International J. Soc. Sci. & Education
2012 Vol. 3 Issue 1, ISSN: 2223-4934 E and 2227-393X Print

Utilizing a Benchmark Formative Assessment to Predict Academic Achievement in a Rural School System

By

[1]Lee James Lesisko and [2]Abdulwahab Sraiheen
[1]Pleasant Valley School System, 2233 Route 115, Suite 100, Brodheadsville, PA 18322
United States of America (lesisko.lee@pvbears.org)
[2]Kutztown University of Pennsylvania, 15200 Kutztown Road, Kutztown, PA 19530
United States of America (sraiheen@kutztown.edu)

Abstract

The focus of this study was to analyze a benchmark formative assessment (4Sight) that predicts how students will perform on the annual Pennsylvania, USA exam that is given in grades three through eight inclusive and 11. Correlation analysis reported a significant positive linear relationship between students' performance in the Pennsylvania System of School Assessment (PSSA) exam and each of the 4Sight exams. Also the authors found significant correlations between PSSA performance by gender, ethnicity, socio-economic status, and grade levels. The research revealed significant differences in the average scores and student grade level, ethnicity, and socio-economic status. This study shows significant gains in student scores and the results clearly support the use of the 4Sight Benchmark Assessment. Not only will 4Sight assist school stakeholders in pinpointing academic achievement, it may also help determine a course of action for staff professional development efforts and spending practices to continue formative benchmark assessments.

Keywords: *Formative Assessment, PSSA, Achievement, 4Sight*

1. Introduction

In the age of accountability where there is a focus on academic achievement, schools across the United States are struggling to meet the No Child Left Behind (NCLB) requirements and at the same time balance extremely tight budgets. Conversations surrounding test scores, high-stakes testing, and meeting Adequate Yearly Progress (AYP) dominate much of today's discussion around the success or failure of public schools. As the proficiency thresholds increase and state sanctions loom over institutions who do not meet NCLB mandates, schools and school systems are seeking early indicators to define areas of academic deficiency.

Many states across the country including Alabama, Arizona, California, Connecticut, Florida, Georgia, Hawaii, Indiana, Mississippi, Missouri, Ohio, Pennsylvania, Texas, and Utah have adopted formative assessments such as the 4Sight Benchmark Assessment developed by the Success for All Foundation of Maryland as an early indicator of student proficiency. This organization defines 4Sight as a benchmark assessment which predicts student achievement in reading and mathematics (Success for All Foundation, 2010).

Although Massachusetts and Connecticut have developed their own version of a benchmark assessment, commercial programs such as Scantron's Performance Series and the Person Benchmark Assessment are also available. However, the focus of this research is to understand how well the Success for All Foundation's 4Sight assessment measures up to the actual NCLB scores from a rural school system in Pennsylvania. The result of having this information will assist school officials in developing plans to adjust or alter current instructional practices in an effort to improve academic achievement on the state's NCLB required yearly assessment.

2. Review of Related Literature

Gronlund and Waugh (2009) postulated that formative assessments are used to monitor student academic progress and designed to measure the extent to which the student has mastered the learning objective(s). The point of formative assessment is to measure all the outcomes of a unit of instruction and to use the results to improve learning. This process allows the educator to briskly identify student learning success or failure and make adjustments throughout instruction in order to improve academic achievement.

Black and William (1998) reported that formative assessment is at the heart of effective teaching especially when educators use it to meet student needs. According to the authors, formative assessment can produce significant and substantial gains in academic achievement. Nevertheless, exams such as high-stakes tests are limited in nature because they only provide overall summaries of achievement rather than useful diagnosis on how to improve.

Benchmark assessments have grown in popularity since the rise of high-stakes testing. Benchmark assessments provide insight into trends in learning and can assist school leaders to understand student needs. The use of benchmarks can proffer entry level performance, and heel academic strengths and weaknesses (Henderson, Petrosino, Guckenburg, & Hamilton, 2007). Strauss and Turner (2009) averred that in response to high-stakes testing requirements, many school systems have implemented intra-school year exams in order to determine student proficiency levels prior to the NCLB exam. These ongoing formative assessments are designed to make students' thinking visible to school officials, educators and students. The interim test can also be a benchmark assessment which seeks to predict a student's performance on an upcoming NCLB exam.

