A Quantitative Case Study Analysis of the 4Sight Benchmark Assessment

By

Daniel R. Castagna

B.S., Clarion University of Pennsylvania, 1998

M.S.Ed., Youngstown State University, 2001

Submitted to the Graduate Faculty of

The School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Education

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH
School of Education

This dissertation was presented

by

Daniel R. Castagna

It was defended on

October 22, 2008

and approved by

Dr. William Bickel, Professor, Administrative and Policy Studies

Dr. Otto Graf, Clinical Professor, Administrative and Policy Studies

Mr. Joseph Werlinich, Associate Professor, Administrative and Policy Studies

Dissertation Advisor: Dr. Sean Hughes, Associate Professor, Administrative and Policy Studies

A Quantitative Case Study Analysis of the 4Sight Benchmark Assessment

Daniel R. Castagna, Ed. D.
University of Pittsburgh, 2008

This study followed students in a suburban middle school (grades 6, 7 and 8) in Western Pennsylvania and confirmed predictions made by the 4Sight benchmark tests (given during the 2007-2008 school year) when compared to the actual student results on the spring 2008 PSSA exam.

The researcher evaluated the ability of the 4Sight tests to accurately predict student's scores on the PSSA exam. The *Success for All Foundation* claimed that this formative assessment can predict (using a linear regression model) how students would have scored on the PSSA test if it had been given on the same day. The researcher confirmed this allegation with a case study with an entire school population. The Western Middle School administered four rounds of 4Sight testing during the 2007-2008 school year with the last round being administered one week after the PSSA exam. A comparison was analyzed between the third and fourth round of tests administered in January and April and the PSSA test in April. In their 2007 Technical Manual, the *Success for All Foundation* used third quarter test results to analyze correlation to the PSSA test. The researcher duplicated this process to compare correlations and added the fourth test analysis for additional findings. A Kappa coefficient formula was also used to compare predicted and actual PSSA classification categories.

Since the 4Sight exam is new, little research has been done to track its effectiveness. Student's raw score results were charted (by grade level) from the four 4Sight exams. Along with their raw score is the predicted categorical classification created by the *Success for All Foundation* that coincides with categories reported after

PSSA testing. The last column of the test score data shows students actual raw score results and actual categorical classifications on the 2008 PSSA test (published near July of 2008). This format allowed the researcher to chart students' progress from the third and fourth 4Sight exams and analyze correlation among predicted and actual results.

ACKNOWLEDGEMENTS

The writing of this dissertation would not have been possible without the personal and practical support of numerous people. To begin I would like to thank the members of my dissertation committee, Dr. William Bickel, Dr. Otto Graf, and Mr. Joseph Werlinich. They all have generously given their time and expertise to better my work. To my advisor, Dr. Sean Hughes, I thank you for your commitment and advice that guided me through this process. To Dr. Elaine Rubinstein of the Office of Measurement, your unselfish dedication to helping students is an inspiration to all.

To my incomparable parents, Robert and Dana Castagna, without them my educational journey would not have been possible. Their unconditional love taught me to continue to move forward regardless of the obstacle. Their belief in public education inspired me to continue this path of learning. They showed me that education is the way to see truth, to realize dreams, and to truly make a difference.

To my incredible sister, Kristen Boni, whose unmatched support of and belief in me carried me through some of the toughest times.

To Nichole Baker, for her assistance in editing my work, I offer sincere appreciation for your time, your feedback, and your endless encouragement.

To my colleagues in the West Mifflin School District, who provided excellent role models in their active pursuit of advanced degrees, I appreciate the concern for my progress in this endeavor.

Last, this dissertation is dedicated to my son Dante. You are the reason. You are my motivation. Figure out where you want to go and how to get there. Then never, never, never give up. When your determination is fixed, nothing is impossible.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

Chapter 1

1.1 Introduction

"Information is the key to holding schools accountable for improved performance every year among every student group. Data is our best management tool. I often say that what gets measured, gets done. If we know the contours of the problem, and who is affected, we can put forward a solution. Teachers can adjust lesson plans. Administrators can evaluate curricula. Data can inform decision-making. Thanks to No Child Left Behind, we're no longer flying blind."

- Margaret Spellings, U. S. Secretary of Education, 2005

In today's era of high stakes testing and accountability, more school districts are forced to continually monitor and adjust instructional methods. Educational standards developed by state departments of education have become the guide in which school districts must follow in order to conform to the demands of high stakes testing. "Standards are like guiding stars that enable us to navigate a successful academic voyage" (Sadker & Zittleman, p 740). To ensure that schools are maintaining an effective standards-based system, districts need to constantly monitor classroom instruction and student progress.

With the passing of the No Child Left Behind (NCLB) in 2001, the relationship between schools and data-driven decision making greatly changed. Schools that are failing to make the requirements of NCLB or districts currently close to not meeting requirements are enacting measure to continuously collect data on student performance. With this focus, school leaders need tools that provide them information that is quick and reliable so they can better see where student's strengths and weaknesses lie. This level of information allows school leaders to place the information in the hands of those who need it the most; the teachers. This knowledge and insight provided by the data allows

teachers to make mid-course adjustments and continuous corrections to assist the academic success of their students (Technology Alliance, 2005).

From the years 2000-2007 the Rand Corporation conducted four studies that examined the use of data in four different educational settings. The studies are reviewed below:

Description of RAND Studies

Study	Funding Source	Purpose	Method
Implementing Standards-Based Accountability (ISBA)[1] 2002–2007	National Science Foundation	To examine the implementation and effects of standards-based accountability systems	• Statewide data collection in California, Georgia, Pennsylvania • Superintendent, principal, teacher surveys • Interviews with state officials • Case studies of 18 schools
Data-driven decision making in South-western Pennsylvania (SWPA)[2] 2004–2005	Heinz Endowments Grable Foundation	To investigate district practices in using data to inform instruc-tional, policy, and evaluation decisions	• Case studies of 6 districts and 1 charter school in SWPA • Superintendent survey • State/regional interviews
Instructional improvement efforts of districts partnered with the Institute for Learning (IFL)[3] 2002–2005	The William and Flora Hewlett Foundation	To examine districtwide efforts to improve teaching and learning as well as the contribution of the IFL, an intermediary organization, to reform efforts	• Case studies of 3 urban districts in the South and Northeast • Principal and teacher surveys • Interviews; focus groups • Observations of trainings • Review of documents
Evaluation of Edison Schools[4] 2000–2005	Edison Schools	To understand Edison's strategies for promoting student achieve-ment and examine how they were implemented; to assess the effect of Edison's management on student achievement	• Case studies of 23 schools • Interviews with Edison staff • Observations of trainings and meetings • Analysis of test scores for all Edison schools

[1] For further details see Stecher and Hamilton (2006); Hamilton and Berends (2006); and Marsh and Robyn (2006).
[2] For further details see Dembosky et al. (2005).
[3] For further details see Marsh et al. (2005).
[4] For further details see Gill et al. (2005).

Retrieved from: http://www.rand.org/pubs/occasional_papers/2006/RAND_OP170.pdf on May 2, 2008.

When summarizing their findings from the Rand Corporation study, authors Marsh, Pane and Hamilton (2007) reported that teachers and administrators in all four settings complained that state test results arrive too late to make meaningful adjustments to curriculum and instruction. Tests taken in the spring usually do not produce results until the middle of July. "By then the tested cohorts of students has moved on to different classes and may have moved to a different school" (p 5). For this reason, many districts moved to benchmark tests to frequently assess students and curricular programs.

"More than 80 percent of superintendents in California, Georgia, and Pennsylvania found results from local assessments to be more useful for decision making than state test results" (p 5). These frequent informal assessments were reported to provide rapid, regular feedback to administrators, teachers and students. While many schools have enacted teacher made in-house assessments and item banks, most enjoyed the convenience of commercially made benchmark tests. Olson (2005) reports that the market for formative assessments is now one of the fastest growing sectors of test publishing companies.

1.2 Statement of the Problem

Currently over 400 school districts in Pennsylvania are purchasing the 4Sight benchmark assessment as a tool to inform instructional practice, identify areas of weakness in the curriculum, and predict student's results on the PSSA test. Administrators and school leaders are using the test to analyze curriculum and measure the effectiveness of instructional interventions. While the test results are viewed as valuable insight, no case study has ever followed an entire school population and compared predicted results to actual results.

Although versions of the 4Sight benchmark assessment have been developed in over twenty states across the country, only in Pennslyania is this test an approved data collection tool necessary for a school district to obtain the accountability block grant. Last year, the school district being reviewed in this study received over one million dollars from this grant.

During the 2007-2008 school year, the Western School District used the 4Sight benchmark test for the first time for students in grades 3-11. The testing and associated scoring/scanning software costs the district over $10,000 annually. Along with the monetary commitment, the district has also dedicated four testing dates throughout the school year to administer the 2 hour exam. Some argue that this measure takes away too much instructional time, leads to frustration with students and staff, and is not cost effective. They complain that the school has giving in to over-testing. Others feel that the testing is a valuable tool that identifies student's individual strengths and weaknesses, exposes curriculum gaps, and gives accurate predictions of how students will score on the PSSA exams in early April.

1.3 Research Questions

1. Do the 4Sight exams accurately predict student's raw scores on the PSSA exam?
2. Is there a high correlation between the predicted categorical classifications of students on the 4Sight tests when compared to actual PSSA Results?
3. Is the information gained from the fourth 4Sight test worth the loss of instructional time?

1.4 Purpose of this Study

This study will follow students in a suburban middle school (grades 6, 7 and 8) in Western Pennsylvania and will confirm or dispute the predictions made by the 4Sight tests (given during the 2007-2008 school year) when compared to the actual student results on the spring 2008 PSSA exam.

The researcher will evaluate the ability of the 4Sight tests to accurately predict student's scores on the PSSA exam. The *Success for All Foundation* claims that this formative assessment can predict (using a linear regression model) how students would have scored on the PSSA test if it had been given on the same day. The Western Middle School will be administering four rounds of 4Sight testing during the 2007-2008 school year with the last round being administered one week after the PSSA exam. The close proximity between the last round of testing and the PSSA test will allow for a fair comparison between the 4Sight tests predictions and actual PSSA results. A comparison will also be analyzed between the third round of tests administered in January and the PSSA test in April. In their 2007 Technical Manual, the *Success for All Foundation* used third quarter test results to analyze correlation to the PSSA test. The researcher will duplicate their process to compare correlations.

Since the 4Sight exam is new, little research has been done to track its effectiveness. The only information available to support the *Success for All Foundations* claim that there predictive scores have a high correlation to PSSA results is the information published by the company. The researcher will conduct a case study that follows the middle school (grades 6, 7, 8) students at Western Middle School as they take the 4Sights exam four times this year. Student's raw score results will be charted (by grade level) from the four 4Sight exams. Along with there raw score is the predicted categorical classification created by the *Success for All Foundation* that coincides with categories reported after PSSA testing. The last column of the test score data will then show students actual raw score results and actual categorical classifications on the 2008 PSSA test (published near July of 2008). This format will allow the researcher to chart

students' progress from the third and fourth 4Sight exams and analyze correlation among predicted and actual results.

1.5 General Procedures

The researcher will:

1. Review the literature on 1) aptitude and achievement tests, 2) the big business of testing, 3) the history of assessments, 4) the background of No Child Left Behind, 5) summative, formative, and benchmark assessments, 6) 4 Sight tests and its' competitors, 7) review other states attempts at benchmark assessments, 8) and examine the values/challenges and drawbacks of using benchmark tests.

2. Oversee the administration of the 4Sight benchmark assessment in the Western Middle School during the 2007-2008 school year. Students in grades 6, 7, and 8 will take the exam in their regular school setting during an adjusted 2-hour delay bell schedule. The assessment will be administered four times throughout the school year.

3. Chart student's raw scores and predicted PSSA categories during each phase of testing.

4. Analyze data using linear regression between students' third 4Sight test results and actual PSSA results to check for correlation.

5. Analyze data using linear regression between students' fourth 4Sight test results and actual PSSA results to check for correlation.

6. Analyze data using kappa coefficient to compare predicted categorical classifications from the third 4Sight test to actual PSSA results.

7. Analyze data using kappa coefficient to compare predicted categorical classifications from the fourth 4Sight test to actual PSSA results.

8. Determine whether the correlation among all data analyzed is positive, negative, or no correlation at all. Findings will then be summarized and reported.

1.6 Operational Definitions

Adequate Yearly Progress (AYP): The minimal level of improvement school districts must achieve each year.

Anchors: A tool developed by the Pennsylvania Department of Education to better align curriculum, instruction and assessment practices throughout the state.

Assessment: An overarching term that covers both testing and assessment.

Achievement Test: Tests that show how well a student has learnt the material that has just been taught.

Aptitude Test: A test showing how well a student is likely to learn a particular skill.

Assessment Anchors: Cluster standards and highlight which areas will be tested on the PSSA exam.

Benchmark Assessments: Measures of the same set of knowledge or skills administered several times throughout the year.

Criterion-Referenced Test: Test results are compared to predetermined criteria and not to the performance of other test takers.

Curriculum: Approved teaching material for classrooms.

Formative Assessments: Assessments for learning designed to focus instruction and identify areas of strengths and weaknesses.

High-Stakes Testing: Assessments that have punitive consequences for school districts and students.

Instructional Strategies: Activities designed to help students learn that are implemented inside of the classroom.

Norm-referenced test: Test scores are compared to the scores of other test takers in a group.

Reporting Categories: Anchors are organized into reporting categories.

Standards: What all students should know and be able to do.

Summative Assessments: Assessments of learning which are normally associated with end of year exams that usually carry high stakes consequences.

1.7 Limitations of the Study

This study will be limited to only middle school students in one school district in Western Pennsylvania. Within that population, the researcher will only analyze test scores of students who took both the third and fourth 4Sight test as well as the 2008 PSSA exam.

The main weakness of the study will be the obvious bias the researcher will bring as the middle school assistant principal in charge of the testing. The students and staff know that he is the person who led an assembly with students and parents about the benefits of using this testing and will be monitoring results closely. Another limitation may lie in the fact that as the teachers better understand how to proctor the 4Sight test, they may become better at explaining to students how to take the assessment. This may then cause correlations of predictions to increase as students' progress through the four exams.

The strength of the study will be the unrestricted access the researcher will have to the students, staff, testing materials, data collection software, and program implementation. Other strengths include the researchers' time on site, his full engagement in the process, and the fact that the taking of the exams by students essentially is a practice run for the yearly PSSA test (students become comfortable in testing rooms, know what materials they need to have for testing, understand expectations, etc.). Since this is the first year of the testing in the chosen school district, the researcher will be able to follow the results throughout the school year to monitor the effect of this intervention.

All three grade levels will be followed so correlations can be compared by building and grade level. This sets up well for further longitudinal research to compare and contrast interventions between grades.

Chapter 2

Review of Literature

2.1 Introduction

The talk of test scores and high stakes accountability dominate much of today's discussion around the success or failure of local public schools. Real Estate agents are using test results to rate different communities and these ratings have been shown to directly raise or lower the prices of homes (Figlio & Lucas, 2000). The effects of test scores carry even higher stakes when viewed at the state and national levels. Millions of dollars depend on the test scores of children to evaluate educational programs and allocate grant money and funds. Educational policies are using high stakes tests to drive reform and elicit changes in schools. School assessment is the centerpiece of many educational improvement efforts. Since most policymakers believe that what gets tested is what gets taught, often these changes have the greatest impact on teachers and students. Sadly these changes have led to many negative consequences for urban and suburban school districts.

Some researchers would argue that high-stakes testing is more proof of a principle of social science known as "Campbell's Law": "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (Campbell, p. 27). High-stakes testing and the accountability attached to the results of such testing has increased the pressure on public school leaders for their districts to perform to the prescribed standards. Many school districts find themselves looking for early indicators to define areas of need so that plans to adjust and

alter instruction can be put in place in time to hopefully change the prescribed outcome. For school districts in Pennsylvania, the 4Sight benchmark assessment is being purchased to predict how students will score on high-stakes tests and outline specific strengths and weaknesses in curriculum plans.

This review of literature will examine a brief history of assessment, identify and explain various types of assessment, and summarize the reauthorization that redefined accountability in public schools. Information about the 4Sight benchmark assessment will be analyzed along with a comparison between Pennsylvania and other surrounding states attempts at periodic assessments. The competitors of 4Sight will also be evaluated and scrutinized.

Finally, the researcher will attempt to highlight the need for a study that follows students in a suburban middle school (grades 6, 7 and 8) in Pittsburgh, Pennsylvania that will confirm or dispute the predictions made by the 4Sight tests and compare them to the actual student results on the yearly (2008) PSSA exam.

2.2 Aptitude and Achievement Tests

Each year high school juniors and seniors aspiring to enter college anxiously take aptitude tests as part of the college application requirement. Aptitude tests are assessments given to measure a student's existing ability in a certain area and to gauge their potential to learn in that same area. These tests are forward looking in that they attempt to describe the ability a student has to become skilled in a specific topic (Reeves, 2002). Currently the Scholastic Aptitude Test (SAT) and the American College Testing

Assessment (ACT) dominate the college entrance exam marketplace. Every year more than a million high school students take these exams (Atkinson, 2001).

The SAT was first administered as a college entrance exam in 1926 and is considered to be America's oldest, most widely used, college entrance exam. It was primarily developed to allow students from all socioeconomic backgrounds a chance to show they have the ability to achieve a college degree. Prior to the use of the test, college entrance was limited to students whose parents attended college (Hebert, 2007). The SAT is broken into three sections that include a math assessment test, a critical reading assessment, and a recently added writing section (to help the SAT compete with the growing popularity of the ACT). Each of the three sections of the SAT is worth up to 800 points for a total combined possible score of 2400. Students who take the SAT are given a point for a correct answer, a point is taken away for a wrong answer, and answers left blank are not counted at all toward the final score. Total completion time for the SAT exam is usually over five hours and the registration cost is around $50.

The ACT originated in 1959 as a direct competitor and challenger to the SAT. Besides the fact that both tests are formalized tests intended as a precursor to college admission, the similarities end there. The ACT is structured and scored differently and costs less than the SAT. The ACT is divided into four multiple choice sections with an optional fifth section of writing. The four sections consist of English (mostly grammar and mechanics), mathematics (includes trigonometry which the SAT does not cover), reading (mostly arts and literature), science, and the optional writing assessment which is very similar to the SAT format. Each of the four sections is graded on a scale from 1 to 36. Unlike the SAT, the ACT does not take points away for incorrect answers. The ACT

exam takes less than three hours to complete and the registration cost is around $29. Hamilton (2005) cited that the main difference between the SAT and ACT is that the SAT is norm-referenced and the ACT is criterion-referenced. A score on a norm-referenced test like the SAT tells whether the test-taker did better or worse than other people who took the same test. By contrast, a criterion-referenced test like the ACT provides for translating the test score into a statement about the behavior to be expected of a person with that score or their relationship to a particular subject matter. Also, the number of students taking the ACT has increased steadily over recent years. "ACT enrollment now virtually equals that of the SAT" (Manya Group, 2008). In 2004, 2.1 million high-school students took the ACT while 1.4 million high school seniors took the SAT (Hamilton, 2005).

While the formats of these two tests differ and the competition over the market share persists, both tests continue to receive objections. The makers of achievement tests complain that aptitude tests show disconnect from what students are actually learning in their high school curriculum (Kirst & Rowen, 1993). Achievement tests are designed to assess student's knowledge or skills in a specific content. These tests allow educators to see what a student can do now, after completing a course of learning. Atkinson (2001) argues that, "achievement tests are fairer to students because they measure accomplishment rather than ill-defined notions of aptitude" (p. 5). He goes on to state that achievement tests are more appropriate for schools because they set clear goals that guide learning and can be used to improve performance (Atkinson, 2001).

While both aptitude and achievement tests can provide a picture of students current levels of performance, aptitude tests play a larger role in the college application

process (Popham, 2008). It is clear that the Educational Testing Service (makers of the SAT) and the California Test Bureau (CTB) division of the McGraw-Hill company (makers of the ACT) have been able to persuade colleges and universities to require their tests for admission. "While colleges have created this market, students and their families pay the freight to take the tests" (Hamilton, p. 2). Only estimates can predict the actual costs families spend each year on preparatory classes, pre-tests and coaching software to assist children in preparing for these exams. In short, the expense for these materials and overall revenues generated by these companies is staggering.

2.3 The Big Business of Testing

As educational testing continues to be a big industry, business leaders are forming coalitions that are adding a corporate influence into public schools. John Stevens, executive director of Texas Business and Education Coalition stated that, "Educators do not dominate the dialogue on education in Texas. The business community and a group of key legislative leaders have been the major players in shaping state education policy" (Gluckman, p.1). The business model of increased testing and accountability has resulted in a tremendous increase in spending for the creation, distribution, and scoring of standardized tests. Spending on testing in Texas rose from 19.5 million in 1995 to 68.6 million in 2001. In Massachusetts, the state department of education saw an increase of spending on testing from $750,000 a year in1990 to 23.2 million in 2002. (Gluckman, 2002). Trends indicate that by the end of 2008, states will be spending up to five billion dollars a year to implement state wide testing programs. Add in the indirect costs of preparatory books and school intervention programs and the costs could be 15 times

higher (Hamilton, 2005). While the spending attached to state mandated testing programs increases, many states are using commercially produced tests to assess students and schools.

As the marketplace for educational testing continues to rise, few companies control the bulk of testing sales. Currently, four companies control most of the testing market. Those companies include Educational Testing Service, CTB McGraw-Hill, NCS Pearson, and Riverside Publishing.

Educational Testing Service (ETS), founded in 1947, is the world's largest non-profit, private testing and measurement organization. In 2007, it operated on a 1.1 billion dollar budget to develop numerous standardized tests in the United States and worldwide. In that year, more than 50 million tests were administered internationally to over 180 countries (Wikipedia, 2008). ETS creates and scores tests in K-12 schools and higher education universities such as the SAT tests, the Graduate Record Examinations (GRE), the Preliminary SAT/National Merit Scholarship Qualifying Test (PSAT/NMSQT), the College Level Examination Program (CLEP), the Praxis Tests, and the California High School Exit Exam. ETS also has a global market for its Test of English as a Foreign Language (TOEFL) and its Test of English for International Communications (TOEIC). In 2007, 6.2 million people registered to take the TOEFL and the TOEIC worldwide (Educational Testing Service, 2008).

CTB McGraw-Hill offers custom achievement and aptitude tests for students in K-12 schools. They design, score and sell the Terra Nova tests, the California Achievement Tests, the Acuity Assessment program, and the Spanish Assessment of

Basic Education. McGraw-Hill acquired the California Testing Bureau (CTB) in 1965 had over 4.2 billion dollars in sales in 2000 (Frontline, 2002).

