or the district will be included. All data will be kept in a locked file cabinet with identification codes kept in a separate location. At the conclusion of the study, the code sheet will be destroyed.

Risks and benefits: There are no foreseeable risks to you from participating in this study, but there are several benefits. You will gain insight into the effectiveness of the Pennsylvania 4Sight Mathematics and Reading Benchmark Assessment scores as predictors of PSSA scaled scores for students in your school district as well as in the other districts participating in this study. Additionally, the results of this research will benefit the Pennsylvania educational community by informing the future design of district and building assessment plans.

The right to withdraw: Your participation is voluntary; refusal to participate will involve no penalty or loss of benefits to which you are otherwise entitled. You may discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled.

Referrals for questions: For answers to questions about the research you may contact the Principal Investigator of the Research Study, Christina Lutz-Doemling at 610.264.5571; lutzdoemlingc@cattysd.org. If you have a concern with the study, please contact Dr. George P. White my doctoral advisor at 610-758-3262; gpw1@lehigh.edu or Ruth Tallman, Office of Research at Lehigh University at 610-758-3024; rt01@lehigh.edu.

To confirm that you have read and understand the foregoing information, that you have received answers to any questions you asked, and to indicate whether you consent to participate in the study, please check the appropriate box and sign below.

☐ Yes, I agree to have my school district participate in this doctoral research study.

☐ No, I do not agree to have my school district participate in this doctoral research study.

Signature Title Date

Please fax the signed informed consent form to: Christina Lutz-Doemling, Director of Curriculum & Assessment, Catasauqua Area School District, Fax #: 610-264-5618.

VITA

Christina K. Lutz-Doemling

Christina K. Lutz-Doemling, daughter of Ronald and Audrey Lutz, was born in Allentown, Pennsylvania on November 25, 1974. Christina graduated from Whitehall-Coplay High School in Whitehall, Pennsylvania in 1992. She graduated from Bucknell University with honors and a Bachelor of Science Degree in Biology and minor in Education in 1996. Christina received her Master of Science in Education Degree in 2001 and earned her teaching certifications in Biology and General Science. While pursuing her Doctorate of Education Degree in Educational Leadership at Lehigh University, Christina garnered a Secondary Principal certification and Superintendent Letter of Eligibility.

In 1997, Christina began her educational career as an eighth grade physical science teacher at Whitehall-Coplay Middle School. After serving five years in this role, she accepted the Whitehall-Coplay Middle School assistant principal position and maintained this position for two years.

In 2003, Christina was hired as the Director of Curriculum and Assessment at Catasauqua Area School District in Catasauqua, Pennsylvania. She is presently employed in this position.

Throughout her professional career, Christina has coordinated curriculum review and development activities in all K-12 content areas. Additionally, she has facilitated a variety of professional development opportunities for teachers and administrators. Furthermore, she has implemented several formative and summative assessment initiatives in the districts in which she has worked. Christina is a member of the

Association for Supervision and Curriculum Development, the Pennsylvania Association for Supervision and Curriculum Development, and the National Staff Development Council.

Christina resides in Allentown, Pennsylvania with her husband, Drew, and daughter Cassandra.