The No Child Left Behind mandate for universal proficiency and competence was an original component of the No Child Left Behind Act (NCLB) (P.L. 107-110, 2002). The central goal of the NCLB Act was to close the achievement-gap and provide all children, from all backgrounds with the tools needed to succeed in today's world. For that reason each identified group of children must reach the mandated level of proficiency in mathematics and reading each year. This benign goal of closing the achievement gap between groups of children has taken on a less benignant aspect with the NCLB mandates that Adequate Yearly Progress (AYP) be shown toward the 2014 goal (Wright, 2008).

The NCLB mandated exam in Pennsylvania is called the Pennsylvania System of School Achievement (PSSA) and is administered in the spring of each year. Many schools and school systems rely on other assessment indicators throughout the year to determine how students will perform on the PSSA in reading and mathematics. One such instrument is the 4Sight Benchmark Assessment. The 4Sight benchmarks are considered to be low-stakes quarterly formative assessments for grades three through eight and 11, and resemble PSSA specific assessments in standards measured, response format, length, and types of passages and questions, and provide an estimate of how students would perform on the exam if taken that same day. The 4Sight assessment offers school districts a blue-print of student proficiency and a tool to guide school leaders and classroom teachers so they can make informed decisions about school programs in an effort to improve academic achievement (PSEA Education Services Division, 2007; Castanga, 2008).

The 4Sight, at least in Pennsylvania, was created in collaboration with the Pennsylvania Center for Data-Driven Reform in Education based at John Hopkins University. The assessment predicts, using a linear regression model, how students would score on the PSSA exam. Furthermore, 4Sight produces scores on key sub-skills that are focused on state standards and are valid and reliable and correlated to the PSSA. Each item on the exam addresses a Pennsylvania standard and/or assessment anchor (Success for All Foundation, 2010; Success for All Foundation, 2008).

In a quantitative study by Strauss and Turner (2009) that analyzed the 4Sight and PSSA for a school district in Pennsylvania found that the 4Sight Benchmark Assessment is not fully aligned with PA standards. In fact, only 40% of the eligible content in reading and 80% in mathematics are covered on the exams. The authors further explained that according to the use of the 4Sight by the Success for All Foundation and the Pennsylvania Department of Education, the 4Sight only provides feedback to educators regarding student performance. By not providing feedback to students on their strengths and weaknesses the 4Sight does not conform to the definition of a true formative assessment as defined by the National Research Council.

Castanga (2008) studied a suburban middle school in western Pennsylvania during the 2007-08 academic year which focused on analyzing both PSSA and 4Sight scores. One of the research questions was to understand if the time spent on 4Sight testing was worth the loss of instructional time in the classroom. The author reported that some educators view 4Sight testing as another example of over-testing students while the opposition views testing as a valuable tool to gain insight about student progress. Castanga found strong correlations between the predicted PSSA scores and raw scores on the 4Sight exams. The researcher indicated that the 4Sight assessment is an effective resource for middle schools.

3. Methods

The research questions for this study are two-fold. First, how close of an indicator are the predicted 4Sight scores to actual students' performance on the annual PSSA exam? Second, can any generalizations be made in terms of actual student achievement when analyzing PSSA scores against gender, ethnicity, social-economic status, and grade level when comparing 4Sight predictions? The answers to these questions will not only help rural school leaders understand the impact of the use of an early indicator systembut also assist their counterparts in suburban and urban areas as well.

Although the data examined in this study are accurate and true, the Superintendent of Schools has requested that the district's name remain confidential. To this end, the school system used for this study will be referred to as the Granite Rock School District. Granite Rock enrolls about 6,000 students in grades Kindergarten through 12 in three elementary, one intermediate and two secondary schools. The communities included in the district cover almost 120 square miles, and 157 school bus routes transport children each day. This district is rural, and the community is dominated by farming, retail, recreation, and light manufacturing businesses. The children are primarily Anglo-white with only 7% of the population of students being English Language learners or ethnic minorities including African-Americans.