NCS Pearson is considered to be the leading scorer of standardized tests in the United States. National Computer Systems (NCS) was created in 1962 as a data processing firm. It grew to become the largest scorer of standardized tests in the 1980s and began selling some psychological and counseling tests. NCS Pearson won several state testing contacts in the 1990s and quickly became a leader in the school testing market. In that time, the revenues of NCS grew dramatically, "from $35 million in 1980 to nearly $630 million in 1999" (Gluckman, p. 1). In 2000, after the Texas State Department of Education signed a 233 million dollar contract with NCS to operate it statewide testing program, Pearson, a British media conglomerate, purchased NCS for $2.5 billion. While neither NCS nor Pearson started out in the testing business, half of NCS Pearson's overall revenue now comes from its assessment and testing services (Frontline, 2002).

Riverside Publishing (a Houghton Mifflin company) was bought by France's Vivendi Universal in August of 2001 for $2.2 billion. Like NCS Pearson, Vivendi Universal does not have a background in education. It instead owns a French publishing empire, Motown Records, and Universal Pictures. Riverside publishing creates and scores tests including the Iowa Test of Basic Skills, the Woodcock-Johnson III, the Gates-MacGinitie Reading Tests, and the Assess2Know program. Riverside was established as a subsidiary of Houghton Mifflin in 1979. Since then, Riverside Publishing's Iowa test has been administered to over 4 million students every year (Frontline, 2002).

With billion dollar revenues and increased testing sales, the question becomes from where does the money for all these testing services come? With testing mandates in place, public schools have been forced to focus much of their school budgets on test preparation and curriculum alignment. The more K-12 schools are forced to conform to mandatory testing requirements, the more money will be spent to supplement instruction. Trends show that each year, more of a school's budget will go directly to testing products and preparatory programs (Gluckman, 2002).

2.4 History of Assessments

Standardized testing in the United States is a billion-dollar industry that impacts the educational decisions of many public schools. Reviewing the history of assessments shows a 150-year-old discourse about its impact on learning and the debate surrounding the use of high-stakes testing.

Link (1919) while writing *Employment Psychology*, complained that standardized testing forced students and teachers to spotlight the needs necessary to prepare for good test scores rather than focus on learning the practical content of the subject matter. Hanson (1993) summarized the thoughts of Dr. Link by stating, "Students are then motivated to work for good examination grades rather than for the intrinsic rewards of learning" (p. 105). These thoughts reveal that the arguments for and against testing are not new to American society. With history showing signs of frustration with standardized testing one must question how today's current landscape of testing came about and why the same arguments are echoed nearly 90 years later.

So where exactly did this term assessment come from? In the late 1940s, the United States Office of Strategic Services coined the term assessment to imply a process much more in-depth and significant than what is thought of in normal testing (Gregory, 2004). Since this time the definition and application of the word assessment has grown. Herrnstein & Murray (1994) realized that the use of standardized assessment will allow researchers to compare any one person's success or failure logically and literally with another person. This thinking is later echoed when Salvia & Ysseldyke (2004) defined assessment as the "process of collecting data for the purpose of making decisions about individuals and groups" (p. 4).

The use of assessments to classify individuals into categories is now a reality for all public school students in the state of Pennsylvania. All students are classified in one of four performance categories based upon their performance on a once a year test of Reading and Mathematics. The categories and definition of each are listed here:

Pennsylvania's General Performance Level Descriptors
Advanced The Advanced Level reflects superior academic performance. Advanced work indicates an in-depth understanding and exemplary display of the skills included in the Pennsylvania Academic Content Standards.
Proficient The Proficient Level reflects satisfactory academic performance. Proficient work indicates a solid understanding and adequate display of the skills included in the Pennsylvania Academic Content Standards.
Basic The Basic Level reflects marginal academic performance. Basic work indicates a partial understanding and limited display of the skills included in the Pennsylvania Academic Content Standards. This work is approaching satisfactory performance, but has not been reached. There is a need for additional instructional opportunities and/or increased student academic commitment to achieve the Proficient Level.
Below Basic The Below Basic Level reflects inadequate academic performance. Below Basic work indicates little understanding and minimal display of the skills included in the Pennsylvania Academic Content Standards. There is a major need for additional instructional opportunities and/or increased student academic commitment to achieve the Proficient Level.

Retrieved from: http://www.pde.state.pa.us/a_and_t/site/default.asp

In compliance with §4.51(b)(4) of the PA School Code the State Board of Education approved "specific criteria for advanced, proficient, basic and below basic levels of performance" (PA Department of Education, 2008). These categories also carry consequences for students who fall into the Basic or Below Basic performance range. The Pennsylvania State Department of Education uses Adequate Yearly Progress (AYP) to measure the performance of public schools at the end of each school year. AYP is determined based upon four standards:

1. Students test scores – Are enough passing the PSSA test?
2. Participation rates – Are enough students taking the PSSA test?
3. Attendance rates – Are enough students attending school on a daily basis?
4. Graduation rates – Are enough students graduating from high school?

(Education Law Center, 2002)

Each of the four categories carries a specific goal for each year. This table defines each goal as outlined in NCLB: (retrieved from http://www.elc-pa.org/pubs/downloads/english/NCLBpubs/ncb-is-your-childs-school-making-the-ayp.pdf May 17, 2008)

Table 2.1 Pennsylvania AYP Standards

State AYP Standards that Schools Must Meet for Each School Year:

School Year	02-03	03-04	04-05	05-06	06-07	07-08	08-09	09-10	10-11	11-12	12-13	13-14
Passing Rate on Reading PSSA	45%	45%	54%	54%	54%	63%	63%	63%	72%	81%	91%	100%
Passing Rate on Math PSSA	35%	35%	45%	45%	45%	56%	56%	56%	67%	78%	89%	100%
Participation Rate - PSSA	95%	95%	95%	95%	95%	95%	95%	95%	95%	95%	95%	95%
School Attendance	95%	90%	90%	90%	90%	90%	90%	90%	90%	90%	90%	90%
Graduation Rate	95%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%

If a public school in Pennsylvania fails to make AYP according to the standards listed above it must then prescribe to corrective stages of reform. The stages are as follows:

1. Warning – first year of not making AYP
2. School Improvement I – second year of not making AYP. Schools must allow students to transfer and offer training for teachers to increase student achievement.

3. School Improvement II – third year of not making AYP. Schools must formalize a plan for improvement, allow student transfers, provide additional professional development for teachers, and offer additional after school tutoring.
4. Corrective Action I – fourth year of not making AYP. School must follow all sanctions of School Improvement I but also show changes in parent involvement, curriculum, school structure, administration and teaching.
5. Corrective Action II – fifth year of not making AYP. School must follow all action of Corrective Action I but also implement a plan for complete restructuring which could include being taken over by a private company or becoming a charter school.
6. Restructuring – sixth year of not making AYP. School must follow all actions provided by Corrective Action I and now implement plans for restructuring.

(Education Law Center, 2002)

Many teachers are frustrated with the way the Pennsylvania Department of Education has chosen to label failing schools and categorize students. Classifying students and branding school districts this way contradicts the Socratic Method; which many teachers would argue is the best type of teaching. Socrates, while teaching in ancient Greece, tested his students through dialogue and conversation. His goal was to always find the path to higher knowledge and not necessarily the right or wrong answer. Socrates' hope was to add to the scholarly dialogue.

Early universities of higher education always used oral examinations as their main form of testing student comprehension of content. The earliest record of oral examinations at the university level in North America was Harvard University's

requirement in 1646 for degree recipients to be able to read in Latin the Old and New Testament and "resolve them logically" (Hanson, p. 101). Oral exams took time and scores were debatable. Educators soon looked toward new ways of assessing a student's abilities. In the early parts of the nineteenth century, Yale began using standard biennial exams with all students at the end of the second and fourth years of college. Oxford and Cambridge soon jumped on board with the new philosophy of biennial testing and now instituted a formal written exam. Harvard, in 1833, then followed with a standard, written mathematics exam. Harvard implemented a written entrance exam in 1851 and Yale began using an annual written examination in 1865. The frequency of written examinations in American Universities continued to grow.

The colleges realized that common assessments were inappropriate for students following different courses of study so they began using separate exams for each course (Hanson, 1993). The ideal of the Socratic method had undergone an enormous transformation. The early written exams used by universities continued to spiral to our modern educational model shaped by testing, charting, comparing and even categorizing many of our youth. In today's fast paced, high-priced testing environment there is no room for dialogue that leads to the process of understanding.

After the use of written exams became common-place at the university level, their use disseminated into American public schools. "Boston's elementary and secondary schools had long followed the practice of annual oral examinations conducted by committees of visiting examiners" (Hanson, p. 102). As the size of Boston public schools grew in the mid nineteenth century, it became impossible to administer these oral exams on a large scale. "Between 1820 and 1860 American cities grew at a faster rate

than in any other period in U.S. history, as the number of cities with a population of over 5,000 increased from 23 to 145" (U.S. Congress, p. 105). The Boston Public School needed a more efficient way to test its 7,000 students. In 1845, Horace Mann, while working with the Boston Public schools, creating a standardized (written) essay test. Mann realized that this format would allow him to pose an identical set of questions to a growing student population while producing comparable scores with unmatched efficiency. Mann also used these test results as "political leverage over recalcitrant headmasters" (Madaus, p. 17) who claimed that Mann was irrational. The debate around testing during this era surrounded the use of oral versus written examinations. Mann asserted that written exams had numerous advantages over oral exams. The same questions being given to all students in all schools allows the students and the schools to be evaluated and compared. Mann also argued that the written exams did not allow teachers to show favoritism or probe students to elicit responses with deeper ideas. This format also allowed students to develop ideas without the interruption of classmates or the time constraints of oral reporting as a group (Hanson (1993).

It is at this time that the classification and grouping of students in public schools was brought to the state level. Mann later became the Secretary of the State Board of Education in Massachusetts. Under his tenure, Mann used the written exam format of the Boston Public Schools to develop a standard exam used by all schools in the State of Massachusetts. From its inception, this large scale use of testing was designed to monitor school systems by external supervisors and to classify students (U.S. Congress, 1992).

Sorting students for comparison and labeling purposes became even more accepted in the early 1900s when Binet introduced the IQ tests (Association for Early

Childhood International, 1991). Alfred Binet's intelligence tests were not designed to be used in American schools though later they would become a main-stay in special education testing. Binet did not believe that intelligence was a measurable trait like weight or height, though his American counterparts adapted his testing for use in American schools (U.S. Congress, 1992).

With Horace Mann having now instituted a standardized test across the entire State of Massachusetts and Alfred Binet's IQ exams being used to sort students by ability, testing was slowly becoming a main stay in public schools. The Thorndike Handwriting Scale, produced in 1909 is considered to be the first widely used achievement test in American public schools to measure students' mastery of correct handwriting form. Edward Thorndike, while working at Columbia University, also developed standardized tests in Math, Reading, Language, Spelling, and Drawing. Following this trend, there were over 100 standardized tests to measure achievement of students in elementary and secondary schools by 1918 (U.S. Congress, 1992).

By the 1930s, the array of standardized tests offered grew but was in no way close to the amount of testing students now face. "Few people who completed high school before 1950, for example, took more than three standardized tests in their entire school careers" (Perrone, p. 2). During this era of American testing, results were not published in newspapers, parents rarely saw the results, and teachers rarely discussed score results with students (Perrone, 1991). After 1950, testing programs began an upward spiral.

American schools in the 1950s through the 1970s saw a dramatic change in demographic population and enrollment numbers due to a huge influx of immigrants. "Enrollment in public schools jumped from 25 million in 1949-1950 to 46 million in

1969-1970" (U.S. Congress, p. 128). With numbers growing and populations changing, the standardized testing movement had to now re-invent itself to be able to handle the growing demand.

Technology's role in public education testing began in the 1950s when the automatic scoring machine was invented by the Iowa Testing Program. This equipment allowed school districts to score tests on a scale and with a level of efficiency that was previously unthinkable (U.S. Congress, 1992). A tremendous jump in the number of tests given in public schoosl soon followed. This new technology also led educational leaders to begin the discussions of testing on the national level. American educational leaders had a renewed interest in the use of large-scale assessments. Technology advancements now allowed reports to be given to schools that showed large scale results for all students. The computer generated information provided school leaders with quick access to data that analyzed students, teacher efficacy, and school curriculum. This later evolved into the computer based testing now commonplace in public schools.

After the launch of Sputnik on October 4, 1957 by The Soviet Union, emphasis on the output of American schools gained national attention and focus. American egos were shocked at how the Soviet Union won the race for space. Looking for answers, Amercans blamed schools for not educating students to their full potential. Federal and state politicians began to advocate large scale use of tests to assess student learning (Kreitzer, Madaus, & Haney, 1989).

This increased testing of student's aptitudes led congress to pass the National Defense Act of 1958. The idea behind this legislation was to provide federal dollars to school districts so they could upgrade mathematics and science education (U.S. Congress,

1992). Along with the upgrade in math and science, the federal money was also used to develop a program of tests that would identify students with high levels of aptitude and provide information for universities about the students seeking admission (National Defense Education Act, Public Law 85-864). With Math and Science education the focus, Social Studies and education of the Arts soon took a back seat. This reform in education began the narrowing of curriculum in order to meet demands of testing.

Rating schools for comparison did not happen much until the mid-1970s when the College Board announced that SAT scores had been falling steady for several years (College Board, 1977). Following the publication in 1983 of *A Nation at Risk* by National Commission on Education, higher expectations for performance, graduation requirements for all students and more testing became an increasing popular trend. This report called for an end to minimal competencies and argues that public standards for student performance are too low. American students were shown to be achieving at much lower levels than students in other countries. This report triggered nationwide panic about the low performance of students in our country (Berliner & Biddle, 1995). As a result, policymakers in many states created educational standards and implemented testing policies as periodic checks of standards. High-stakes consequences were now attached to these assessments to hold administrators, teachers, and students accountable (Quality Counts, 2001).

Politicians' role in education increased more when in 1988, congress created the 26-member National Assessment Governing Board to set policy for the National Assessment of Educational Progress, commonly known as the "The Nation's Report Card." The Nation's Report Card, "is the only nationally representative and continuing

assessment of what America's students know and can do in various academic subjects" (Retrieved from http://www.nagb.org/). It is the only measure of student achievement in the United States where you can compare the performance of students in one state with the performance of students across the nation or in other states.

The role of testing increased in our schools once more in 1994 when congress reauthorized the Elementary and Secondary Education Act (ESEA) of 1965 to require all public schools to use standardized testing. In 1965, ESEA provided Title I funds to schools as a way to assign money to schools with higher numbers of students who qualify for free and reduced lunch. Its design was intended to increase the opportunities for enrichment and smaller class sizes to poorer school districts, but it also encouraged school districts to use standardized testing to assess all students (Public Law 107-110, Section 1001). The 1994 reauthorization of ESEA further solidified the federal government's role in public education because it clearly defines academic standards and high stakes testing as a necessary means for school districts to receive Title I Funds. For the first time, reports had to be disaggregated along racial and socioeconomic lines so it could be determined if poorer students where benefiting from the use of Title I money. While the terminology of high stakes testing and accountability now had greater emphasis, rarely were sanctions imposed and public display of failure was limited (Kafer, 2004).

The No Child Left Behind (NCLB) Act of 2001, again revised the Elementary and Secondary Education Act. This new legislation created original incentives, stringent requirements, and posed significant challenges for states. For the first time in our nation's history "NCLB moves the federal government from being primarily a source of

funding--now about 9% of every public school dollar--to being a major factor in shaping the substance of K-12 instruction" (Bloomfield and Cooper, p. 1). The federal government is now holding schools, districts and states accountable for student outcomes. This legislation marked the most dramatic increase of the federal government's function in public education in almost 40 years. Of all the revisions of the ESEA, NCLB of 2001 carries with it specific sanctions for schools who fail to meet its standards of success. For the first time, school districts were classified in categories such as warning, school improvement, and corrective action.

2.5 Background of No Child Left Behind

With specific sanctions and public comparisons in place, President George W. Bush in 2002 signed into law the No Child Left Behind Act. NCLB Act of 2001, the reauthorization of the original 1965 Elementary and Secondary Education Act completely changed the language, focus, and objective of most public schools. Some researchers view this legislation as having a positive impact on classroom instruction. Raymond & Hanushek (2003) conducted a study that showed that students in states with high stakes accountability systems in place score higher in mathematics on the National Assessment of Educational Progress than students who test in states which do not have serious consequences attached to testing. Winter (2003) agrees that states with accountability measures in place score higher on national tests. Advocates of high-stakes testing believe that testing helps to raise the expectations for teacher and student performance in the classroom. Another argument for the accountability movement is that it adds credibility to being a high school graduate. "There ought to be a way to ensure to

graduates, future employers, institutions of higher education, and the public that a high school diploma means that students have the skills they need to succeed" (Edweek, 2004).

Opponents of the high stakes consequences attached to NCLB believe the reauthorization was politically designed so people would view its intent as the betterment of education. "How could any reasonable person oppose the idea that schools should show some measure of success…and produce evidence that children are learning?" (Sirotnik, p. 67) Naming it "No Child Left Behind" evidences the political spin and hidden agenda behind its creation. It can be implied from its title that schools have been forgetting about students for years and now, with this legislation, that will stop. What the reauthorization did was completely change the relationship between the federal government and local school agencies. Public schools in Pennsylvania now have to define their success in terms of each student's level of performance on the once a year Pennsylvania System of School Assessment (PSSA). Sirotnik (2004) states that, "failing schools have been compelled to enact measures that have actually undermined the education and social well-being of students" (p. 72). Poor test scores on an annual exam result in reduced funding and public humiliation in local papers. "Standards based reform has drawn more attention to the achievement gap. Such patterns have been evident for decades, but because NCLB requires that test scores be disaggregated by race and released to the public, the issue has garnered considerably more attention" (Sirotnik, p. 69). Scwartz (2003) feels that testing is now used from the time we are born till the time we die to separate, identify and add to the social gap between classes of people. In his paper tilted, *Standardized Testing in a Non-Standardized World*, he quotes Gerda Steele of the National Association for the Advancement of Colored People as saying that

standardized tests, "Are used in ways to keep certain segments of the population from realizing their aspirations. Most of all they limit the access of blacks and other minorities to higher education" (p. 2).

While the debate between advocates and adversaries of NCLB will always continue, one fact is not under debate. NCLB completely changed the federal government's role in public education. Basically, the federal government is now holding schools, districts and states accountable for student outcomes. Another step towards accountability, annual state-wide assessments were aligned with the curriculum to provide an external, independent measure to determine the extent of learning (U.S. Department of Education, 2004). "Among other things, NCLB placed a federal mandate on states to test every child in grades 3-8 every year in Math and Reading (Smith, p.7). All public schools, regardless of their demographics or socio-economic status, would be tested and held to the same expectations and prescriptive sanctions if they failed to meet the criteria established by NCLB.

NCLB defines an accountable educational system as having these important steps:

- States created their own standards for what a child should know and learn for all grades.
- With standards in place, states tested every student's progress toward those standards by using tests aligned with the standards.
- Each state, school district and school was expected to make adequate yearly progress toward meeting state standards. Progress was measured by sorting test results for students who were economically disadvantaged, were from racial minority groups, had disabilities or limited English proficiency.
- School and district performance was publicly reported in district and state report cards.
- If the district or school failed to make adequate yearly progress toward the standards, additional sanctions were imposed.

Note: From Introduction: No Child Left Behind. Pennsylvania Department of Education / located at http://www.pde.state.pa.us/nclb/cwp/view.asp?a

In April of 2005, the United States Department of Education Secretary, Margaret Spellings announced a set of "common-sense" approaches to guide states as they measure their progress toward the goals outlined in NCLB. Her plan included assessing all students in grades 3-8 and once in high school every year, breaking down students' results by subgroups and improving teacher quality. Spellings (2005) stated that the goals of school districts, "Must lead to all students achieving at grade level or better in reading and mathematics by 2014" (p. 1). To achieve this goal, every student must score proficient on the PSSA exam in the spring of 2014.

The reality for public schools is that we all must understand the implications that result from not meeting state wide performance standards in accordance to NCLB. All districts need to evaluate every classroom and every student to understand what they can do to succeed. The struggle for many administrators is to find the balance so that students are placed in a creative, positive learning environment while gaining exposure to necessary content that allows them to prepare for the PSSA exams.

2.6 Summative and Formative Assessments

The PSSA exams are just one form of summative assessment used in schools across Pennsylvania. With NCLB's emphasis on high stakes, standardized testing, these summative assessments are used to measure what students are learning. The results are gathered and used for comparison outside of the school setting. Summative assessments are often referred to as "assessments OF learning" because they accurately measure what a student has learned or what they have or have not mastered. These types of assessments let students know how they did and are usually carried out at the end of a course of study.

Essentially, these assessments are used to judge a student's competency or to evaluate a program at the end of an academic year. They are holistic, final, content-driven and rigorous. Often, they are used to assign students with a formal grade or ranking (Starkman, 2006). Summative assessments are viewed to be a main-stay in today's era of accountability. It is important to "think of summative assessments as a means to gauge, at a particular point in time, student learning relative to content standards" (Garrison & Ehringhaus, p. 1). Besides state assessment tests, other examples of summative assessments include midterm/final exams, end of unit tests, SAT tests and accreditation tests. The problem with summative assessments is that they happen at the end of the learning path so it is normally too late to adjust instructional methods or techniques (Aronson, p. 1). Generally, the primary purpose of summative assessments is to appraise a student's overall performance (McAlpine, 2002).

Formative assessments are usually conducted at the beginning and during the administration of a certain program, so progress can be charted and adjustments can be made. Formative assessments are often called "assessments FOR learning" because they identify student needs and provide input to highlight individual strengths and weaknesses. These types of assessments let students know how they are doing. In a classroom setting, formative assessment can be as simple as a teacher or peer providing feedback to a student about his/her work. These assessments are not necessarily used for grading or classifying purposes. Instead, these assessments are suggestive, goal-directed, and non-judgmental. The first priority of this type of assessment is to promote learning and student feedback. Examples of formative assessments include quizzes, journals, benchmark test, portfolios, KWL charts and pretests. The results of these checks are used

to identify areas that need additional attention. This information is then used to guide instructional methods and inform practice. "When anyone is trying to learn, feedback about the effort has three elements: recognition of the desired goal, evidence about present position, and some understanding of a way to close the gap between the two" (Black and William, p. 143). Using formative assessment provides feedback to both teachers and students. This feedback supports the learning process, is developmental in nature and helps to make decisions on the readiness of learners to perform on a summative assessment.

Plato Learning (2005) outlined uses of formative assessment at the district, school and classroom levels. Districts use formative assessments to track trends and gaps in the curriculum, compare grade level achievement between buildings and monitor student progress towards state assessment goals. Schools use formative assessments to track student progress, create profiles of learners and to identify the needs and challenges of specific classes and grade levels. Classroom level formative assessments are used to assess prior knowledge, create baselines to chart growth and to evaluate student learning during a set of learning activities.

2.7 Benchmark Assessments

> "Benchmarks allow you to take a pulse and not an autopsy"
> – Robert Slavin, *Chairman of the Success for All Foundation*

Another type of formative assessment, benchmark assessment, has grown in popularity since the rise of high stakes testing. Benchmark assessment data is sometimes referred to as a snapshot of student knowledge. The snapshot metaphor is useful for considering the values of and limitations related to different tiers and different types of assessment data.

Consider the following:

Snapshots (like this one to the left) reflect how this flower looks at a specific point in time. This can be used as a metaphor to compare the results shown on once-a-year summative assessments like the PSSA exams.

Benchmark assessments provide insight into learning trends and make it easy to chart student growth. This series of snapshots, taken over time, may reveal changes that could not be monitored once a year.

It is only with a series of assessment data, sampled over time, that changes in student knowledge, or learning, may be ascertained (Tananis, 2007).

Administrators and teachers use Benchmark assessments to:

- Establish the entrance-level performances of students when the school year begins.
- Indicate progress with strengths and weaknesses in a particular content area in the middle of the year.
- Create groupings of students based on their changing skill needs.

- Identify which students need enrichment or which students need special assistance at any point during the school year.

If teachers and principals know which students exhibit gaps in specific areas, they can direct efforts to fill in those gaps early enough to make a difference on the next state assessments. Benchmark assessments currently being implemented vary widely in form and structure. Some are given as often as monthly, while others are given just once or twice a year. Some are given online, others on paper. Some are designed to assess progress in a designated curriculum, and test the skills expected to have been mastered at a given point in time. Others are survey assessments of the body of knowledge expected (according to state standards) to be learned in a given school year. "Benchmark assessments occur within, between, and among instructional units" (Wisconsin Department of Public Instruction, 2007). These assessments can be teacher made documents such as common 9-week assessments, midterms, or finals, or commercially purchased materials. Some examples of commercially produced benchmark assessments include the 4Sight test, Scranton's Performance Series, and Pearson's benchmark assessment.

Even with the rapid implementation of benchmark assessments, there is little research on their effects. There is some verification that regular formative evaluation can help teachers and principals guide their instructional decisions (Schmoker, 1999; Trumbull & Farr, 2000), and that frequent assessment is better than infrequent assessment (Dempster, 1991; McMillan, 2004). However, there remains much to learn about how the use of benchmark assessments truly affects teaching and learning.

Throughout Pennsylvania more and more school districts are purchasing the 4Sight benchmark assessments to give them feedback on how students are progressing towards the proficiency levels of the PSSA exam. In the January/February 2008 newsletter published by the *Success for All Foundation* tilted, "Pennsylvania and the 4Sight" it is reported that 391 school districts across the commonwealth and 1,785 schools are now using the 4Sight test in Pennsylvania. The newsletter states, "Pennsylvania schools nearly doubled the number of 4Sight tests in the system. Clearly, the number of reporting tools being used demonstrates that Pennsylvania understands the power of the program: using data to drive instruction" (p. 1). Administrators and teachers in these districts rely heavily on the predictability of these assessments to highlight specific areas of weakness in the curriculum and in student learning. These results guide professional development planning and determine the spending practices of many local districts. The *Success for all Foundation* feels that the thought behind this is straightforward.

When Dr. Nancy Madden of the *Success for all Foundation* was asked (through personal communication) if there has ever been a case study done in a school district in Pennsylvania to compare the predictions made by the 4Sight to the actual PSSA results, she responded, "Yes." She continued to say, "Results for the state last year were compared to 4Sight scores collected in the spring of the year. Basically, the percent of students achieving proficiency on PSSA was very close to the percent predicted to be proficient by 4Sight. I will have to request the details from a colleague. Also, correlations between 4Sight and PSSA are reported in the 4Sight Technical manual."

The July 2007 newsletter published by the *Success for All Foundation* titled, "Pennsylvania and the 4Sight", stated that a check was done by the members of *Success for All* to see how well the benchmarks performed on the 2007 PSSA test. They provided this graph:

Comparison of 4Sight Third Quarter Scores to PSSA Scores			
Grade/Subject	4Sight Percent Passing		PSSA Percent Passing
3 Math	73		78
4 Math	80		78
5 Math	69		70
6 Math	73		69
7 Math	65		66
8 Math	68		67
3 Rdg	68		72
4 Rdg	72		70
5 Rdg	62		60
6 Rdg	74		63
7 Rdg	69		66
8 Rdg	70		74

They compared, "The percentage of students estimated to score proficient and/or advanced by 4Sight scores entered into the member center during the 3rd grading period" to the initial database of students who scored proficient or advanced on the 2007 PSSA. While the results are similar, these findings are non-conclusive. Since not all districts use the 4Sight test but all take the PSSA this comparison shows deficiencies.

2.8 4Sight Testing

> ***In today's high-stakes testing environment, you can't afford to wait to see how your students perform on your state assessments. You need to estimate how students are likely to perform throughout the year. That's why the Success for All Foundation created 4Sight, a benchmark assessment tool that enables you to predict your students' reading – and in some states, math – achievement multiple times throughout the year. These predictions allow you time to take action in the areas in which students need help.***
>
> ***- Success for All Foundation***
>
> **Retrieved from: http://successforall.com/ayp/4sight.htm on February 29, 2008.**

With state sanctions in place for schools who fail to meet NCLB mandates, many public schools in Pennsylvania have looked to the 4Sight benchmark assessments to prepare students for the PSSA test. The 4Sight benchmark assessment is a series of tests that are connected to Pennsylvania Department of Education's *Assessing to Learn: PA Benchmark Initiative*. All school districts in Pennsylvania have access to this formative assessment by simply completing an application through their Intermediate Units. Participation is voluntary but the 4Sight assessment can be listed as an approved data collection tool for numerous state grants and programs. 4Sight benchmarks for Reading and Mathematics were created to provide school districts with a blue-print of the upcoming state exams as well as a tool to track individual student's progress toward proficiency on the same exam. The *Success for All Foundation*, in collaboration with the *Pennsylvania Center for Data-Driven Reform in Education*, created this formative assessment to predict (using a linear regression model) how students would have scored on the PSSA test if it had been given on the same day (Success for All, 2007). It is

important to note that the *Center for Data-Driven Reform in Education* is housed at the John Hopkins University College of Education, and was founded by Dr. Robert Slavin and Dr. Nancy Madden. The *Success for All Foundation* is an educational reform non-profit organization also founded by Dr. Robert Slavin and Dr. Nancy Madden (2008 January/February *Success for All* newsletter p. 3).

Students taking the 4Sight are assessed using multiple choice and open-ended response questions that are administered within a 60-minute time frame. The *Success for All Foundation* recommends using the 5 exams (one benchmark and four evaluations) periodically throughout the school year. Questions and scoring for each exam are developed from the Pennsylvania Academic Standards, Assessment Anchors and Eligible Content for both Reading and Math. Chapter 4 regulations state that all Pennsylvania Academic Standards should be taught by all teachers in the commonwealth. Anchors were created by the Pennsylvania State Department of Education to cluster standards and highlight which standards are tested on the annual PSSA exam. Reading and Mathematics anchors are then organized into reporting categories. On the 4Sight exam, students are scored in the following "Reporting Categories":

Reading
- Comprehension and Reading Skills
- Interpretation and Analysis of Fictional and Nonfictional Text
- Learning to Read Independently
- Reading Critically in all Content Areas
- Reading, Analysis, and Interpreting Literature

Math
- Number and Operations
- Measurement
- Geometry
- Algebraic Concepts
- Data Analysis and Probability

(Anderson, 2007)

Reports and item analysis that show where there are curriculum gaps and trends throughout grade levels are available to schools. In the spirit of NCLB, disaggregated subgroup data is also available for each reporting category. This tool identifies curriculum gaps and problems with scope-and-sequence in Reading and Math. Another benefit of this assessment is that it is realigned (updated) to the standards every year and reports are generated automatically. Users of the reports state that the data is easy to access and useful in informing instruction.

The 4Sight benchmark assessments were designed to mirror the look, design and feel of the PSSA test. While it is given on grade level 3-5 times per year, it provides information of student progress towards the PSSA test and considered to be a low stakes formative assessment. It is not a teacher evaluation tool, a diagnostic assessment nor is it administered on a student's instructional level.

2.9 4Sight Competitors

After calling several schools in the Pittsburgh area the researcher realized that some schools are using a product from the Scantron Corporation in place of the 4Sight Tests. Dr. Christine Assetta, Assistant to the Superintendent, West Allegheny School District informed the researcher (in an interview conducted on October 13, 2007) that they decided to use Scantron's Performance Series because it is administered once throughout the year. She felt that the 4Sight testing took too much time away from instruction and preferred an online test.

This Performance Series that is created by the Scantron Corporation seems to be 4Sights closest competitor in Pennsylvania. On Thursday, November 1, 2007 Mr. Michael Marchionda, the Director of K-12 Testing and Assessment for the Scantron Corporation, was reached through personal communications. Mr. Marchionda was quick to point out the time it takes to administer the 4Sight tests throughout the school year and the quick results gained from online testing. He also pointed out that while the 4Sight tests claim to predict how students would have scored if they would have taken the PSSA test on the same day (using a linear regression model) the Performance Series tests use a predictability analysis to show how students will perform on the PSSA at the end of March. Basically, he was promising that by giving students a one time test in November, his company can predict what student's PSSA results would be five months later. When the researcher asked him about showing student growth and monitoring gains throughout the year he was informed that for an additional cost, a second test can be administered at the end of February. On, November 21, 2007 Mr. Marchionda sent a cost analysis of his program and even offered a free pilot of his software to the Western Middle School in the spring of 2008. The cost analysis is shown here:

SCANTRON
A HARLAND COMPANY

PRICE PROPOSAL

Important Requirement For Processing Purchase Orders:

Please submit your PO to match the product description, product code, unit price and total exactly as itemized below. If you have any questions, please contact:

Mike Marchionda, Account Executive

Date: 11/21/2007
Contact: Dan Castagna, Assistant Principal
Name of Institution: WESTERN SCHOOL DISTRICT
Address: 515 CAMP HOLLOW RD * *WEST MIFFLIN, PA 15122-2697*
Phone: 412-466-1473 Fax: Email: castagnad@wmasd.org

Qty	Product Description	Product Code	List Unit Price	List Total
1363	**Performance Series-Annual Subscription**	PSSubject2	**$8.00**	**$10,904.00**
1	**Performance Series On-Site Training (6 Hours) Limit 15 participants, maximum - each participant must have access to a PC during training.** **30 Day advance notice required to schedule training.** *Performance Series*	PS-TRANON	**$2,500.00**	**$2,500.00**
	SUBTOTAL			**$13,404.00**
	Tax			
	Ground Shipping Costs			
	TOTAL			

Terms and Conditions

- Terms are pre-paid, credit card or Net 30 with a valid purchase order.
- **Please fax Purchase Order to Scantron Corporation at 619-615-0516.**
- Pricing does not include sales tax or shipping.
- Pricing is valid for 30 days unless extended in writing by Scantron.
- Scantron reserves the right to change prices at any time without prior notice.
- All Training and Services must be scheduled within **nine months** of invoice date. Any unused time will be forfeited. Scantron shall notify customer of unused time prior to expiration date.

After reviewing the Scantron information with the Assistant Superintendent of the Western School District, he informed the researcher that the Western School District is

currently using the 4Sight assessment four times per year with all students in grades 3-11 and the total cost is under $9,700. The Assistant Superintendent also informed the researcher that the 4Sight assessment is an approved data collection tool that needs to be present in our district in order to qualify for certain grant money from the state department of education. He stated that the district looked into using the Scantron series assessment until they realized that it was not an approved data collection tool recognized by the Pennsylvania State Department of Education.

The Accountability Block Grant, which the Western School District received over one million dollars from in the 2007-2008 school year is one grant where the use of the 4Sight assessment is needed to show approved procedures of data collection. The Accountability Block Grant was designed to provide schools in Pennsylvania with monetary aid to implement educational initiatives that improve student achievement. The Block Grant supports "in-depth implementation of improvement strategies and allows districts to select from a breadth of programs to meet the specific learning needs of their students" (http://www.pde.state.pa.us/svcs_students/cwp/view.asp?a=175&q=111226 retrieved April 11, 2008).

The Educational Assistance Program (EAP), developed by the Pennsylvania Department of Education and designed in 2003-2004 to accelerate learning in struggling school districts, is another program where the 4Sight assessment is approved to be used. The EAP is a tutoring program targeted to the most challenging and needy districts in Pennsylvania. The program provides funds for tutoring of students before, during or after school hours. The 82 schools districts that were eligible for the program during the 2005-2006 school year had to meet the following criteria:

- Enrolled full-time in grades K-12 and
- Scored below proficient on a PSSA test in a subject required by the NCLB Act of 2001 or
- Enrolled in kindergarten through third grade and scored below the score approved by the Department on an approved eligibility test

Retrieved from: http://www.pde.state.pa.us/svcs_students/cwp/view.asp?a=296&q=114418 on February 2, 2008

As part of the design of the program, approved assessments must be used to show a baseline score of students participating in the program and an exit score. The 4Sight benchmark assessment is one of the approved assessments for the EAP tutoring program. So it is fair to say that these 86 school districts in Pennsylvanian receive funding for tutoring programs and in return purchase the 4Sight tests.

While conducting an online search of benchmark assessments the researcher located a test created by NCS Pearson (subsidiary of Pearson Education), a leading scorer of standardized testing (retrieved at: http://www.ncspearson.com/ on November 1, 2007), that claimed to be a formative benchmark tool. After speaking to three representatives using a toll-free number provided, a return phone call was received on Tuesday, November 27, 2007. Mrs. Barbara Peterson, Regional Director of the Pearson Company informed the researcher that her company could offer the same type of data tools as the 4Sight but could not even come close to the price. She stated that since the 4Sight testing is affiliated with the Pennsylvania Department of Education, they are able to provide the services at about half of what the Pearson assessment would cost. Further research showed that in February of 2008, Scantron Corporation purchased the former Data Management Group of Pearson and is now offering survey and research tools to schools as part of the Scantron business (retrieved at http://www.ncspearson.com/ on April 10, 2008). As a result, the research concluded that the 4Sight test is the most cost efficient

program available to schools in Pennsylvania to predict a student's performance on the PSSA exam.

2.10 Review of Other States Use of Benchmark Assessments

Before reviewing the research on how other states are using periodic assessments to prepare for the state exams, the researcher first wants to compare the current summative assessments used in Pennsylvania to the ones used in Ohio, New Jersey and New York. These states were chosen due to their close proximity.

	Pennsylvania (PSSA)	**Ohio Achievement Tests**	**New Jersey (NJ ASK)**	**New York State Testing Program**
Math	Grades 3-8, 11	Grades 3-8	Grades 3-8	Grades 3-12
Reading	Grades 3-8, 11	Grades 3-8		
Writing	Grades 5, 8, 11	Grades 3-8		
Language Arts			Grades 3-8	Grades 3-12
Science	Grades 4, 8, 11	Grades 3-8	Grades 4, 8	Grades 4, 8-12
Social Studies		Grades 3-8		Grades 5, 8-12
Foreign Language				Grades 9-12
Ohio Graduation Test (OGT)		Grade 10		
High School Proficiency Assessment (HSPA)			Grade 11	
High School Regent Examinations				Grades 9-12

(Pennsylvania Department of Education, 2007)

(Ohio Department of Education, 2007)

(State of New Jersey Department of Education, 2007)

(New York State Education Department, 2007)

The data organized in the above table shows that all four states use summative assessments in all grades 3-8 and one additional test in high school to meet the mandated requirements of NCLB. The first row of the table highlights Pennsylvania's testing schedule for PSSA administration. The Pennsylvania Department of Education has decided to administer the Math and Reading PSSA exam to all students in grades 3 through 8 and also in grade 11. The only other summative assessments used are a Writing test given in grades 5, 8, and 11 and a Science test in grades 4, 8 and 11. The other rows of the table show how each state's department of education has decided to formally assess students. While all states follow the guidelines outlined by NCLB each state has chosen to make their own decisions about what (besides Reading and Math) is tested. Ohio and New York have a formal Social Studies test that Pennsylvania and New Jersey currently do not offer. New York has a formal Foreign Language assessment that no other state offers while New Jersey has combined the Reading and Writing exams into a Language Arts assessment.

While all four states use yearly exams to assess achievement towards state standards and graduation requirements, preparation for the exams differs greatly. Each of the departments of education offers a plethora of information regarding how students can prepare for the exams. What varies is the way schools within those states choose to prepare students. For the purpose of this review, the researcher will summarize the findings from each state.

Pennsylvania

Administrators in Pennsylvania are flooded with data to help them analyze and predict PSSA results. Speaking from his own experience as a middle and high school

principal, the researcher finds the 4Sight exams a useful and extremely user-friendly tool that quickly provides the information needed to help guide instruction. The problem lies in what to do with the results. Basically it is totally up to building administrators and teachers if the results are used to guide instruction and inform practices. With more districts every year using the 4Sight exam, it is obvious to say that others find it an effective tool.

Pennsylvania 4Sight Numbers

2005/2006		2006/2007
• **1,445,833** 4Sight benchmark scores were recorded in the Member Center; • **445,040** students took 4Sight benchmarks; • **835** schools administered the 4Sight benchmarks.		• **2,832,258** 4Sight benchmark scores were recorded in the Member Center; • **762,743** students took 4Sight benchmarks; • **1,324** schools administered the 4Sight benchmarks.

Note: taken from *Pennsylvania and 4Sight*, newsletter published by Success for All, October 2007.

Ohio and New Jersey

The researcher ran into some difficulty trying to locate a formative assessment used to prepare students for the state exams in both Ohio and New Jersey. Phone calls where then made to a principal from Ohio and a Superintendent in New Jersey to hear their feelings about periodic assessments in their state. Two phone interviews were conducted on Thursday, October 4, 2007 with Mr. Bob Alsett, Principal, New Philadelphia School District, New Philadelphia, Ohio and Mr. Vince Palmieri, Superintendent, Upper Township School District, Petersburg, New Jersey. The conversations and answers to the questions were very similar. Both administrators cited numerous data collection tools. Such tools included preparatory classes, tutoring, DIBELS, SRA, words of the day, problems of the day which are taken from previous

state tests, and online preps like www.studyisland.com . These are used to assess instructional goals. Neither administrator used a formative assessment that was constructed outside of their district. In these districts, teacher made documents are created, reviewed by administration, and then approved as a periodic check of standards and eligible content. Teachers in these districts rely heavily on practice tests and released items from the state department of education as well as sample items from older tests. These materials created the foundation for which their assessments are built. Missing from their programs are the diagnostic tools and reports generated from the 4Sight exams. These administrators were very interested in the 4Sight program and each asked for a sample report to view. They explained that the biggest problem with giving teacher-made assessments on a large scale is the fact that it takes a lot of time, energy, and money to get people together to score reports and analyze results. They also said that assessing four times a year was unrealistic using their current format so both schools relied heavily on pre- and post-test in content areas to assess growth and curriculum gaps.

New York

The researcher conducted a phone interview with Mrs. Connie Evelyn, Middle School Principal, Mrs. Laura Ryder, Director of Literacy, and Mrs. Carrie Platz, Director of Mathematics for the Oswego County Schools, on Monday, October 22, 2007. New York seems to be the most aggressive of all four states reviewed in terms of their testing schedule and demands. This is the only state reviewed that has a formal Social Studies exam as part of their yearly state-wide summative assessments and a completely different set of requirements for the graduation tests. Students in New York State must pass a series of Regent Exams at the completion of each grade level (9-12) to be able to advance

to the next grade/subject. After enduring four years of testing in all subject areas, students are awarded types of diplomas based upon their test results. Students in New York State upon graduation can earn a Regents Diploma, a Regents Diploma with Advanced Designation, or a Regents Diploma or Advanced Diploma with Honors. Along with these varied diplomas is an extremely high stakes consequence for failure. Students who do not pass 5 Regents test do not graduate high school. While the interviewees informed me that their school uses only teacher made benchmark assessments, they knew of an Acuity program used in New York City.

New York City public schools use a system of periodic assessments called the Acuity Assessment Management System. While the system offers many of the same benefits of the 4Sight exam used in Pennsylvania (predictive assessments, detailed reports, item analysis) is also offers what the researcher considers to be the one factor missing from the 4Sight program. The Acuity system offers teachers two very useful tools to change instructional practices:

1. Customized Item Banks – teachers are given access to 20,000 questions aligned to the New York State standards so they have the ability to design their own classroom assessments. Teachers can choose questions based upon the needs of their class that is evidenced by this periodic assessment.
2. Aligned Instructional Materials – personalized instructional materials are provided for teachers to complement lessons. These exercises can be used with the whole group or assigned individually to students who are struggling in a certain area.

(Mc-Graw Hill, 2007)

This system seems to be the most similar to the information provided by the 4Sight test. In public education, nothing changes if classrooms do not change; school reform must impact teaching and learning. This Acuity system is similar to the 4Sight in that it puts the tools in the hands of the people who need them the most; the teachers.

Sirotnik (2004) agrees that, “Reasonable accountability systems will require a long-term focus”. A focus driven by “multiple forms of information… that informs present practice” (p. 158). Although the similarities to 4Sight are numerous, the customizable item banks, vocabulary lists, and direct links to professional development provided by this Acuity system seems to be the missing advantage that the Acuity system has over the 4Sight tests. In both tests, teachers and administrators must work together to use the data from the tests and emphasis a culture that values the assessment as a useful tool.

The researcher attempted several times to contact the Acuity Assessment company to see if they could provide him with a cost analysis of using such a program in a Pennsylvania school yet those messages were never returned.

On March 5, 2008, Dr. Nancy Madden, President and CEO of the *Success for All Foundation* informed the researcher (through email) that the following benchmarks have been established by her company in Ohio, New York and New Jersey:

Ohio: Reading grades 3-8 Math grades 3-8

New York: Reading grades 3-8

New Jersey: Reading grades 3-8 Writing grades 3-8

The list Dr. Madden provided showed 22 states where benchmarks have been established and 4Sight tests are available. Of those 22 states only in Pennsylvania, Hawaii and Mississippi were four tests available to be used. It is also important to note that Pennsylvania was the only state listed where 4Sight tests are developed beyond grade eight.

While commenting on the list of states she confirmed, “4Sight tests have been developed and are available. There are some schools using them in each state. They are

not supported by the state departments of education for statewide use as they are in Pennsylvania."

2.11 Values and drawbacks of High Stakes Testing

> "What we choose to evaluate and how we choose to evaluate delivers a powerful message to students about what we value" - Staylor & Johnson (1990)

As an administrator working in a school district on the state "warning" list, the researcher has personal experience with the pros and cons of high stakes testing. While most educators are extremely bothered by the way public schools are funded and the monetary sanctions and testing stress felt by "failing" schools, few can say that the testing is all bad. While the complaints of teachers that they are forced to teach to the test and narrow their focus is heard, it is important to also recognize the benefit to a common assessment across the state. "Some positives that arise from high-stakes testing are: 1. greater emphasis of state standards; 2. standardization across schools of what is taught and when; 3. and higher expectations" (Van Maele, p. 16). Accountability has forced educators to analyze their own instruction to find what is essential to learning. Teachers who once taught a unit on Australia in third grade because they just really had an interest in the content realized that they no longer had the time in their schedule to do this. Instead they have a grade level curriculum that focuses on specific eligible content which needs to be taught and tested on pace with the other classrooms in the school. While high stakes testing and accountability measures have not leveled the playing field between school districts, it has definitely leveled the field among classrooms in the same school. Teachers are now teaching the same content within the same grade level and can work together to plan units and themes. Parents can see that their child in a certain homeroom

is covering the same material as a student in the classroom up the hall. This common curriculum has led to a common language that can be used to create periodic common assessments. These short tests can show how students are progressing toward mastery and identify any curriculum gaps. This small, frequent assessment style can be more effective in identifying a student's strengths and weaknesses as well as an evaluative tool to inform instruction. This era of accountability and testing has created a common problem with schools across the country. Administrators who are working in affluent and progressive school districts, as well as those in struggling districts, complain about the same pressures in regards to PSSA testing and NCLB. This commonality may lead public school leaders to work together to overcome the negative aspects of high stakes testing and resist the competitiveness it has created among school districts.

Hart and Teeter (2001) examined the public's perception of educational reform using surveys and focus group research for the *Educational Testing Service*. In their study they surveyed 431 parents, 401 educators, 200 education policy makers, and created eight focus groups in California, New York, Florida and Texas. The focus groups consisted of one group of education policy makers, two groups of teachers and administrators and five groups of parents including one African-American and one Hispanic group. One of their major findings was that "testing is supported; benefits outweigh concerns" (Hart-Teeter, 2001). The following tables illustrates their findings:

Table 2.2 Testing Concerns and Beliefs

Testing Concerns and Benefits (% all adults who are very concerned/who find each very valuable)	
Concerns	Benefits
Overemphasize scores (38%)	Identify low-perform schools (46%)
Possible test bias (38%)	Guarantee basics mastered (45%)
Teaching to the test (37%)	Identify low-perform students (40%)
Too much fed influence (29%)	Comparable measurements (38%)
Too much time testing (28%)	Raise standards of excellence (30%)

Table 2.3 Testing Benefits

Benefits of Testing Outweigh Concerns Based on concerns/values you have heard about standardized testing, do you support or oppose greater use of testing as part of a broader education initiative?	
Oppose Testing 22%	Favor Testing 68%

Retrieved from ftp://ftp.ets.org/pub/corp/2001_survey_presentation.pdf on February 28, 2008.

This new era of accountability has also shined a spotlight on some of the longstanding inequalities found in many public schools in lower socio-economic areas. It is no secret that the quality of instruction delivered by teachers significantly influences the achievement of students. Since NCLB has forced districts to hire only highly qualified staff, "the law has stimulated recruitment efforts in states where low-income and 'minority' students have experienced a revolving door of inexperienced, untrained teachers" (Darling-Hammond p 2). This law truly marks the first time in history where, "student's right to qualified teachers is historically significant" (Darling-Hammond p 2).

Knowing that the tests do hold some positive outcomes for education as a whole it is also very important to highlight some of the negative effects. Administrators feel the stress and pressure of preparing students for the exam. The job of a Principal years ago is

completely different from today's role and a lot of that change can be credited to NCLB. Many Principal's job performance will be judged based upon their school's results on this year's PSSA exam. It is scary that a single measure can carry so much weight and influence so many careers. "In short, accountability becomes synonymous with a public display of judgment" (Foote, p.360).

Critics of NCLB claim that the federal government has forced an unfunded mandate that is narrowing the curriculum of many school districts and forcing teachers to focus solely on state test results. They argue that this legislation mistakenly labels successful schools as failing and even has driven middle-class students away from the public schools and into private schools (Darling-Hammond 2007).

Classroom teachers seem to be the ones with the most frustration. A survey completed by Farr (2001) with teachers in Ohio reports that, "91 percent believe that students spend too much free time preparing for tests and 91 percent feel that high-stakes tests do not support developmentally appropriate practices for students" (p. 2). Teachers are frustrated with the labels and headlines that humiliate them. "Real or authentic instruction is being replaced by teaching to the test, with the presumption being that good instruction is being squeezed out by bad" (Bror, 2006).

Most teachers lack the ability to extract data from information. The job of many building level principals is to make the piles of data make sense and use it to inform practice. If districts do not designate someone to commit to this specific task, much of the data goes unused.

So then… what is the compromise? Gulek (2003) discussed ways that teachers can help students prepare for the state exams without feeling like they are "teaching to

the test". He outlines the following chart to define appropriate and inappropriate test measures:

Appropriate and Inappropriate Test Preparation Practices

Appropriate Test Preparation	Inappropriate Test Preparation
Teaching the content of the domain to which the user wishes to infer	Engaging in instruction that limits one's ability to infer from the test score to the domain of knowledge/skill/ability
Teaching test-taking skills	Limiting content instruction to a particular item format
Teaching toward test objectives if the test objectives match the domain objectives	Teaching only those objectives from the domain that are sampled on the test
Ensuring that students understand the test vocabulary	Using an instructional guide that reviews questions from the latest issue of the test
Assessing students on various aspects of the content domain	Limiting instruction to the actual test questions

Source: Mehrens (1991, April)

The 4Sight testing plan implemented by the Western School District closely resembles the recommendations made above. The data is used to inform instruction instead of narrowing instruction. The Acuity assessment used by New York Public Schools, which combines the use of periodic assessments with user-friendly teacher items and direct links to specific professional development seems to show the most comprehensive plan for implementation. This is where building level teams and central administration need to focus efforts to provide a complete plan to effectively use the 4Sight tests.

Since Pennsylvania seems to be in the middle as far as its' use of periodic assessment when compared to Ohio, New York, and New Jersey, it is evident that we will eventually move into an even more high stakes testing environment similar to that in New York State. With recent talk of a mandatory graduation test surfacing in Pennsylvania, it is clear that we may soon be mirroring the New York State model of summative testing.

2.12 The Need for This Study

Interestingly, both the 4Sight tests and Scantron assessments in Pennsylvania along with the Acuity Assessments used in New York Public Schools claim to be able to predict how students will score on the state exams. Working with middle school students, the researcher wishes to follow students in grades six, seven, and eight through their four 4Sight exams in the 2007-2008 school year, track their predicated results on the PSSA, and then compare them to the actual results that will be received in July of 2008. Anyone with experience in a middle school building can tell you that there are many factors that account for a student's results. It will be curious to see if all of those elements truly affect outcomes. Basically, the research will be analyzing how accurate the formulas used by the *Success for All Foundation* are at actually predicting student's scores. Students scores will be charted as they progress through the testing this year to confirm the predictability component 4Sight claims for student's PSSA results. A case study format with students of Western Middle School in Pittsburgh, Pennsylvania will be used to answer the following question: Do the 4Sight exams accurately predict student's scores on the PSSA exam? For further investigation, the researcher will also use a Kappa Coefficient to analyze the ability of the 4Sight test to accurately predict student's categorical classification on the PSSA exam.

Cimbricz (2002) while reporting from a review of literature on teacher beliefs and practices and state testing stated that most of the literature related to these issues contained theory and rhetoric. Lutz-Doemling (2007) while analyzing this review stated, "Very little research included objective data collected through classroom observations" (p. 18). Even Cimbricz (2002) felt that the relationship between teacher's practices and

high-stakes testing could not be established as positive or negative until further observational research can be conducted using specific teachers and students.

While reviewing the 2005 technical manual provided by the *Success for all Foundation* Lutz-Doemling (2007) identified several limitations. The researcher highlighted that the technical report only shows Reading data gathered from students in grades three, five and eight. Probably the most alarming finding was that, "The math correlations were determined as a result of a pilot study where the 4Sight Mathematics Benchmark Assessments were administered 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" (p. 40). This study will be administered to students on grade level in grades six, seven and eight and their predicted results as stated by the *Success for all Foundation* and reported on their members center web site will be compared to the same students actual results on the 2008 PSSA Reading and Mathematics exam that will be administered April 1, 2, and 3 in the Western Middle School. The results that will be received near the first week of July 2008 will be compared to the predictions made by the third and fourth 4Sight tests that will be administered in January and then one week after the PSSA exam. This close time proximity of the fourth test sets the base for a level comparison while the comparison of the third test coincides with the previous study conducted by the *Success for All Foundation*.

As the current middle school Assistant Principal at Western School District the implementation and use of 4Sight testing is one of the researchers' main responsibilities. This entails the creation of a testing schedule and the designing of a systematic way to

use the data from the exam to inform instruction and school reform. Since the Western Middle School is currently on the "warning" list for not making adequate yearly progress, these benchmark assessments are viewed to be a major resource to make informed curriculum decisions, identify specific areas of need, and offer predictions of how students will score on the upcoming PSSA exams. The specific interest in this study will be the degree of correlation that exists between the predictions made by the *Success for All Foundation* and the actual PSSA results. Incorporating the time into the school calendar to administer four to five 4Sight tests sacrifices crucial learning time out of students' regular classrooms. Teachers and parents argue that this is just another test, unless we can prove to them that the information gained is valuable and the predictions made are accurate. These adults do not want to hear about what the company says the test can do or what the ads in the newsletter claim. They want to see actual results within a normal school setting. These correlations will validate or dispel the claim that there is not a need for any additional testing in Pennsylvania middle schools. A positive, strong correlation will show the time and money is well spent. A weak correlation that evidences inconsistent results of prediction will strengthen the argument that the mandatory PSSA exams are enough testing for middle school students.

2.13 Conclusion

The United States in now at an important crossroad in educational testing. Today's public schools are seeing another shift in demographic population, an increase in the consequences from not meeting testing requirements, and a new awareness of global competition. The history of educational testing shows that as political and federal roles

in public school testing increase, so do the negative consequences many districts face. The lessons gained from looking into the history of testing in the United States provide useful background information necessary to the development of future testing policies.

There is a lot of research that supports that the use of assessments for learning (formative assessments) can chart student progress and highlight weak areas in overall instruction. The ability of a benchmark assessment to predict students results on a once a year summative assessment like the PSSA exam needs further research and investigation. Most of the research published that shows comparison between predicted 4Sight results and actual PSSA results was either published by the Success for All Foundation or with the affiliation of Dr. Robert Slavin or Dr. Nancy Madden. A study independent from direct corporate influence and with no real benefit of either confirming or denying predictions could yield tremendous insight.

If the 4Sight benchmark assessment truly can predict student's results on the PSSA exam, Dr. Slavin and Dr. Maddden have discovered a breakthrough in educational testing. Their ability to formulate an exam that nullifies the external factors that some argue affect a child's performance on standardized testing would quiet many of the critics of high-stakes testing. Instead, this exam may show a pattern or progression to a child's current level of achievement that would prove that educational success is a continuum built on assessment driven instruction.

Chapter 3

Methodology

3.1 Problem Statement

In the January/February 2008 newsletter published by the *Success for All Foundation* tilted, "Pennsylvania and the 4Sight" it is reported that 391 school districts across the commonwealth and 1,785 schools are now using the 4Sight test in Pennsylvania. The newsletter states, "Pennsylvania schools nearly doubled the number of 4Sight tests in the system. Clearly, the number of reporting tools being used demonstrates that Pennsylvania understands the power of the program: using data to drive instruction" (p. 1). Administrators and teachers in these districts rely heavily on the predictability of these assessments to highlight specific areas of weakness in the curriculum and in student learning. These results guide professional development planning and determine the spending practices of many local districts.

No study has ever followed an entire school population through the 4Sight and PSSA exams being administered on grade level and analyzed predicated results to actual results. A case study format with students of Western Middle School in Pittsburgh, Pennsylvania will be used to answer the following question:

1. Do the 4Sight exams accurately predict student's raw scores on the PSSA exam?
2. Is there a high correlation between the predicted categorical classifications of students on the 4Sight tests when compared to actual PSSA Results?
3. Is the information gained from the fourth 4Sight test worth the loss of instructional time?

3.2 Setting

The school district being reviewed is located in South Western Pennsylvania in the suburbs of Pittsburgh. It is made up of five elementary schools, an Early Childhood Education Center, a Middle School and a High School. The total school population for the 2007-2008 school year is 3,281 students in grades kindergarten through 12 and the district employs over 220 full-time teachers. The district spends approximately $9,560 per student per school year.

As of the 2000 Census, the school community had a population of 22,464 people with 6,475 families residing there. The population density is about 1,586.2 people per square mile. The racial makeup of the borough is 89.64% White, 8.85% African American, 0.12% Native American, 0.25% Asian, 0.06% Pacific Islander, and 0.25% Other. Of the 9,509 households, 26.8% have children under the age of 18 years old, 50.6% are married couples, 13.7% have a female with no male involved, 31.9% are non families, 29.0% are individuals living alone, and 15.3% are elderly persons of the age of 65 or older living alone. The average household size in 2000 was 2.35 and the average family size was 2.89 and the median income per household $36,130.00 (United States Census Bureau, 2000).

3.3 Research Population

The researcher will follow the 6th, 7th, and 8th grade students enrolled during the 2007-2008 school year as they take the 4Sight benchmark assessment and the 2008 PSSA exam. Permission from the Superintendent of Schools to use the 4Sight data and PSSA information has been granted through written confirmation. No students' names will be

used and results will be generalized to the total student population by building and grade level. The school will be referred to as the Western Middle School.

3.4 Data Collection Procedures

In April of 2008 there were 731 students enrolled in the Western Middle School and the distribution among grade levels is as follows:

Grade 6- 258 students
Grade 7- 242 students
Grade 8- 231 students

Following the recommendation of the *Success for All Foundation*, four 4Sight benchmark assessments of Reading and Mathematics were administered periodically throughout the school year. The dates of testing were as follows:

4Sight Test #1 – September 11, 2007
4Sight Test #2 – November 22, 2007
4Sight Test #3 – January 12, 2008
4Sight Test #4 – April 14, 2008

While setting the examination schedule as shown above agreed with recommendations made by the *Success for All Foundation*, it also allowed for proper observation of student growth throughout the testing. An additional observation made by the researcher while using this administration schedule was that each time students were administered the exam, they were more familiar with the location of their testing room and materials necessary for the testing session.

Students in grades 6, 7, and 8 also took the 2008 PSSA exam in Reading and Mathematics on April 1, 2, and 3, 2008.

For the purpose of this study, the third and fourth 4Sight test raw scores and categorical classifications will be correlated to the actual PSSA results to determine the relationship between the two exams. Only students who took all four 4Sight exams and the PSSA test will be used in the correlation analysis. This decision was made because the researcher felt that students who missed one or more administrations of the 4Sight test would be less familiar with its format and method of administration than students present for all four test administrations. Using this criterion, 94% of the total student population qualified for the correlation analysis. The table below shows the testing breakdown by grade level:

Table 3.1 Students Qualifying for Study

Grade	Total Enrollment	Number of students who qualify	Percent of total enrollment
6	258	234	91%
7	242	232	96%
8	231	220	95%
Totals	731	686	94%

3.5 Data Analysis Procedures

Students will be grouped by grade level and assigned a number. Individual test results will be charted for each round of testing. A raw score and categorical classification (advanced, proficient, basic, below basic) will be assigned to each student based upon the results as reported on the Member's Center website (located at https://members.successforall.net/) provided by the *Success for All Foundation*. Student tests will be scored using the scanning software provided by the *Success for All*

Foundation as part of the 4Sight testing package that was purchased by the Western School District.

The researcher will specifically analyze the third and fourth 4Sight test results that will be used to measure correlation with actual PSSA results. An example of the data table is shown here:

Table 3.2 Organization of Student Data

GRADE 6 Student ID	4-Sight Math 3 Raw	4-Sight Math 3 Category	4-Sight Math 4 Raw	4-Sight Math 4 Category	4-Sight Read 3 Raw	4-Sight Read 3 Category	4-Sight Read 4 Raw	4-Sight Read 4 Category	PSSA Math Raw	PSSA Math cat	PSSA Read Raw	PSSA Read Cat
Student #1	27	Proficient	29	Advanced	23	Basic	25	Basic	1125	Proficient	1257	Proficient
Student #2	24	Basic	26	Basic	23	Baisc	27	Proficient	1243	Proficient	1288	Proficient

Results will be analyzed by grade level and then generalized to the entire building population.

After all data is charted in the above format, the researcher will:

1. Use Linear Regression to calculate how well the 4sight scores from the third and fourth administrations respectively predict the PSSA results in Reading and Mathematics. The researcher will use the third 4Sight because the *Success for All Foundation* states in the 2007-2008 Technical Report that third test results were used in their field study to assess prediction correlation with the PSSA (Success for All, 2007). The researcher is using the fourth 4Sight test as an addition measure of correlation because of the close time proximity between the fourth 4Sight test and the PSSA exam.

2. Use a Kappa coefficient formula to compare the classification categories predicted by the third and fourth 4Sight tests to the actual classifications from the PSSA results (Advanced, Proficient, Basic, and Below Basic.)

In the January/February 2008 Newsletter published by the *Success for All Foundation* it was stated that, "The 4Sight correlations are determined by a statistical process called linear regression" (p 2). Two sets of scores, the 4Sight and PSSA, are used to develop an equation that relates one score to the other. "Once the equation is determined, one can plug in a 4sight score (X) and get a PSSA score (Y)." (Success for All, 2008). In Linear Regression, the independent variable (X) is used to predict the dependent variable (Y). This analysis is a statistical technique for investigating and modeling the relationship between two variables (Sun, 2008). Although Linear Regression does not imply that one variable causes another variable; it does indicate how accurately scores on the dependent variable can be predicted by the independent variable (Montgomery, Peck & Vining, (2001). The researcher chose to replicate this process in this study to mirror the design used but the *Success for All Foundation*. Results will be graphed to show the form (shape) of the relationship.

To analyze correlation among the 4Sight test results and the PSSA exam further, the researcher is also using a Kappa Coefficient formula to measure the agreement between predicted categorical classifications of students and actual PSSA categories. Like the PSSA exam, the 4Sight test lists student's categorical classifications as Advanced, Proficient, Basic, or Below Basic. The Kappa Coefficient is a statistical measure of the agreement between two raters who classify items into exclusive categories (childrensmercy.org). "Kappa has long been used in content analysis and medicine to

assess the reliability of tagging" (Di Eugenio, 2000). In other words, this was used to assess how well medical students' diagnosis based upon a set of facts would agree with expert diagnosis. The researcher believes that this technique of analysis will offer interesting insight into the categorical predications made by the 4Sight company when compared to actual results. The following formula will be used to compute kappa's possible value (Shoukri, 2004):

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

In the above equation, P(A) is the observed agreement among raters and P(E) is the likelihood that the agreement is due to chance. The possible values of Kappa are between 0 and 1. When K equals 0, this means that the agreement is no greater than would be expected by chance and when K equals 1, this means perfect agreement.

Landis & Koch (1997) suggests the following kappa interpretation scale:

Kappa Value	**Interpretation**
Below 0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

Gardner (1995) recommends that kappa exceed .70 before you continue with any additional data analyses.

The researcher will use the data table (A) shown above to run both the linear regression and kappa coefficient. Findings will be analyzed and summarized in later chapters.

Chapter 4

Findings

4.1 Introduction

This study followed students in a suburban middle school (grades 6, 7 and 8) in Western Pennsylvania and was meant to compare and contrast the predictions made by the 4Sight tests (given during the 2007-2008 school year) to the actual student results on the spring 2008 PSSA exam. The Western Middle School administered four rounds of 4Sight testing during the 2007-2008 school year with the last round being administered one week after the PSSA exam. Student's results from the third and fourth round of 4Sight testing were analyzed and compared to actual student results on the 2008 PSSA exam. The raw scores were analyzed using linear regression between students' third and fourth 4Sight test results and actual PSSA results to check for correlation. Student's predicted categorical classifications from the third and fourth 4Sight test were then compared to actual PSSA results using a kappa coefficient formula.

Both the linear regression and Kappa coefficient were used to answer the primary research questions of the study:

1. Do the 4Sight exams accurately predict student's raw scores on the PSSA exam?
2. Is there a high correlation between the predicted categorical classifications of students on the 4Sight tests when compared to actual PSSA Results?
3. Is the information gained from the fourth 4Sight test worth the loss of instructional time?

Presentation of the Data Related to Research Questions #1and #3

Research Question #1 **Do the 4Sight exams accurately predict student's raw scores on the PSSA exam?**

4.2 Descriptive Statistics

The following descriptive statistics highlight students' mean score, median score, the standard deviation between scores, and the minimum and maximum score for each round of testing. The results were tabulated and organized by grade level and then shown together to reflect the entire student population. Each test is labeled on the left side of the table so results from the third and fourth 4Sight Math and Reading test along with the PSSA Reading and Math tests can be observed. Appendix A outlines the performance level cut scores for the 2008 PSSA exam so that students' mean (average score) results can be compared to the expectations set by the Pennsylvania Department of Education.

Table 4.1 Grade 6 Descriptive Statistics

Grade 6 – Descriptive Statistics					
RAW SCORES	Mean	Median	Standard Deviation	Minimum	Maximum
Third 4Sight Math	20.79	21	6.225	3	35
Fourth 4Sight Math	26.59	28	5.866	7	36
Third 4Sight Reading	21.17	22	5.146	5	29
Fourth 4Sight Reading	22.21	23	4.732	4	30
PSSA Math	1551.91	1554	225.818	948	2050
PSSA Reading	1360.24	1362	199.282	829	2044

The results from the sixth grade testing show a 5.8 point increase in the mean score and a 7 point increase in the median score from the third to the fourth 4Sight Math exam. The 4Sight Reading results show a 1.04 point increase in the mean score and a 1

point increase in the median raw score. The researcher feels that these results show what was expected since the third round of testing was administered in January 2008 and the fourth round of testing was administered in April 2008. Students had time to cover much more of the curriculum during that time so higher results should be likely. It is interesting to note that the minimum score for the 4Sight Math test rose 4 points while the minimum score for the 4Sight Reading test dropped 1 point.

In grade 6, the average raw score on the PSSA Math exam was 1551.91 which places the mean score in the advanced performance range. The average raw score on the PSSA Reading exam was 1360.24 which places the mean score in the Proficient performance range.

Table 4.2 Grade 7 Descriptive Statistics

Grade 7 – Descriptive Statistics					
RAW SCORES	Mean	Median	Standard Deviation	Minimum	Maximum
Third 4Sight Math	19.42	19	6.746	6	36
Fourth 4Sight Math	25.10	26	6.825	6	36
Third 4Sight Reading	21.34	22	4.403	7	29
Fourth 4Sight Reading	21.16	21.43	4.932	7	29
PSSA Math	1465.78	1453	196.508	985	2407
PSSA Reading	1377.42	1379	193.582	913	1961

The results from the seventh grade testing show a 5.68 point increase from the third to fourth 4Sight Math test in the mean score and a 7 point increase in the median score. The results from the third and fourth Reading test show a decrease of .18 in the mean score and a decrease in .57 in the median score. It is interesting to note that the minimum and maximum scores for both test remained exactly the same.

In grade 7, the average raw score on the PSSA Math exam was 1465.78 which places the mean score in the proficient performance range. The average raw score on the

PSSA Reading exam was 1377.42 which places the mean score in the Proficient performance range.

Table 4.3 Grade 8 Descriptive Statistics

Grade 8 – Descriptive Statistics					
RAW SCORES	Mean	Median	Standard Deviation	Minimum	Maximum
Third 4Sight Math	17.76	18	6.252	2	32
Fourth 4Sight Math	21.69	22.5	6.410	4	33
Third 4Sight Reading	22.38	24	5.233	6	30
Fourth 4Sight Reading	21.96	23	5.147	6	30
PSSA Math	1432.05	1427	220.730	933	2270
PSSA Reading	1481.84	1507	243.09	780	2049

The results from the eighth grade 4Sight Math test show a 3.93 increase in the mean score and a 4.5 point increase in the median score. The results from the 4Sight Reading test show a 0.42 point decrease in the mean score and a 1 point decrease in the median score. It is interesting to note that both the minimum and maximum score for the Math test increased while the minimum and maximum score for the Reading test stayed the same from the third to fourth 4Sight test.

In grade 8, the average raw score on the PSSA Math exam was 1432.05 which places the mean score in the proficient performance range. The average raw score on the PSSA Reading exam was 1481.84 which places the mean score in the advanced performance range.

Table 4.4 Total School Descriptive Statistics

Total School Population – Descriptive Statistics					
RAW SCORES	Mean	Median	Standard Deviation	Minimum	Maximum
Third 4Sight Math	19.35	19	6.523	2	36
Fourth 4Sight Math	24.52	26	6.687	4	36
Third 4Sight Reading	21.61	22	4.957	5	30
Fourth 4Sight Reading	21.77	23	4.949	4	30
PSSA Math	1484.29	1475	220.161	933	2407
PSSA Reading	1404.96	1406	218.743	780	2049

When the results are generalized to the entire school population (who qualified for the study) it reflects a much higher increase in the Math scores than the Reading scores. The overall Math scores for the building show a 5.17 point increase in the mean score and a 7 point increase in the median score. The Math scores also show a 2 point increase in the minimum score while the maximum score stayed the same.

The results of the Reading test show a slight 0.16 point increase in the mean score and a 1 point increase in the median raw score. The minimum score decreased 1 point while the maximum score stayed the same.

Histograms that show the frequency of student's scores for the third 4Sight and fourth 4Sight exam as well as the 2008 Math and Reading PSSA exam can be found in Appendix B.

When analyzing the entire school population, the average raw score on the PSSA Math exam was 1484.25 and the average raw score on the PSSA Reading exam was 1404.96. Both scores average building results to be near the high proficient performance level on the 2008 PSSA exam.

4.3 Linear Regression Results

The researcher used Linear Regression to calculate how well the 4sight scores from the third and fourth administrations respectively predict the PSSA results in Reading and Mathematics. The correlation analysis produced a Pearson correlation coefficient (r) score and reflected the correlation between 2 variables (one of the 4Sight test and the same subject PSSA exam). The (r) score can range from 0 to 1 with scores closer to 1 showing a stronger correlation and scores close to 0 showing a weaker correlation. An (r) value of 1 shows that a linear equation describes the relationship perfectly and a value of 0 shows that there is no linear relationship between the two variables (Glass & Hopkins, 1996). The results were tabulated and shown using the table below:

Table 4.5 Linear Regression Values

Linear Regression R (Pearson Correlation) Values				
Raw Scores	**Grade 6**	**Grade 7**	**Grade 8**	**Total School Population**
Third 4Sight Math/PSSA Math	0.817	0.829	0.846	0.831
Fourth 4Sight Math/PSSA Math	0.806	0.812	0.851	0.820
Third 4Sight Reading/PSSA Reading	0.721	0.756	0.845	0.775
Fourth 4Sight Reading/PSSA Reading	0.755	0.777	0.800	0.753

The table above shows that when the constant variable (third 4Sight Math raw scores – Grade 6) was compared to the dependent variable (PSSA Math raw scores – Grade 6) it resulted in a Pearson correlation of 0.817. This shows a very strong agreement between the predicted raw scores from the third 4Sight Math test and actual raw scores on the PSSA Math exam. In each case above, the constant variable was one of the 4Sight tests (listed first in the first column) and the dependent variable was one of the PSSA tests (listed second in the first column).

The strongest correlation shown is between the Grade 8 fourth 4Sight Math raw scores and the Grade 8 PSSA Math raw scores with an (r) value of 0.851. The weakest correlation shown (which is still a strong correlation agreement) is between the Grade 6 third 4Sight Reading test and the Grade 6 PSSA Reading test with an (r) value of 0.721.

While analyzing the entire school as a whole, the correlation between the third 4Sight Math raw scores and the PSSA Math raw scores was 0.011 points higher than the correlation between the fourth 4Sight Math raw score and the PSSA Math raw score. The same correlation is also shown on the Reading test. The third 4Sight Reading test shows a 0.022 point stronger correlation to the PSSA Reading exam than the fourth 4Sight test does. Overall, the results for the entire building show a higher correlation between predicted and actual Math scores when compared to Reading scores.

Summary of the Findings Related to Research Questions #1and #3

Research Question #1 Do the 4Sight exams accurately predict student's raw scores on the PSSA exam?

The results of the linear regression clearly show a strong, significant correlation between predicted raw scores on the third and fourth 4Sight test and the actual results on the PSSA exam. This would confirm the claim made by the *Success for All Foundation* that the 4Sight test can accurately predict PSSA raw score results. In addition, this analysis also provides valuable information to administrators and teachers who are using the 4Sight test to inform instructional practice.

While the analysis in this one case shows that the 4Sight test can predict PSSA results, secondary findings demonstrate that four tests in addition to the PSSA exam may

support the argument of over-testing. As stated in earlier chapters, the third round of 4Sight testing was administered on January 12, 2008 while the fourth round of 4Sight testing was administered April 14, 2008 (two weeks after the PSSA exam). While the exams were nearly three months apart, the correlation above shows that a stronger agreement exists between the predicted raw score results from the third 4Sight test than the predicted results of the fourth 4Sight test. While students are exposed to more of the curriculum over that three month time span, the higher correlation on the third 4Sight exam opens the argument that administering the fourth 4Sight exam is unnecessary.

Presentation of the Data Related to Research Questions #2 and #3

Research Question #2 Is **there a high correlation between the predicted categorical classifications of students on the 4Sight tests when compared to actual PSSA Results?**

The researcher used a Kappa coefficient formula to compare the classification categories predicted by the third and fourth 4Sight tests to the actual classifications from the PSSA results (Advanced, Proficient, Basic, and Below Basic). Results were organized and separated by each 4Sight test and comparisons where shown between predicted categorical classifications and actual results on the PSSA exam. The tables below show results comparing each 4Sight exam to the PSSA exam and a table for each grade level and test was created. Finally, a table showing the combination of the entire school population was used to show the data as a whole.

A Kappa coefficient for each exam was also listed to show the level of agreement between the predicted and actual scores. As stated earlier, Landis & Koch (1997) suggests the following kappa interpretation scale:

Kappa Value	Interpretation
Below 0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

Gardner (1995) recommends that kappa exceed .70 before you continue with any additional data analysis.

The yellow highlighted areas of the tables below show when predicted categorical classifications from the 4Sight test matched perfectly to actual PSSA classifications. The areas outlined in red show where students did worse on the PSSA test than predicted by the 4Sight exam. The areas outlined in green show where students scored in a higher categorical classification on the PSSA test than was predicted on the 4Sight exam. In summary, when a 4Sight predicted categorical classification matched perfectly with the actual PSSA classification it is listed in the yellow highlighted column. When the PSSA categorical classification was higher than the predicted categorical classification from the 4Sight test it is outlined in the green area. When the PSSA categorical classification was lower than predicted by the 4Sight test it is outlined in the red area.

Table 4.6 Grade 6 Math Regression Results – 3rd 4Sight

Grade 6 – Math		PSSA Math Category			
		Below Basic	Basic	Proficient	Advanced
Third 4Sight Math Category	Below Basic	14	14	29	17
	Basic	0	3	21	56
	Proficient	0	0	2	62
	Advanced	0	0	0	16

This table compares the third 4Sight Math predicted categorical classifications to the actual categorical results on the PSSA exam. The left column lists the 4 possible classification categories and the top column shows the actual PSSA results. The kappa coefficient shows that:

- 14 students who were predicted to score Below Basic (by the third 4Sight test) on the PSSA exam did score Below Basic.
- 3 students who were predicted to score Basic did score Basic.
- 2 students who were predicted to score Proficient did score Proficient
- 16 students who were predicted to score Advanced did score Advanced.
- 35 students out of 234 scored in the same categorical classification on the PSSA exam that was predicted by the third 4Sight exam. The measure of agreement using the Kappa Coefficient is .001 which is considered to be a poor/slight agreement.
- 0 students scored in a lower categorical classification than predicted
- 199 out of 234 sixth grade students scored in a higher categorical classification than was predicted.

Table 4.7 Grade 6 Math Regression Results – 4th 4Sight

Grade 6 – Math		PSSA Math Category			
		Below Basic	Basic	Proficient	Advanced
Fourth 4Sight Math Category	Below Basic	11	6	2	0
	Basic	3	6	21	12
	Proficient	0	5	25	84
	Advanced	0	0	4	55

The above table compares the fourth 4Sight Math test to the PSSA Math exam. The results show:

- 97 students out of 234 scored in the same categorical classification on the PSSA exam that was predicted by the fourth 4Sight exam. The measure of agreement using the Kappa Coefficient is .177 which is considered to be a slight agreement.
- 12 students out of 234 scored worse than predicted.
- 125 students out of 234 scored better than predicted.

Table 4.8 Grade 6 Reading Regression Results -3rd 4Sight

Grade 6 – Reading		PSSA Reading Category			
		Below Basic	Basic	Proficient	Advanced
Third 4Sight Reading Category	Below Basic	16	9	2	1
	Basic	4	35	25	8
	Proficient	1	10	59	60
	Advanced	0	0	2	2

The above table compares the third 4Sight Reading test to the PSSA Reading exam. The results show:

- 112 students out of 234 scored in the same categorical classification on the PSSA exam that was predicted by the third 4Sight exam. The measure of agreement using the Kappa Coefficient is .260 which is considered to be a fair agreement.
- 17 students out of 234 scored worse than predicted.
- 105 students out of 234 scored better than predicted

Table 4.9 Grade 6 Reading Regression Results – 4th 4Sight

Grade 6 – Reading		**PSSA Reading Category**			
		Below Basic	Basic	Proficient	Advanced
Fourth 4Sight Reading Category	Below Basic	20	13	5	0
	Basic	0	32	26	3
	Proficient	1	9	55	59
	Advanced	0	0	2	9

The above table compares the fourth 4Sight Reading test to the PSSA Reading exam. The results show:

- 116 students out of 234 scored in the same categorical classification on the PSSA exam that was predicted by the fourth 4Sight exam. The measure of agreement using the Kappa Coefficient is .291 which is considered to be a fair agreement.
- 12 students out of 234 scored worse than predicted.
- 106 students out of 234 scored better than predicted

Table 4.10 Grade 7 Math Regression Results – 3rd 4Sight

Grade 7 – Math		PSSA Math Category			
		Below Basic	Basic	Proficient	Advanced
Third 4Sight Math Category	Below Basic	15	28	25	4
	Basic	0	6	35	27
	Proficient	0	0	17	43
	Advanced	0	0	1	33

The above table compares the third 4Sight Math test to the PSSA Math exam. The results show:

- 71 students out of 234 scored in the same categorical classification on the PSSA exam that was predicted by the third 4Sight exam. The measure of agreement using the Kappa Coefficient is .114 which is considered to be a slight agreement.
- 1 student out of 234 scored worse than predicted.
- 162 students out of 234 scored better than predicted.

Table 4.11 Grade 7 Math Regression Results – 4th 4Sight

Grade 7 – Math		PSSA Math Category			
		Below Basic	Basic	Proficient	Advanced
Fourth 4Sight Math Category	Below Basic	14	9	4	0
	Basic	0	16	14	1
	Proficient	1	9	43	21
	Advanced	0	0	17	85

The above table compares the fourth 4Sight Math test to the PSSA Math exam. The results show:

- 158 students out of 234 scored in the same categorical classification on the PSSA exam that was predicted by the fourth 4Sight exam. The measure of agreement

using the Kappa Coefficient is .514 which is considered to be a moderate agreement.

- 27 student2 out of 234 scored worse than predicted.
- 49 students out of 234 scored better than predicted.

Table 4.12 Grade 7 Reading Regression Results – 3rd 4Sight

Grade 7 – Reading		**PSSA Reading Category**			
		Below Basic	Basic	Proficient	Advanced
Third 4Sight Reading Category	Below Basic	14	3	2	0
	Basic	9	19	22	0
	Proficient	2	19	79	44
	Advanced	0	0	2	17

The above table compares the third 4Sight Reading test to the PSSA Reading exam. The results show:

- 129 students out of 232 scored in the same categorical classification on the PSSA exam that was predicted by the third 4Sight exam. The measure of agreement using the Kappa Coefficient is .318 which is considered to be a fair agreement.
- 32 students out of 232 scored worse than predicted.
- 71 students out of 232 scored better than predicted.

Table 4.13 Grade 7 Reading Regression Results – 4th 4Sight

Grade 7 – Reading		PSSA Reading Category			
		Below Basic	Basic	Proficient	Advanced
Fourth 4Sight Reading Category	Below Basic	20	11	3	0
	Basic	5	14	27	3
	Proficient	0	15	71	36
	Advanced	0	1	4	22

The above table compares the fourth 4Sight Reading test to the PSSA Reading exam.

The results show:

- 127 students out of 232 scored in the same categorical classification on the PSSA exam that was predicted by the fourth 4Sight exam. The measure of agreement using the Kappa Coefficient is .333 which is considered to be a fair agreement.
- 25 students out of 232 scored worse than predicted.
- 80 students out of 232 scored better than predicted

Table 4.14 Grade 8 Math Regression Results – 3rd 4Sight

Grade 8 – Math		PSSA Math Category			
		Below Basic	Basic	Proficient	Advanced
Third 4Sight Math Category	Below Basic	24	29	14	1
	Basic	1	11	33	17
	Proficient	0	1	17	50
	Advanced	0	0	1	31

The above table compares the third 4Sight Math test to the PSSA Math exam.

The results show:

- 83 students out of 220 scored in the same categorical classification on the PSSA exam that was predicted by the third 4Sight exam. The measure of agreement using the Kappa Coefficient is .195 which is considered to be a slight agreement.
- 3 students out of 220 scored worse than predicted.
- 135 students out of 220 scored better than predicted.

Table 4.15 Grade 8 Math Regression Results – 4th 4Sight

Grade 8 – Math		**PSSA Math Category**			
		Below Basic	Basic	Proficient	Advanced
Fourth 4Sight Math Category	Below Basic	21	6	2	0
	Basic	4	20	9	0
	Proficient	0	5	45	23
	Advanced	0	0	9	76

The above table compares the fourth 4Sight Math test to the PSSA Math exam. The results show:

- 162 students out of 220 scored in the same categorical classification on the PSSA exam that was predicted by the fourth 4Sight exam. The measure of agreement using the Kappa Coefficient is .619 which is considered to be a substantial agreement.
- 18 students out of 220 scored worse than predicted.
- 40 students out of 220 scored better than predicted.

Table 4.16 Grade 8 Reading Regression Results – 3rd 4Sight

Grade 8 – Reading		PSSA Reading Category			
		Below Basic	Basic	Proficient	Advanced
Third 4Sight Reading Category	Below Basic	14	3	1	0
	Basic	8	10	24	6
	Proficient	0	4	26	121
	Advanced	0	0	0	3

The above table compares the third 4Sight Reading test to the PSSA Reading exam. The results show:

- 53 students out of 220 scored in the same categorical classification on the PSSA exam that was predicted by the third 4Sight exam. The measure of agreement using the Kappa Coefficient is .060 which is considered to be a poor/slight agreement.
- 12 students out of 220 scored worse than predicted.
- 155 students out of 220 scored better than predicted

Table 4.17 Grade 8 Reading Regression Results – 4th 4Sight

Grade 8 – Reading		PSSA Reading Category			
		Below Basic	Basic	Proficient	Advanced
Fourth 4Sight Reading Category	Below Basic	22	10	7	1
	Basic	0	5	24	16
	Proficient	0	2	20	91
	Advanced	0	0	0	22

The above table compares the fourth 4Sight Reading test to the PSSA Reading exam. The results show:

- 69 students out of 220 scored in the same categorical classification on the PSSA exam that was predicted by the fourth 4Sight exam. The measure of agreement using the Kappa Coefficient is .129 which is considered to be a slight agreement.
- 2 students out of 220 scored worse than predicted.
- 149 students out of 220 scored better than predicted

After each grade level was shown individually, the researcher decided to combine the entire school population to compare Kappa agreement. The following tables highlight those results.

Table 4.18 Total School Math Regression Results – 3rd 4Sight

Total School Population – Math		**PSSA Math Category**			
		Below Basic	Basic	Proficient	Advanced
Third 4Sight Math Category	Below Basic	53	61	68	22
	Basic	1	20	89	100
	Proficient	0	1	36	155
	Advanced	0	0	2	80

The above table compares the third 4Sight Math test results for the entire building to the results of the PSSA Math exam. The results show:

- 189 students out of 688 scored in the same categorical classification on the PSSA exam that was predicted by the third 4Sight exam. The measure of agreement using the Kappa Coefficient is .093 which is considered to be a slight agreement.
- 4 students out of 688 scored worse than predicted.
- 495 students out of 688 scored better than predicted.

Table 4.19 Total School Math Regression Results – 4th 4Sight

Total School Population – Math		PSSA Math Category			
		Below Basic	Basic	Proficient	Advanced
Fourth 4Sight Math Category	Below Basic	46	21	8	0
	Basic	7	42	44	13
	Proficient	1	19	113	128
	Advanced	0	0	30	216

The above table compares the fourth 4Sight Math test results for the entire building to the results of the PSSA Math exam. The results show:

- 417 students out of 688 scored in the same categorical classification on the PSSA exam that was predicted by the fourth 4Sight exam. The measure of agreement using the Kappa Coefficient is .712 which is considered to be a substantial agreement.
- 57 students out of 688 scored worse than predicted.
- 214 students out of 688 scored better than predicted.

Table 4.20 Total School Reading Regression Results – 3rd 4Sight

Total School Population – Reading		PSSA Reading Category			
		Below Basic	Basic	Proficient	Advanced
Third 4Sight Reading Category	Below Basic	44	15	5	1
	Basic	21	64	71	14
	Proficient	3	33	164	225
	Advanced	0	0	4	22

The above table compares the third 4Sight Reading test results for the entire building to the results of the PSSA Reading exam. The results show:

- 294 students out of 686 scored in the same categorical classification on the PSSA exam that was predicted by the third 4Sight exam. The measure of agreement using the Kappa Coefficient is .201 which is considered to be a slight/fair agreement.
- 61 students out of 686 scored worse than predicted.
- 331 students out of 686 scored better than predicted.

Table 4.21 Total School Math Regression Results – 4th 4Sight

Total School Population – Reading		**PSSA Reading Category**			
		Below Basic	Basic	Proficient	Advanced
Fourth 4Sight Reading Category	Below Basic	62	34	15	1
	Basic	5	51	77	22
	Proficient	1	26	146	186
	Advanced	0	1	6	53

The above table compares the fourth 4Sight Reading test results for the entire building to the results of the PSSA Reading exam. The results show:

- 312 students out of 686 scored in the same categorical classification on the PSSA exam that was predicted by the third 4Sight exam. The measure of agreement using the Kappa Coefficient is .250 which is considered to be a fair agreement.
- 39 students out of 686 scored worse than predicted.
- 335 students out of 686 scored better than predicted.

4.4 Analysis of Kappa Coefficient Agreement

Table 4.22 Kappa Coefficient Analysis

Kappa Coefficient				
Categorical Classification	**Grade 6**	**Grade 7**	**Grade 8**	**Total School Population**
Third 4Sight Math/PSSA Math	0.001 Poor	0.114 Slight	0.195 Slight	0.093 Slight
Fourth 4Sight Math/PSSA Math	0.177 Slight	0.514 Moderate	0.619 Substantial	0.712 Substantial
Third 4Sight Reading/PSSA Reading	0.260 Fair	0.318 Fair	0.060 Poor	0.201 Slight
Fourth 4Sight Reading/PSSA Reading	0.291 Fair	0.333 Fair	0.129 Slight	0.250 Fair

The above table organizes the Kappa Coefficient analysis between all the 4Sight and PSSA exams previously discussed. Included in each description are the kappa score and the kappa interpretation of the agreement. Each 4Sight and PSSA exam is aligned horizontally to show agreement and each grade level was given a column to show comparison. The final column shows agreement (kappa score) generalized to the entire school population.

Summary of the Findings Related to Research Questions #2 and #3

Research Question #2 **Is there a high correlation between the predicted categorical classifications of students on the 4Sight tests when compared to actual PSSA Results?**

According to the Kappa interpretation scale developed by Landis & Koch (1997) only the Grade 8 fourth 4Sight Math test and the PSSA Math test have a substantial agreement. While the Grade 7 fourth 4Sight Math test had a moderate agreement with the PSSA Math results, all other correlations ranged from fair to poor. Six of the twelve

correlations were considered to have poor to slight correlation. These results show a low correlation between predicted classification categories and actual PSSA results.

While findings show a low correlation between predicted categorical classifications and actual results, the correlation provided information useful to school administrators and teachers. While viewing the total school population the researcher found that while the kappa relationship between the third 4Sight Math test and the PSSA Math test only showed an agreement of 0.093, only four students out of 688 scored in a lower category on the PSSA exam than was predicted by the 4Sight test. When comparing the fourth 4Sight Math test to the PSSA Math exam results, a higher correlation of 0.712 exists. However, 57 students out of 688 scored in a lower category than predicted by the 4Sight test. While the argument with over-testing still exists in public schools, it is valuable for administrators and teachers to note that useful predictions can be made in January (or whenever the 3rd 4Sight test is administered). While the results of the Kappa analysis show that categorical predictions were not perfect or even substantial, in most cases the test did show the lowest reporting category for student results on the PSSA exam. This information will allow administrators and teachers to focus last minute interventions around students who are predicted to score at the high Basic level during the 3rd 4Sight exam.

When viewing the entire school population, both the third (0.201) and fourth (0.250) 4Sight Reading test had a fair correlation to the actual categorical classification on the PSSA exam. When comparing the third and the fourth 4Sight to the actual PSSA exam results, 61 out of 688 students scored in a lower category than the third 4Sight test

predicted and 39 out of 688 scored in a lower category than what the fourth 4Sight test predicted.

In both the Reading and the Math exams, a higher kappa correlation exists between the fourth 4Sight test and the PSSA exam than in the third. However, it is hard to argue that the time spent testing the last round is worth the loss of class instruction. Much of the argument around benchmark assessments is based in the theory that students spend too much time testing. The results show that schools can have a clear idea of where their students will categorically classify on the PSSA exam from the 3rd 4Sight test in January. Administrators and teachers can expect students to score in that range or better on the April PSSA exam. These results make it hard to argue that administering the fourth round of 4Sights tests would be worth the loss of class instruction time. While the *Success for All Foundation* recommends the last test as a way to finish the blueprint for curriculum and achievement analysis of the school year, the reality is that most schools are only concerned with the number of students who will score proficient or above on the PSSA exam. The researcher recommends that the PSSA exam be used in place of the last round of 4Sight to gauge student achievement, identify weaknesses in the curriculum, and highlight scoring trends.

Chapter 5

Conclusion

5.1 Introduction

This study was conducted in a suburban middle school where data was shared with teachers who then used the data to inform instructional practice. This communication and feedback among administrators and practitioners proved the most effective way to create a result focused plan. It is clear to the researcher that the same study conducted without the intended implementation of the 4Sight test may not show the same outcome. This chapter will serve to summarize the major findings of the study and make recommendations for practitioners currently using 4Sight tests in public schools. A best practices model will be outlined to highlight steps taken by the administrators and teachers of Western Middle School to effectively use the benchmark exams. Since the main goal of the Western School District was to make the 2008 AYP requirements, the researcher will also share the final PSSA results of the Western Middle School. In closing, ideas for future research will be highlighted and concluding remarks will be shared.

5.2 Overview of the Findings

The Western Middle School began the 2007-2008 school year on the Pennsylvania state "warning list" for failing to achieve proficiency on the PSSA exams in all sub-group populations. The administration began the year with an immediate sense of urgency knowing that if they failed to make the proficiency targets of 56% in Math and 63% in Reading, it would mean they would fall into the AYP category of School Improvement I.

The researcher's goal was to create a plan to successfully use the 4Sight benchmark assessment as part of an overall plan to make AYP. During the design of the results-focused plan, questions arose as to the validity of the 4Sight test. The main research questions of this study served to challenge the *Success for All Foundation's* claim that the 4Sight test can accurately predict students' results on the PSSA exam. This study was also designed to reveal if the time spent using the four rounds of 4Sight testing was worth the loss of instructional time in the classroom.

The Western School District purchased the 4Sight test as a data collection tool to inform instructional practice and increase student performance on the PSSA exam. While some educators view the exam as another example of the over-testing of students, others view it as a valuable tool to gain insight about students that other wise would be unknown. The review of literature revealed supporters for both sides of this argument so this study is pivotal in that it provided school administrators and teachers with a real-life example of student based interventions in a Western Pennsylvania school district

This study set out to confirm or dispute the allegations made by the *Success for All Foundation* that the 4Sight test can accurately predict student's score on the PSSA exam. The researcher used two main research questions to challenge the claim:

1. Do the 4Sight exams accurately predict student's raw scores on the PSSA exam?
2. Is there a high correlation between the predicted categorical classifications of students on the 4Sight tests when compared to actual PSSA Results?
3. Is the information gained from the fourth 4Sight test worth the loss of instructional time?

Findings in relation to the first research question showed a strong correlation between predicted PSSA raw scores from both rounds of 4Sight testing and actual raw score

results. Surprisingly, a higher correlation existed between the third round of testing in January than between the 4Sight test given in April close to the PSSA exam. The findings clearly support the claim made by the *Success for All Foundation* that the 4Sight test can accurately predict PSSA raw score results with a high level of correlation. The findings also show that the third 4Sight test administered near January showed a high enough correlation for the researcher to now suggest that the Western Middle School eliminate the use of the last round of 4Sight tests after the PSSA exam. The researcher feels that the PSSA results can be used as the last measure of student's progress. In essence, the PSSA exam can finish the blueprint of growth that the three rounds of 4Sight tests begin to draw.

Findings in relation to the second research question were not so apparent. While the correlation between the categorical classifications predicted by the 4Sight exam did not have a strong correlation to actual PSSA results, useful information still emerged. The third round of 4Sight testing elicited predictions that correctly showed where the majority of students will score or improve on their score on the PSSA exam. This information is valuable to administrators and teachers who are looking to identify students at-risk for not passing the PSSA exam. The 4Sight testing can highlight students most susceptible to last minute interventions aimed at moving them into a higher categorical classification. Similar to the results found with the raw score data, predicted categorical classifications from the January 4Sight test showed enough data to suggest not using the last round of 4Sight testing (in addition to the PSSA exam). Since this finding emerged from both data analysis it will be the researcher's recommendation to the Superintendent of the Western School District to stop using the last round of 4Sight testing. Three rounds of 4Sight

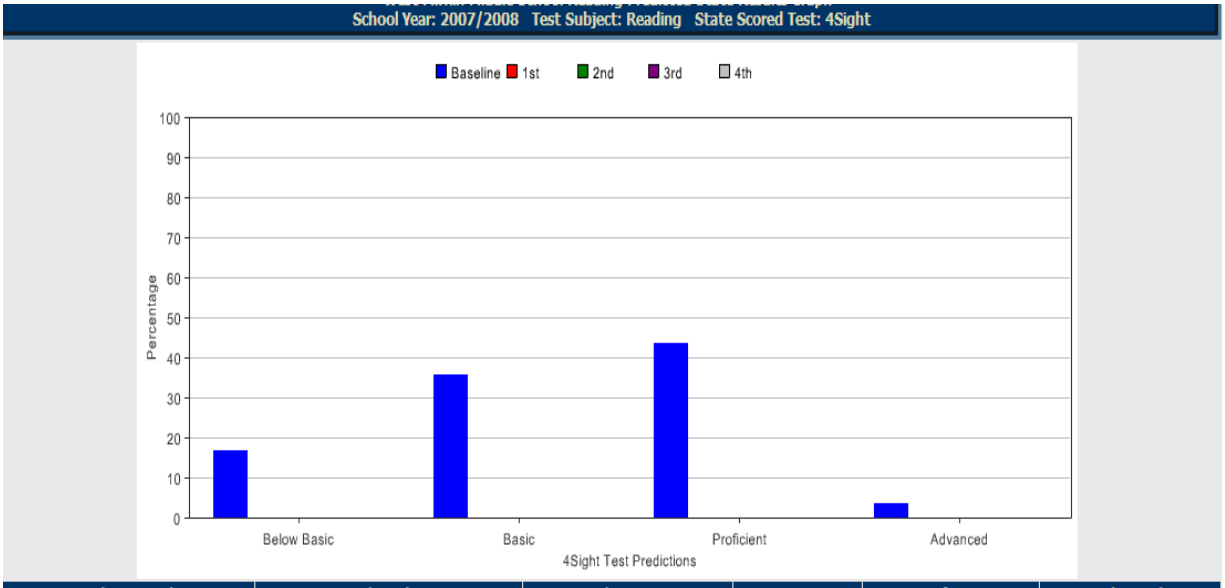
testing combined with the PSSA exam provide a wealth of data to examine curriculum and instructional practice.

5.3 Recommendations for Practice

To prepare staff members for the first administration of the 4Sight testing, the researcher discussed the assessment tool on the first in-service day of the 2007-2008 school year (August 23, 2007). The information provided by the *Success for All Foundation* was reviewed with teachers to highlight the importance of the testing and the valuable insight that can be gained. The researcher knew that he had to show a belief in the testing and an enthusiasm for its use to have full staff buy-in to the process.

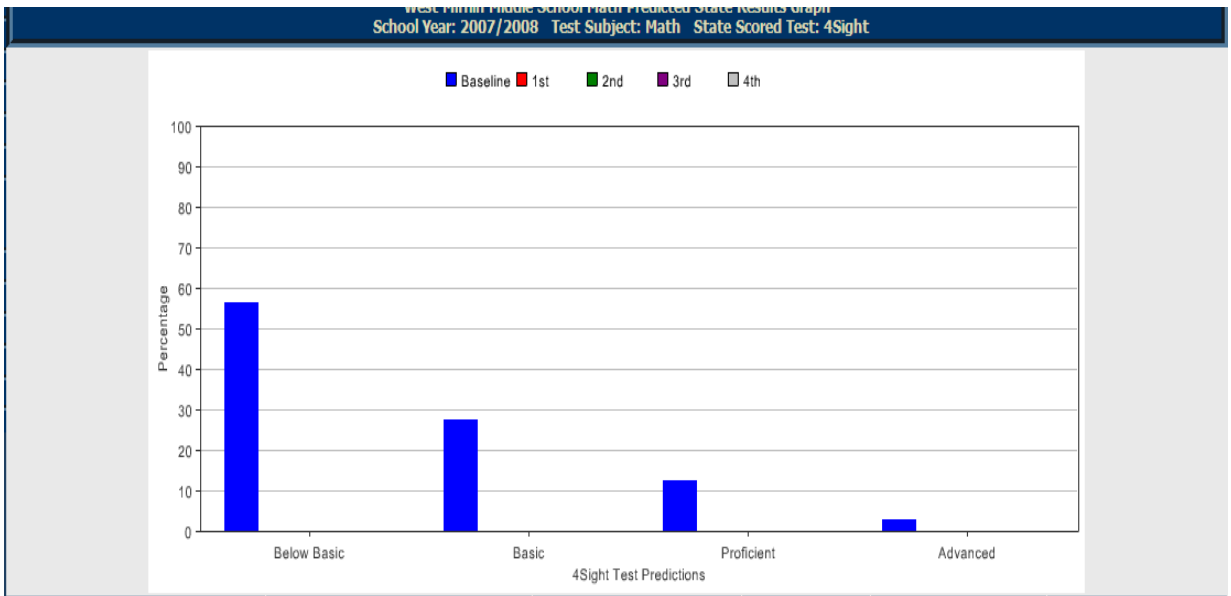
To begin to outline a plan for intervention strategies, teachers were given a list of the updated Pennsylvania Assessment Anchors for Reading and Math to review. The researcher first wanted to know what anchors the curriculum was currently covering and what areas are being overlapped or neglected. It was also important to note which anchors may be covered in September and then never reviewed again before the PSSA exam. This leg work at the beginning of the year helped to build a foundation where teacher support for the project grew. Teachers began to comment that they know what they are covering but they do not know (on a large scale covering all anchors) how students are progressing toward mastery. These questions led perfectly into the first round of 4Sight tests in September 2007. After the testing was complete the researcher shared the following two graphs with staff members:

4Sight Reading Baseline Scores from September 11, 2007



Grading Period	Total Students	Below Basic	Basic	Proficient	Advanced
Baseline	716	16.90 %	35.75 %	43.72 %	3.63 %
1st					

4Sight Math Baseline Scores from September 11, 2007



Grading Period	Total Students	Below Basic	Basic	Proficient	Advanced
Baseline	718	56.69 %	27.72 %	12.53 %	3.06 %
1st					

These charts were retrieved from the Western Middle School's Member Center Web Site (https://members.successforall.net) that was provided by and is managed by the *Success for All Foundation*. The member center site allows users to manage student profiles, create reports and charts and lists students who performed below basic, basic, proficient and advanced for each reporting category of the PSSA exam. According to the *Success for All Foundation* the results shown in the above charts reflect how students would have scored on the PSSA test if it had been given on the same day. When the researcher reviewed the results with each of the departments of teacher in the Western Middle School, an immediate panic ensued. According to the first 4Sight test scores, only 46% of the total population of students (all grade levels combined) scored at the proficient level in Reading and only 15% of students scored at the proficient level in Math. If the predictions made by the *Success for All Foundation* were accurate, the teachers and administration of the Western Middle School clearly had a lot of work to do. Somehow, a plan had to be in place to move students to meet the AYP targets of 63% in Reading and 56% in Math before the PSSA exam that will be administered on April 1, 2008.

After scores were gathered and reported, here are the measures that were put into place to inform instructional practice and raise student achievement:

1. Near the third week of September 2007, all teachers reviewed the assessment anchors sheets and completed a chart showing the month each anchor was taught or reviewed in their classroom instruction
2. Monthly department meetings were held to review data and identify trends across students. Also, anchor sheets were again reviewed to identify

curriculum gaps and identify areas that teachers would not have time to cover. These anchors (not covered in regular subject areas) were then shared with all special area teachers (music, gym, art, computers, etc.) so they could work these areas into their curriculums.

3. Immediately following each 4Sight test, item analysis breakdowns for each question (and corresponding anchor) were provided for all teachers. This item analysis (available from the members center web site) showed individual questions and answers for the test and the percentage of students who selected each answer option. Teachers reviewed this information with students and often used exact questions as challenge or warm-up questions. For example, why would 88% of all seventh graders choose this answer? This showed students that the teachers and administrators valued the test as an importance piece of classroom instruction.

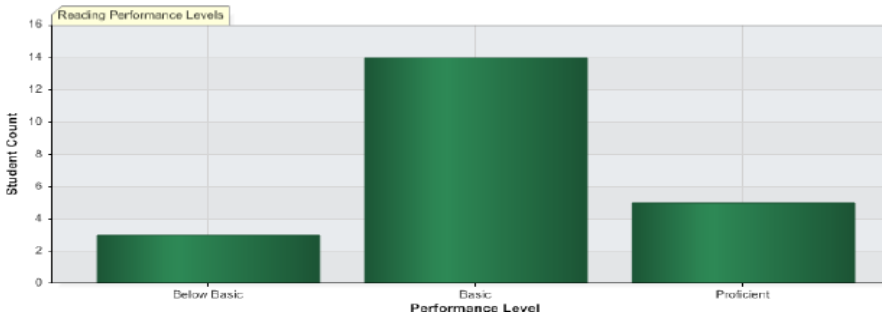
4. All data was uploaded into EdInsight software so teachers could view strengths and weaknesses of students in all of their classes. An example of what teachers were provided access to on their personal computers is shown here:

Subject Scores

ScoreType	Score Test Per 0	Gains
Read Score 4Sight	20.36	
Read Raw Score	20.36	
Read Pct Score	67.88	
Math Score 4Sight	13.77	
Math Raw Score	13.77	
Math Pct Score	38.26	

Reading

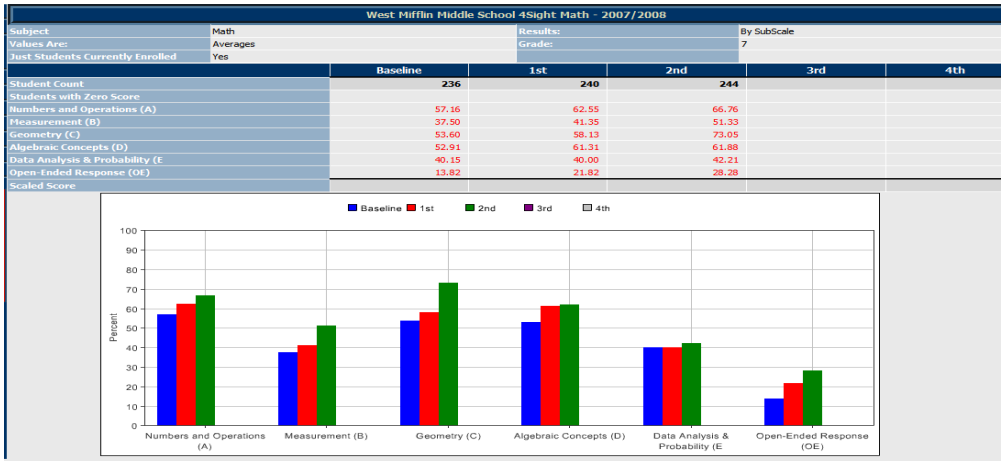
Performance Level	Baseline	Overall Gains
Below Basic	3	-3
Basic	14	-14
Proficient	5	-5
Total Students	22	



5. Flexible groups were created within all subject to highlight students with common areas of concern. Student scores individually, by class, and by grade level were charted by reporting category to identify common areas of need. Reporting category information is provided by the *Success for All Foundation* through their member's center website (customized for the Western Middle School). This example (shown below) that was provided to teachers shows grade seven Math results after the third (two tests and one baseline) test. Teachers then used

this information to create the flexible grouping within their regular classroom. This information allows teachers to group students with common needs for enrichment and supplemental activities. Most importantly, with this amount of information, teachers can more easily embed these skills and knowledge in the larger curriculum. This way classroom curriculum was enhanced rather than narrowed and teachers and administrators can argue that they are not teaching to the test; they are teaching to the needs of our students.

Retrieved from https://members.successforall.net/ on January 28, 2008.



West Mifflin Middle School 4Sight Math - 2007/2008

Subject	Math	Results:	By SubScale
Values Are:	Averages	Grade:	7
Just Students Currently Enrolled	Yes		

	Baseline	1st	2nd	3rd	4th
Student Count	236	240	244		
Students with Zero Score					
Numbers and Operations (A)	57.16	62.55	66.76		
Measurement (B)	37.50	41.35	51.33		
Geometry (C)	53.60	58.13	73.05		
Algebraic Concepts (D)	52.91	61.31	61.88		
Data Analysis & Probability (E	40.15	40.00	42.21		
Open-Ended Response (OE)	13.82	21.82	28.28		
Scaled Score					

On February 5, 2008 the researcher emailed a copy of the above information (in the form of a detailed PowerPoint) to Dr. Nancy Madden of the *Success for all Foundation*. At this point in the school year, three rounds of 4Sight tests had been administered to all students in grades 6, 7, and 8. The goal was to be certain that the researcher and the middle school staff are using the 4Sight tests correctly and more importantly, how the designers of the test intended. After reviewing the information provided, Dr. Madden responded, “It looks as if your district has taken a very productive

approach to the use of 4Sight." The original email to Dr. Robert Slavin, creator of the 4Sight assessments was directed to Dr. Madden (Appendix C).

5.4 2008 Western Middle School Final PSSA Results

The researcher feels that it is important to report the significant increases in the 2008 PSSA results which followed a year of the intensive interventions outlined above. The Western School District purchased the 4Sight test as an improved data collection tool to inform instructional practice and provide insight to teaching interventions. The assessment was purchased as a tool to assist the district in attaining the goal of passing the PSSA exam and getting the middle school off of the "Warning" list. The results that follow show the increase in student PSSA scores from the 2007 PSSA exam to the 2008 PSSA exam. It is important to note that the 4Sight testing was used for the first time in the Western Middle School during the 2007-2008 school year. The results show dramatic increases that the researcher credits in part to the information gained from the 4Sight benchmark test.

Table 5.1 2008 Western Middle School Final PSSA Results Including Sub-groups

PSSA Math	2007	2008	Net Increase
All Students	56.8	78.8	+22
White	62.5	82	+19.5
Black	32.6	64.7	+32.1
IEP	26.9	38.5	+11.6
Econ.	45.3	72.9	+27.6

PSSA Reading	2007	2008	Net Increase
All Students	62.6	72.1	+9.5
White	67.8	76.4	+8.6
Black	40.7	54.4	+13.7
IEP	21.3	34.1	+12.8
Econ.	48.6	63.4	+14.8

The categories in the left column first show all students and then each subgroup within the Western Middle School. The numbers next to each student group show the percentage number of students who scored above the proficiency level on the PSSA exam. The final PSSA results from 2008 when compared to results in 2007 show student achievement increases in all subgroups and in the overall population of students in grades 6, 7, and 8 combined. The largest increase in student performance was in the Math test with the Black student subgroup population. Students in this subgroup showed an over 32% increase in proficiency. While students in the IEP subgroup in Math and in the Black and IEP subgroups in Reading did not meet the AYP targets of 56% proficiency in Math and 63% proficiency in Reading, they did make AYP because of the Safe Harbor provision. Safe Harbor allows districts to make AYP if they show a 10% increase in the number of students achieving proficiency in a particular subgroup from the previous year (Pennsylvania Department of Education, 2007). The results above show that each subgroup comfortably scored above the required 10% increase. As a result, the Western Middle School made AYP by all standards outlined by the Pennsylvania Department of Education and was removed from the "Warning" status. The researcher credits the 4Sight assessment as one of the main components of the districts plan to get immediate results. The testing combined with effective communication and ongoing curriculum review provided the results that met the expectations of both the district and the state department of education. From a leadership standpoint, the researcher feels that the results do not show the administrators ability to lead; but instead the administrators ability to listen to staff recommendations and the needs of the students. The 4Sight tests identified trends and weaknesses in student learning and the researcher presented this

information to all teachers. Collectively a plan was made to supplement curriculum, re-teach areas of concern, and attack student needs. The combination of an effective testing regimen and use of data by instructors to inform classroom instruction created the environment conducive to increasing student achievement.

5.5 Other Benefits of Using the 4Sight Exam

As the adminstrator in charge of organizing the testing, anlayzing the results, and disaggregating data for teachers, the researcher also realized unexpected benefits of using the 4Sight exam. These included:

- Practice run for the PSSA testing. Beneficial for both students and staff. Reminders went out to parents for students to get a good nights rest, eat breakfast, show up to school on time, etc.
- Students get used to “testing room”. By the second round of testing, it was a very smooth transition when students were told to report to their testing rooms.
- Scheduling issues such as (what do we do with teachers who do not have homerooms?) were worked out. It was decided by the researcher to use extra/free teachers to create small group testing rooms across the entire middle school. This design was used during all rounds of 4Sight testing and also with the PSSA exam.
- Do we have enough pencils, paper and calculators? During the first rounds of testing, the researcher realized that while preperation was thought to be complete, some teachers were scrambling at the last minute to round up

pencils and calculators. This was worked out for the second round of 4Sight testing and maintained consistantly through the PSSA exam.

- The researcher observed students to be much less nervous for the 2nd round of the 4Sight tests and more comfortable with the adjusted schedule. By the time of the PSSA exam, this testing schedule became routine. The researched used a 2-hour delay bell schedule so that a consistent time frame for the testing of the entire building was set. This 2-hour delay schedule was used for all four rounds of 4Sight tests and the PSSA exams.
- Found/identified building level trends in open-ended responses. This opened communication between departments to discuss how Social Studies and Science teachers could begin to evaluate open ended responses the same way the Language Arts and Reading teachers were using in their classrooms. Teachers began using common language between subjects areas to teach students correct writing skills. For example, students often only used one supporting detail when the question asked for two.
- Created incentives for attendance and participation (grade level challenges). This rewards system added to the culture that valued testing for its positive rewards. The researcher committed the school to celebrating the success and progress of each grade level after every round of testing. The researcher also feels that it is essential for building administrators to share the results with the students.
- Better use of the PSSA Coach books materials. While not every benchmark or skill covered on the PSSA exam is assessed on every 4Sight test,

weaknesses in reporting categories is shown. The researcher then instructed teachers to look to the PSSA coach books for lessons across the reporting category. Since the PSSA coach books are divided into sections by reporting category, it was easy for teachers to use these short lessons as warm-up exercises in their regular instructional plan. This ensured that teachers were covering an array of skills that can be tested on the PSSA and not limiting students to only skills assessed on the 4Sight test.

5.6 Essentials Elements of Effective use of Benchmark Assessments

Based upon the findings of this study and the review of literature, the following recommendations are made for districts to effectively use the 4Sight benchmark assesment:

- Districts can benefit from the use of teacher-friendly data use systems like EdInsight software.
- Districts can benefit from creating time for teachers to meet collectively to disucss data findings.
- Districts can benefit by having an adminisrator assigned to analyzing testing data, disaggregating results, and sharing analysis with staff.
- Districts can benefit from creating a school culture that values testing as a guide to inform instruction and not to replace instruction.
- Districts can benefit from creating an alternative testing schedule that stays consistant through all testing.

- Districts can benefit by emphasizing the use of testing data to reflect the effectiveness of interventions and encourage changes based on best-practices.

5.7 Recommendations for Future Research

Based upon the findings of this study and the review of literature, the following recommendations are made for future research:

- Investigate a longitudinal study that follows the use of the 4Sight test with a cohort of students over several years.
- Identify a model of best practices to use when implementing 4Sight assessments.
- Compare specific classroom interventions used in conjunction with benchmark assessment with student achievement growth.
- Expand this study to include all students in grades 3-8 and 11.
- Narrow the focus of the study to a particular sub-group or groups.
- Analyze specific skills assessed on the PSSA and compare them with skills tested on the 4Sight exam.

5.8 Concluding Remarks

This process began with an issue of validity and an interest in finding the answer to a question that many asked but few answered. Administrators knew that the 4Sight test was gaining popularity and produced predicted PSSA results but no one ever followed up on its' predictions. Educational leaders seem to ride the wave of new trends and quickly spend money on anything believed to raise test scores. When the Pennsylvania Department of Education supports a testing regimen (like the 4Sight test) and approves

grant monies to be used in its purchase, few school administrators have the time to challenge any of the accusations of its effectiveness. This study served to test claims, examine effectiveness, and uncover best practices of the 4Sight benchmark test. Through the process, the researcher found the 4Sight tests to be an effective resource to add to a school's overall testing plan. The researcher believes that the testing alone is not the answer to the question of how to raise test scores. The 4Sight testing combined with teacher collaboration, student enthusiasm, and administrative leadership can result in the creation of an effective plan to quickly raise PSSA results. All four components (testing plan, collaboration, enthusiasm, leadership) equally accounted for the success seen by the Western Middle School during the 2007-2008 school year. No one quality could have be removed or neglected. The plan encompassed a united staff with dedicated students and the researcher can proudly say the results speak for themselves.

Appendix A

2008 PSSA Math and Reading Performance Level Cut Scores

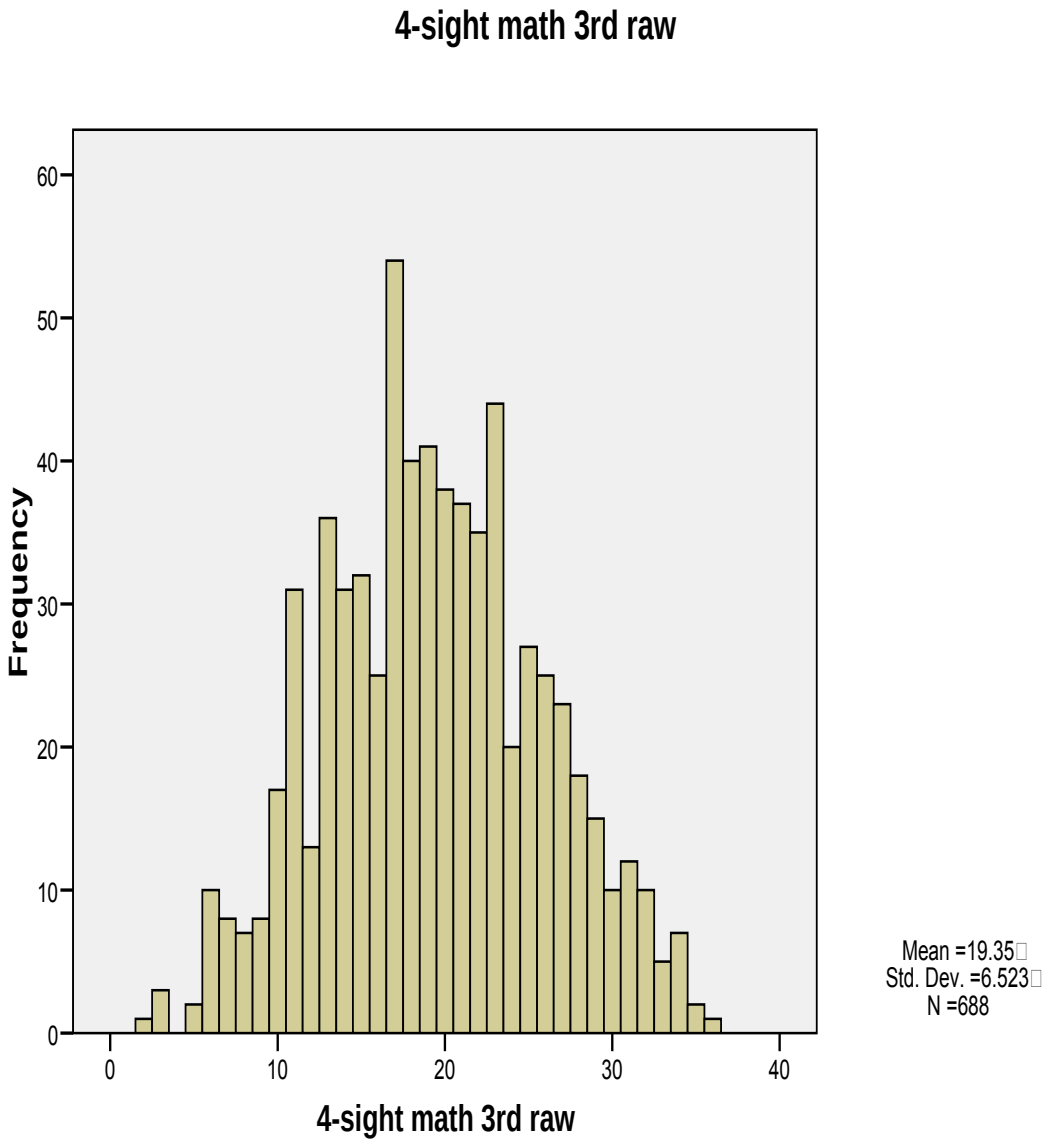
	Grade 6		Grade 7		Grade 8	
MATH	low	high	Low	high	low	high
Advanced	1476	and up	1472	and up	1446	and up
Proficient	1298	1475	1298	1471	1284	1445
Basic	1174	1297	1183	1297	1171	1283
High Basic	1236	1297	1240	1297	1227	1283
Low Basic	1174	1235	1183	1239	1171	1226
Below Basic	700	1173	700	1182	700	1170
High Below Basic	937	1173	941	1182	935	1170
Low Below Basic	700	936	700	940	700	934

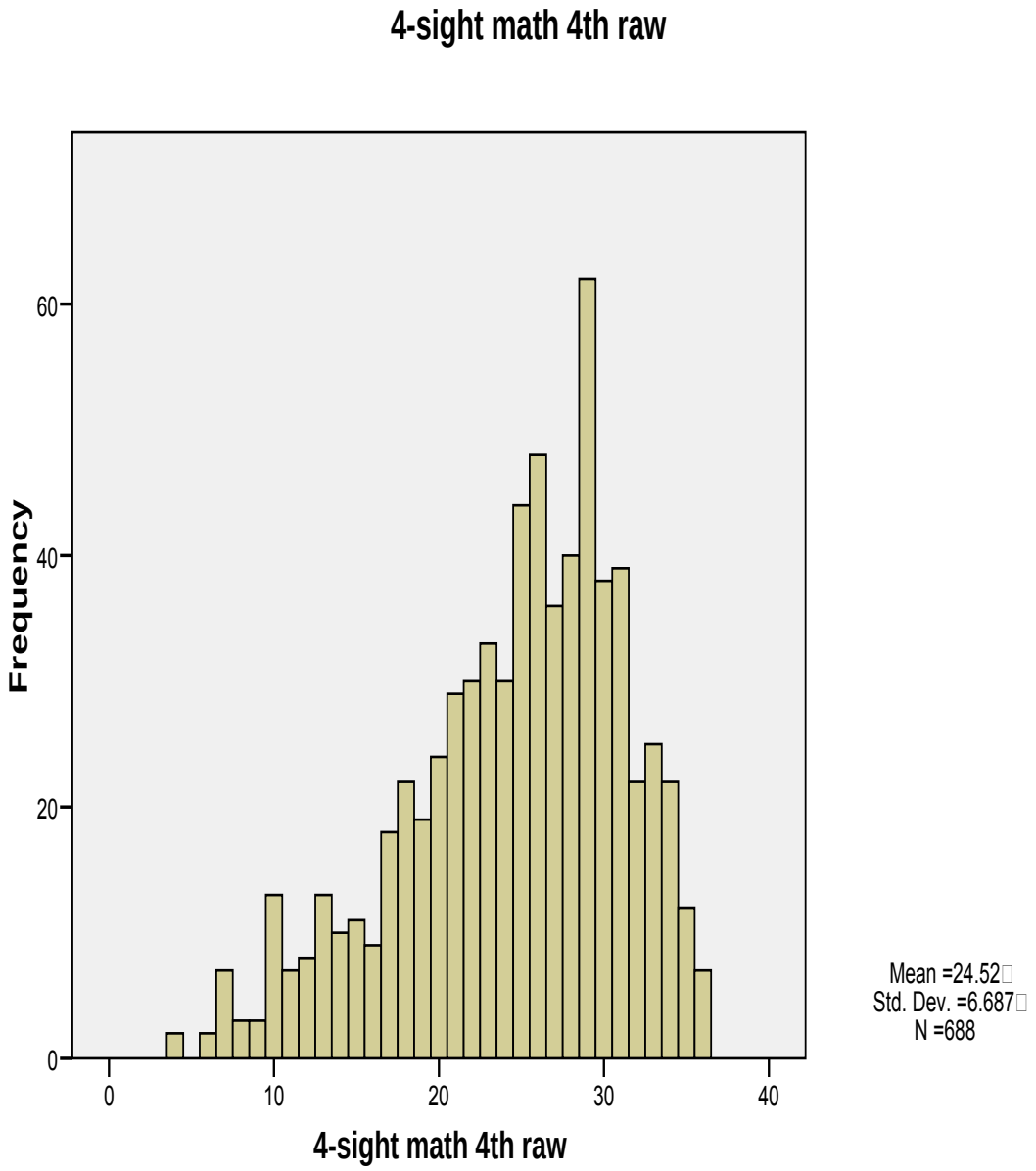
	Grade 6		Grade 7		Grade 8	
READING	low	high	Low	high	low	high
Advanced	1456	and up	1470	and up	1473	and up
Proficient	1278	1455	1279	1469	1280	1472
Basic	1121	1277	1131	1278	1146	1279
High Basic	1199	1277	1205	1278	1213	1279
Low Basic	1121	1198	1131	1204	1146	1212
Below Basic	700	1120	700	1130	700	1145
High Below Basic	910	1120	915	1130	923	1145
Low Below Basic	700	909	700	914	700	922

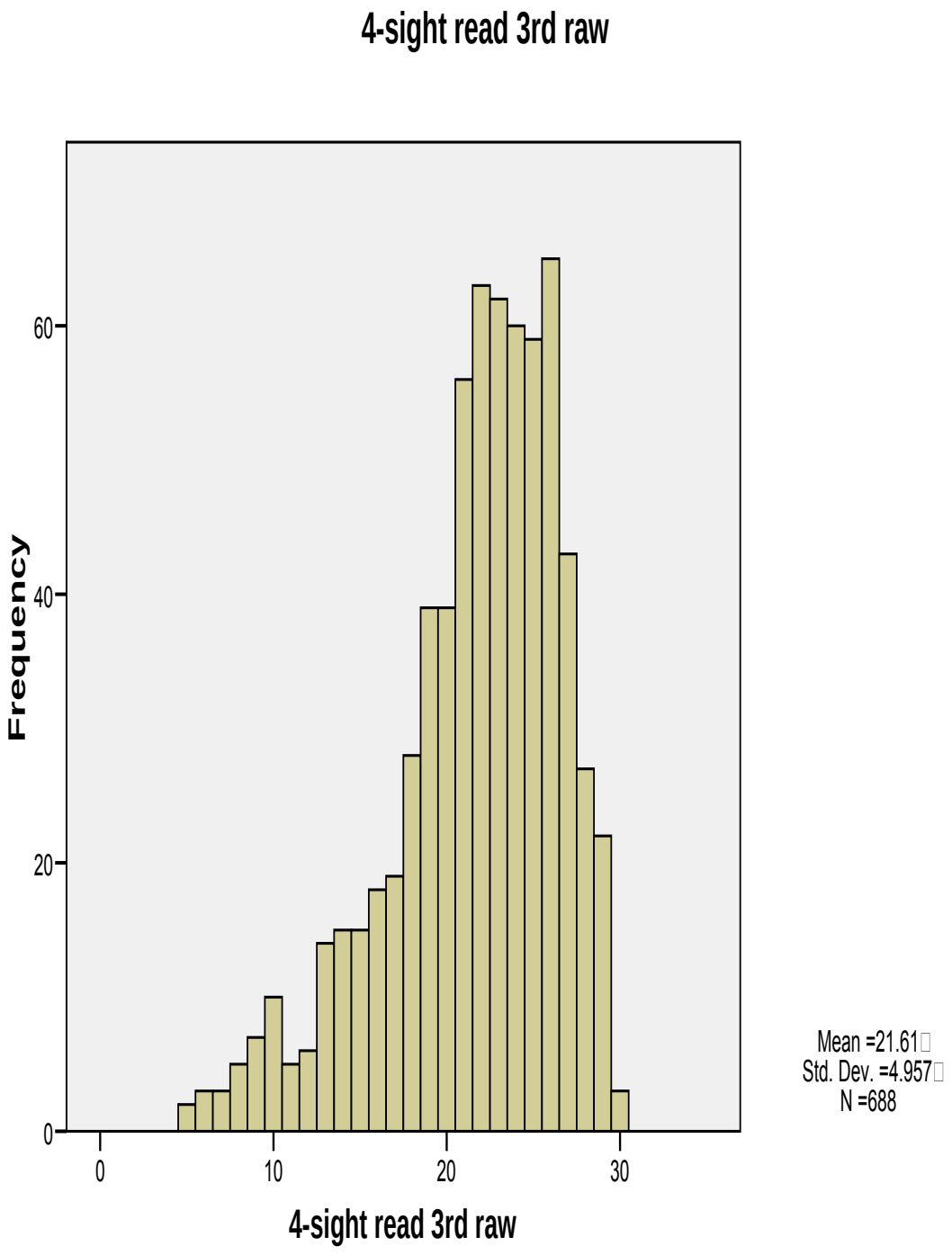
Pennsylvania Department of Education (2008)

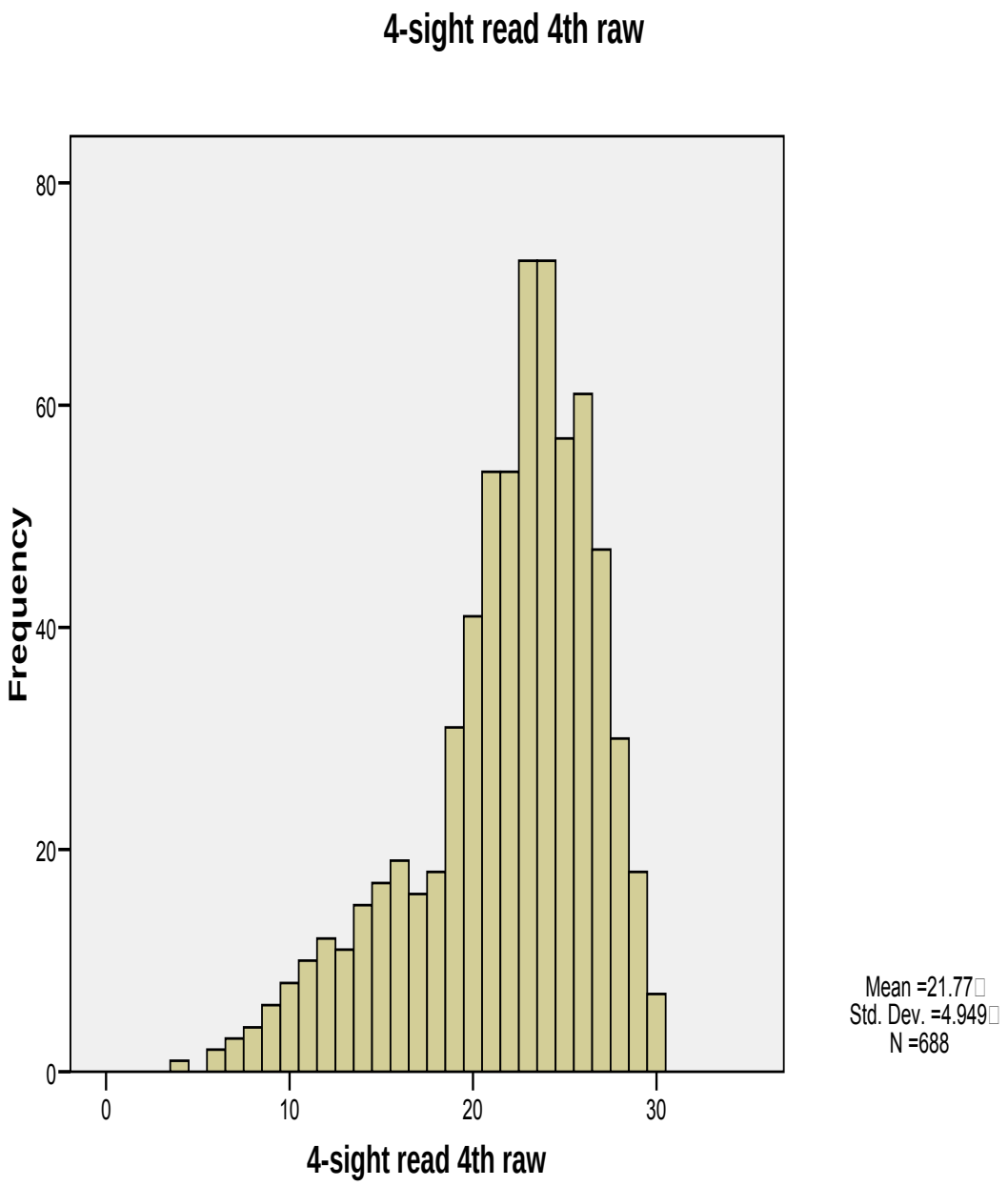
Retrieved from: http://www.pde.state.pa.us/a_and_t/lib/a_and_t/Cut_Scores_07.xls July 20, 2008.

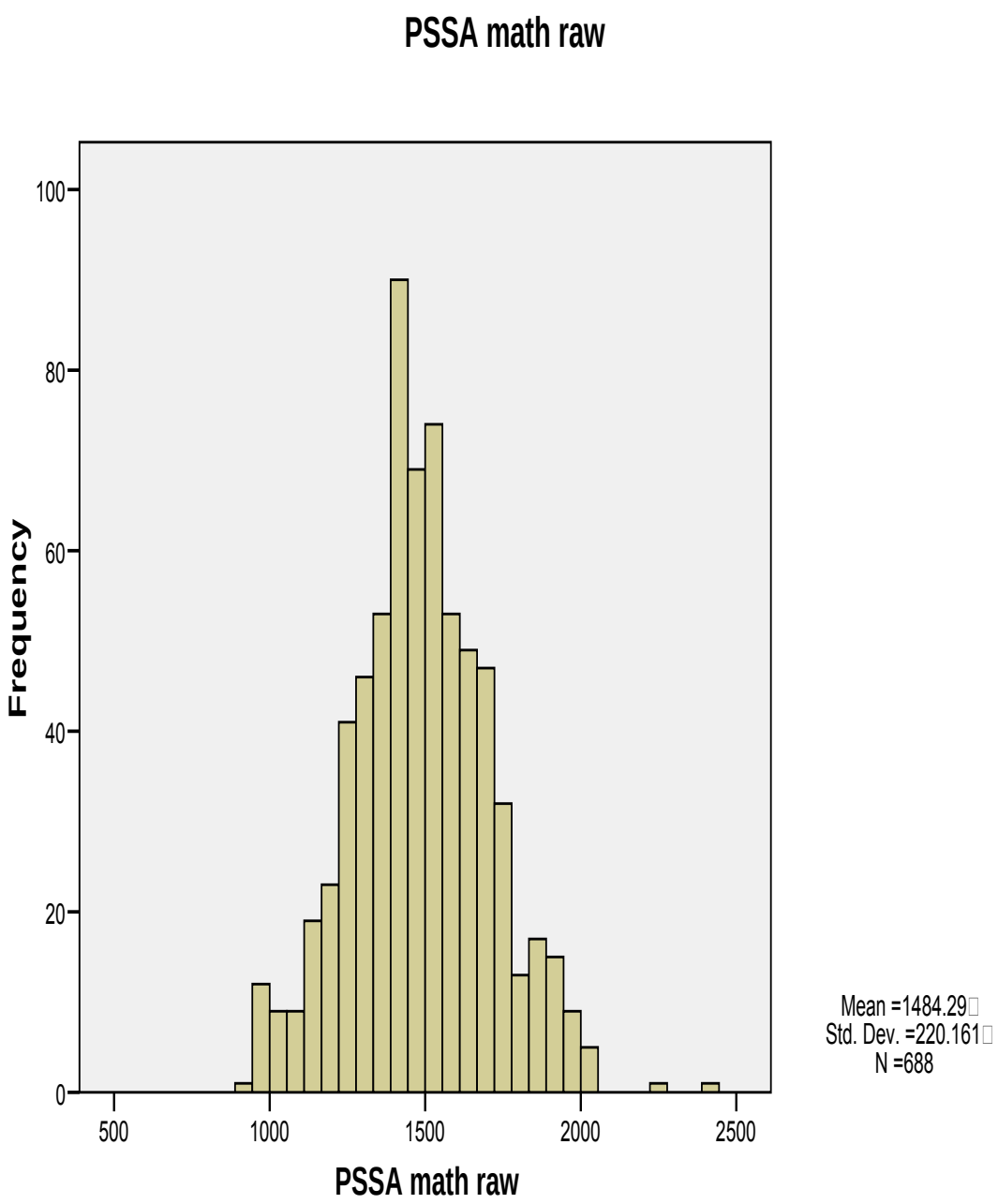
Appendix B

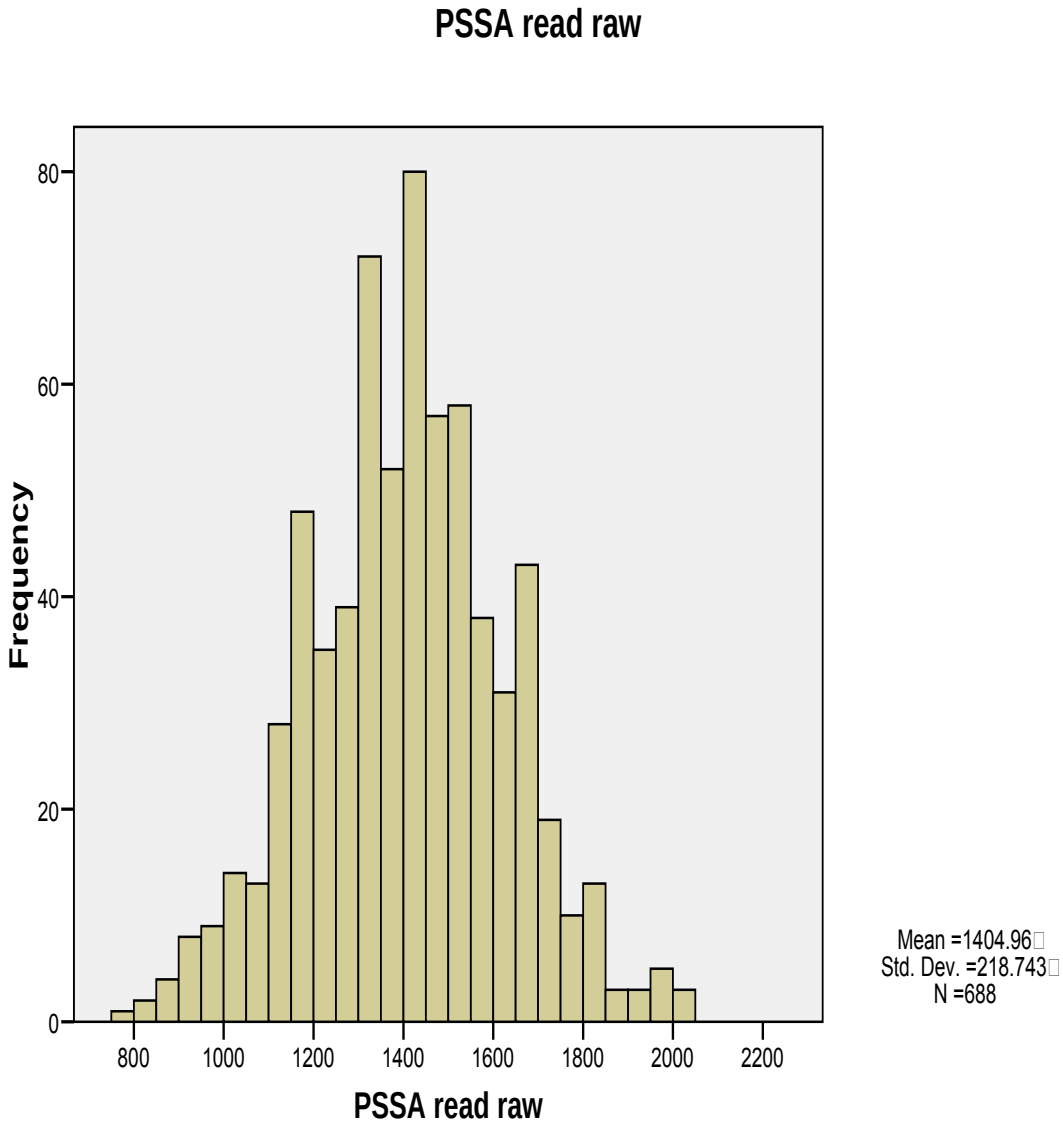

4-sight math 3rd raw
Frequency
60
50
40
30
20
10
0
0
10
20
30
40
4-sight math 3rd raw
Mean =19.35
Std. Dev. =6.523
N =688


4-sight math 4th raw
Frequency
60
40
20
0
0
10
20
30
40
4-sight math 4th raw
Mean =24.52
Std. Dev. =6.687
N =688

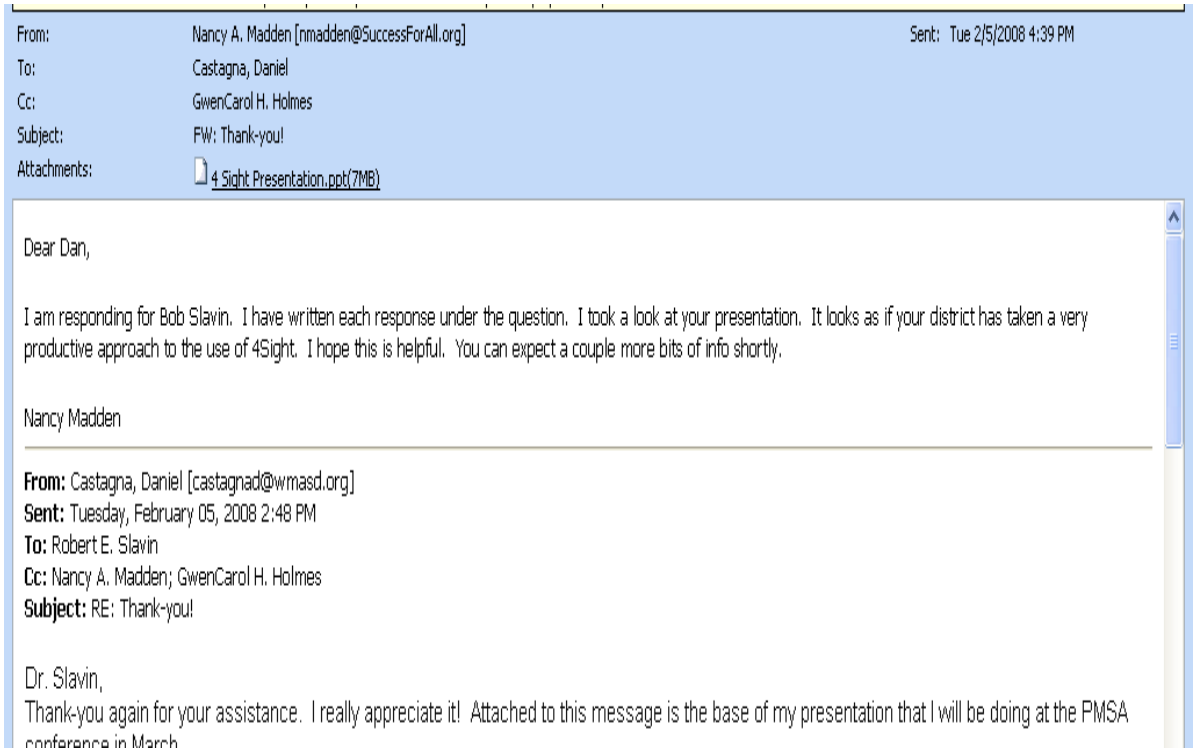

4-sight read 3rd raw
Frequency
60
40
20
0
0
10
20
30
4-sight read 3rd raw
Mean =21.61
Std. Dev. =4.957
N =688


4-sight read 4th raw
Frequency
80
60
40
20
0
0
10
20
30
4-sight read 4th raw
Mean =21.77
Std. Dev. =4.949
N =688


PSSA math raw
Frequency
100
80
60
40
20
0
500
1000
1500
2000
2500
PSSA math raw
Mean =1484.29
Std. Dev. =220.161
N =688


PSSA read raw
Frequency
80
60
40
20
0
800
1000
1200
1400
1600
1800
2000
2200
PSSA read raw
Mean =1404.96
Std. Dev. =218.743
N =688

Appendix C



From: Nancy A. Madden [nmadden@SuccessForAll.org] Sent: Tue 2/5/2008 4:39 PM
To: Castagna, Daniel
Cc: GwenCarol H. Holmes
Subject: FW: Thank-you!
Attachments: 4 Sight Presentation.ppt(7MB)

Dear Dan,

I am responding for Bob Slavin. I have written each response under the question. I took a look at your presentation. It looks as if your district has taken a very productive approach to the use of 4Sight. I hope this is helpful. You can expect a couple more bits of info shortly.

Nancy Madden

From: Castagna, Daniel [castagnad@wmasd.org]
Sent: Tuesday, February 05, 2008 2:48 PM
To: Robert E. Slavin
Cc: Nancy A. Madden; GwenCarol H. Holmes
Subject: RE: Thank-you!

Dr. Slavin,
Thank-you again for your assistance. I really appreciate it! Attached to this message is the base of my presentation that I will be doing at the PMSA conference in March

Bibliography

Aronson, D. (2007). *Formative assessment: helping students grow.* The National Council of Teachers of English. Retrieved February 9, 2008, from http://www.ncte.org/pubs/chron/highlights/126802.htm

Association for Early Childhood International. (2001). *AECI position paper on standardized testing* . Olney, MD: Author.

Atkinson, R. (2001). Achievement versus aptitude tests in college admissions. Retrieved April 17, 2008 http://www.ucop.edu/pres/speeches/achieve.htm

Berliner, D. C. & Biddle, B. J. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. Reading, MA: Addison-Wesley Publishing Company, Inc

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2002). *Working inside the black box: Assessment for learning in the classroom.* London, UK: King's College London Department of Education and Professional Studies.

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

Black, P. & William, D. (1996). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment. *British Educational Research Journal, 22*, 537-548.

Bloomfield, D. & Cooper, B. (2003). NCLB: A New Role for the Federal Government An Overview of the Most Sweeping Federal Education Law since 1965. *THE Journal (Technological Horizons in Education), 30.* Retrieved March 22, 2008, from http://thejournal.com/articles/16365

Boston, Carol (2002). The concept of formative assessment. *Practical Assessment, Research & Evaluation, 8*(9). Retrieved March 9, 2008, from http://PAREonline.net/getvn.asp?v=8&n=9

Bowling, B. (2007, September 13). Wilkinsburg, West Mifflin score big on scores. *Pittsburgh Tribune Review*. Retrieved November 21, 2007, from http://www.pittsburghlive.com/x/pittsburghtrib/news/today/s_527079.html

Bror, S. (2006, November 17). Learning and preparing for state tests – in conflict? Message posted to https://communitychest.k12.com/node/761

Campbell, D. T. (1975). Assessing the impact of planned social change. *Social research and public policies: the Dartmouth/OECD Conference*. Hanover, NH: Dartmouth College.

Clarke, M., Madaus, G., Horn, C., & Ramos, M. (2001). The marketplace for educational testing. *Statements 2*(3). Retrieved April 17, 2008 from http://www.bc.edu/research/nbetpp/publications/v2n3.html

College Board (1977). On further examination: report of the advisory panel on the scholastic aptitude test score decline. New York: College Board.

Collett, P., Gyles, N., & Hrasky, S. (2007). Optional formative assessment and class attendance: their impact on student performance. *Global perspectives on accounting education, 4,* 41-59.

Cimbricz, S. (2002). State-mandated testing and teachers' beliefs and practice. *Education Policy Analysis Archives, 10*(2). Retrieved February 22, 2008, from http://epaa.asu.edu/epaa/v10n2.html

Darling-Hammond, L. (2007). Evaluating no child left behind. *The Nation.* Retrieved February 5, 2008, from http://www.thenation.com/doc/20070521/darling-hammond

Dempster, F. N. (1991). Synthesis of research on reviews and tests. *Educational Leadership, 72*(8), 71–76.

Di Eugenio, B. (2000). On the usage of Kappa to evaluate agreement on coding tasks. In LREC2000: Proceedings of the Second International Conference on Language Resources and Evaluation, pages 441-444, Athens.

Edinformatics. (2007) *State tests preparation.* Retrieved October 2, 2007, from http://www.edinformatics.com/testing/testing2.htm

Education Law Center (2002). *Is your child's school making "adequate yearly progress" (AYP)?* Retrieved May 2, 2008 from http://www.elc-pa.org/pubs/downloads/english/NCLBpubs/ncb-is-your-childs-school-making-the-ayp.pdf

Education Week (2004). *Assessment*. Retrieved April 1, 2008, from http://www.edweek.org/rc/issues/assessment

Educational Testing Service (2008). *ETS reports record 6.2 million TOEFL® and TOEIC® test takers in 2007*. Retrieved April 17, 2008 from http://www.hoovers.com/free/co/news/detail.xhtml?ID=54718&ArticleID=200803282800

Figlio, D. N. & Lucas, M. E. (2000). *What's in a grade? School report cards and house prices.* National Bureau of Economic Research [On-line]. Available: http://papers.nber.org/papers/w8019

Foote, M. (2007). Keeping accountability systems accountable. *Phi Delta Kappan, 88*(5), 359-363.

Frontline (2002). *The testing industry's big four*. Retrieved April 14, 2008 from http://www.pbs.org/wgbh/pages/frontline/shows/schools/testing/companies.html

Gardner, W. (1995). On the reliability of sequential data: measurement, meaning, and correction. *In John M. Gottman (Ed.), The analysis of change.* Mahwah, N.J.: Erlbaum.

Garrison, C. & Ehringhaus, M. (2007). Formative and summative assessments in the classroom. *National Middle School Association.* Retrieved February 29, 2008, from http://www.nmsa.org/Publications/WebExclusive/Assessment/tabid/1120/Default.aspx

Glass, G. V. & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). New York: Allyn & Bacon

Gluckman, A. (2002). Testing…testing…one, two, three: The commercial side of the standardized testing boom. *Dollars & Sense: The Magazine of Economic Justice*. Retrieved April 17, 2008 from http://www.dollarsandsense.org/archives/2002/0102gluckman.html

Gratz, D. (2000). High standards for whom? *Phi Delta Kappan, 81,* 681-687.

Gulek, C. (2003). Preparing for High Stakes Testing. *Theory Into Practice, 42*(1), 42-50.

Hamilton, K. (2005). *Big business: Educational testing is a multimillion-dollar industry, with revenues only expected to increase with NCLB mandated tests.* Retrieved April 17, 2008 from http://findarticles.com/p/articles/mi_m0DXK/is_8_28?ai_n15399787/print

Hanson, F. (1993). *Testing testing: Social consequences of the examined life*. Berkeley: University of California Press

Hart, P. & Teeter, R. (2001). *A measured response: Americans speak on educational reform*. Princeton, NJ: Educational Testing Service. Retrieved ftp://etsis1.ets.org/pub/corp/2001_executive_report.pdf .

Hattie, J. & Jaeger, R. (1998). Assessment and classroom learning: a deductive approach. *Assessment in Education*, *5*(1), 111-112.

Hebert, E. (2007). *ACT vs. SAT – which test should I take and why?* Retrieved April 17, 2008 from http://www.collegedegree.com/library/act-sat

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

Johnston, P. (1995). Assessment of teaching and learning in literature-based classrooms, *Teaching and Teacher Education*, *11*(1), 359.

Jones, J. M. (1996). *The standards movement – past and present*. Milwaukee, WI: PRESS Publication.

Kafer, K. (2004). No child left behind. *World and I*, 19(5), 253-266

Kirst, M. & Rowen, H. (1993). *Why we should replace aptitude tests with achievement tests*. Education Week. Retrieved April 17, 2008 from http://www.edweek.org

Kreitzer, A. E., Madaus, G. F., & Haney, W. (1989). Competency testing and dropouts. In L. Weis, E. Farrar, & H. G. Petrie (Eds.) *Dropouts from school: Issues, dilemmas, and solutions*. Albany, NY: State University of New York Press.

Landis, J. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* (33), 159-174.

Link, H (1919). *Employment Psychology: The application of scientific methods to the selection, training, and grading of employees*. New York: Macmillan.

Lipson, M. & Wixson, K. (2003). *Assessment & instruction of reading and writing difficulty*. Boston: Allyn & Bacon.

Lutz-Doemling, C. (2007) An examination of the Pennsylvania 4sight benchmark assessments as predictors of Pennsylvania system of school assessment performance (Doctoral dissertation, Lehigh University, 2007). Dissertation Abstracts.

Madaus, G. F. (1993). A national testing system: Manna from above? An historical/technological perspective. [Electronic version]. *Educational Assessment*, pp. 9-26.

Manya Group (2005). SAT vs. ACT: How do the tests compare? Retrieved April 17, 2008 from http://www.manyagroup.com/modules.php?name=20a3c_sat_articles_act

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). Making sense of data-driven decision making in education. RAND Occasional Paper Series. Retrieved May 10, 2008 from http://www.rand.org/pubs/occasional_papers/OP170/

McAlpine, M. (2002). *Principles of assessment*. Glasgow: University of Glasgow, Robert Clark Center for Technological Education.

McGraw-Hill (2007). *Acuity Assessment Management System.* Children First. Retrieved October 8, 2007, from http://www.2.ctb.com/acuity/NYCAcuity.shtml#instructionalexercises

McMillan, J. H. (2004). *Classroom assessment: Principles and practice for effective instruction.* Boston, MA: Pearson.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2001). *Introduction to linear regression analysis* (3rd Ed.). New York: John Wiley & Sons.

National Assessment Governing Board (2008). located at http://www.nagb.org/

National Defense Education Act: Public Law 85-864 (1964). *The South Central Bulletin (24)*, pp. 66-76.

New York State Department of Education (2007). *State assessment examination schedules.* Retrieved September 27, 2007, from http://www.emsc.nysed.gov/osa/sched.html

Neill, M (1997). Transforming student assessment. *Phi Delta Kappan,* pp. 35-36.

Ohio Department of Education (2007). *Resources for Ohio's achievement test*. Retrieved September 27, 2007, from http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=222&ContentID=19484&Content=36892

Olson, L. (2005). ETS to enter formative-assessment market at k-12 level. *Education Week 24*(25), p. 11.

Pennsylvania Department of Education (2007). *2007-2008 PSSA assessment handbook.* Retrieved September 27, 2007, from http://www.able.state.pa.us/a_and_t/lib/a_and_t/2007-2008_Assessment_Handbook.pdf

Perrone, V. (1991). *ACEI position paper on standardized testing*. Retrieved April 11, 2008, from http://www.acei.org/onstandard.htm#index

Plato Learning (2005). The need for formative assessment in education. *Plato Teaching and Learning Enterprise.* Retrieved March 2, 2008, from http://www.plato.com/media/Technical-White%20Papers/2/The%20Need%20for%20Formative%20Assessment%20in%20Education.pdf

Popham, W.J. (2008). *Ten "must-know" facts about educational testing*. Retrieved April 17, 2008 from http://www.pta.org/archive_article_details_1117837372328.html

Quality Counts 2001. (2001). *Education Week* [On-line]. Available: http://www.edweek.org/sreports/qc01/

Rapp, D. (2001). Ohio teachers give tests an 'f'. *Rethinking School Online, 15*(4). Retrieved October 8, 2007, from http://www.rethinkingschools.org/archive/15_04/Ohio154.shtml

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*(1), 4-13.

Reeves, J. (2002). *Aptitude assessment for career and educational guidance*. The Work Suite. Retrieved April 17, 2008 from http://www.theworksuite.com/id15.html

Sadker, D. & Zittleman, K. (2004). Test anxiety: Are students failing tests – or are tests failing students? *Phi Delta Kappan, 85*(10), pp. 740-744, 751.

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

Salvia, J., & Ysseldyke, J. (2004). *Assessment*. 9th ed. Boston: Houghton Mifflin.

Schmoker, M. (1999). *Results: The key to continuous school improvement (2nd ed.).* Alexandria, VA: ASCD.

Schwartz, E (2003). Standardized testing in a non-standardized world. Retrieved April 7, 2008 from http//www.waldorflibrary.org/Journal_Articles/RB1203.pdf

Sebatane, E. (1998). Assessment and classroom learning: A response to Black &William. *Assessment in Education, 5*(1), 123-130.

Shields, R (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.

Shoukri, M. (2004). *Measures of interobserver agreement*. Boca Raton, Florida: CRC Press LLC.

Simon, S. (2005). What is a kappa coefficient? Creative Commons Attribution. Retrieved May 23, 2008 from *http://www.childrensmercy.org/stats/definitions/kappa.htm*

Sirotnik, K. (2004). *Holding accountability accountable: what ought to matter in public education*. New York: Teachers College Press.

Sloane, F. & Kelly, A. (2003). Issues in high stakes testing programs. *Theory into Practice, 42*(1), 12-17.

Smith, M. (2004). *Political spectacle and the fate of american schools*. New York: RutledgeFalmer.

Spellings, M. (2005). Secretary Spellings Announces Growth Model Pilot, Addresses Chief State School Officers' Annual Policy Forum in Richmond. *United States Department of Education.* Retrieved February 29, 2008, from http://www.ed.gov/news/pressreleases/2005/11/11182005.html

Starkman (2006). Formative assessment: Building a better student. *The Journal*. Retrieved March 2, 2008, from http://thejournal.com/the/printarticle/?id=19174

State of New Jersey Department of Education (2007). *Statewide assessment schedule.* Retrieved September 27, 2007, from http://www.state.nj.us/education/assessment/schedule.shtml

Staytor, F. & Johnson, P. (1990). Evaluating the teaching and learning of literacy. *Reading and writing together: New perspectives for the classroom.* Norwood, MA: Christopher-Gordon Publisher.

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

Success for All Foundation. (2007). *4Sight reading and math benchmarks 2007-2008 technical report.* Success for All Foundation. Retrieved September 20, 2007, from https://members.successforall.net/index.cfm?CFID=321553&CFTOKEN=77521886

Sun, Y. (2001). "Introduction to linear regression." PowerPoint presentation. California State University, Sacramento, CA.

Tananis, C. (2007). Information technologies for educational leaders. PowerPoint presentation. University of Pittsburgh, Pittsburgh, PA. Fall 2007.

Technology Alliance (2005). *Data driven decision making in k-12 schools*. Retrieved May 2, 2008 from http://www.technology-alliance.com/pubspols/dddm/dddm.html

Trumbull, E., & Farr, B. (2000). *Grading and reporting student progress in an age of standards.* Norwood, MA: Christopher–Gordon.

United States Census Bureau (2000). *Fact sheet: West Mifflin borough, Pennsylvania*. Retrieved May 2, 2008 from http://factfinder.census.gov/servlet/SAFFFacts?_event=Search&geo_id=&_geoContext=&_street=&_county=west+mifflin&_cityTown=west+mifflin&_state=04000US42&_zip=&_lang=en&_sse=on&pctxt=fph&pgsl=010&show_2003_tab=&redirect=Y

U.S. Congress, Office of Technology Assessment (1992). *Testing in American schools: Asking the right questions*. OTA-SET-519. Washington, DC: U.S. Government Printing Office.

U.S. Department of Education (2004). *Helping families, schools, and communities understand and improve student achievement*. Retrieved September 29, 2007, from http://www.ed.gov/print/nclb/accountability/ayp/testingforresults.html

Van Maele, D. (2006). Data use by teachers in high-performing, suburban middle schools to improve reading achievement of low-performing students. (Doctoral dissertation, University of Pittsburgh, 2006).

Wikipedia (2008). Educational Testing Service. Retrieved April 15, 2008 from http://en.wikipedia.org/wiki/educational_testing_service

Wisconsin Department of Public Instruction (2007). *Balanced assessment system*. Retrieved May 2, 2008 from http://dpi.wi.gov/oea/pdf/bas.pdf

Yeh, S. (2005). Limiting the unintended consequences of high-stakes testing. *Education policy analysis archive, 13*(43), 1-23.