The data collected for this report were extracted from Granite Rock's raw PSSA and 4Sight data sources during the 2009-10 school year. Employed in this study were individual results consisting of all students who took both the PSSA and 4Sight mathematics and reading assessments during that time frame. This was a total of 2,780 mathematics and 2,600 reading PSSA test scores for a combined total of 5,380 records. The 4Sight scores consisted of three exams, one administered in the fall, winter, and spring months of that same year. The number of data records examined was equivalent to the PSSA records in reading and mathematics. However, the grade six mathematics 4Sight data for the spring assessment was not used due to corrupt data elements from the source file and the winter and spring grade three reading assessments were not available as well due to the district switching to an alternative assessment program. The raw data were downloaded from both databases and copied into a single spreadsheet then uploaded into the Statistical Analysis System (SAS) for disaggregated reporting and data analysis to identify significant patterns, differences and commonalities.

To answer the research questions, the authors begin by providing descriptive statistics (mean, standard deviation, minimum and maximum) for the four variables of interest (PSSA, along with the first, second and third 4Sight) broken down by gender, grade, school level, ethnicity, and socio-economic status. Second, Pearson correlations were produced between the four variables of interest and different students'

attributes and other variables contained in the data set. Third, t-tests were used for PSSA scores by gender, socio-economic status, and those who took all three 4Sight exams, as well as for those students who did not take all three tests. Fourth, an Analysis of Variance was employed for the difference of means using PSSA by grade, ethnicity, and school level. Finally Regression Analysis was used to explain the variability in PSSA scores and provide a statistical formula (an estimated regression equation) to predict a student's PSSA score using the first, second and third 4Sight assessments and the student's attribute gender, socioeconomic status, and ethnicity. Several hypotheses were tested including the following:

Correlation hypotheses:

- H1: No significant relationship between students' PSSA scores and the 4Sight scores.
- H2: No significant relationship between students' PSSA scores and the 4Sight scores in different grade levels.
- H3: No significant relationship between students' PSSA scores and the 4Sight scores for different ethnic groups.
- H4: No significant relationship between students' PSSA scores and the 4Sight scores for socio-economic status.
- H5: No significant relationship between students' PSSA scores and the 4Sight scores for male and female students.

Test of difference between two means (averages), t-test:

- H6R: There is no significant difference between the average score of male and female students in reading.
- H6M: There is no significant difference between the average score of male and female students in mathematics.
- H7: There is no significant difference between the average score of those students who are social-economically disadvantaged and those who are not in mathematics and reading.
- H8: There is no significant difference between the average score of students who took one 4Sight test and students who took all three 4Sight tests in mathematics and reading.

Test of difference between more than two means (averages), Analysis of Variance (F-test):

- H9: There is no significant difference between the average score of students in different grade levels in mathematics and reading.
- H10 There is no significant difference between the average score of students in different ethnic groups in mathematics and reading.
- H11: There is no significant difference between the average score of elementary, intermediate, middle and high school students in mathematics and reading.

3. Discussion

Table 1 and Table 2 present descriptive statistics for some of the variables in this study.

Table 1

Descriptive Statistics for Mathematics Scores by School Level

Level	***Variable***	***N***	***M***	***SD***	***Minimum***	***Maximum***
Elementary	PSSA	707	1415.65	181.42	826	2198
	1st 4Sight	707	1164.57	112.19	887	1518
	2nd 4Sight	707	1257.94	106.98	965	1534
	3rd 4Sight	707	1313.03	105.66	866	1534
Intermediate	PSSA	1245	1420.59	210.89	700	2063
	1st 4Sight	1245	1193.50	135.04	820	1563

	2nd 4Sight	1245	1314.22	145.71	820	1605
	3rd 4Sight	828	1350.18	135.46	865	1605
Middle School	PSSA	465	1416.30	210.11	970	2090
	1st 4Sight	465	1325.92	155.42	967	1763
	2nd 4Sight	465	1357.23	162.99	1018	1737
	3rd 4Sight	465	1408.58	154.76	1044	1763
High School	PSSA	363	1406.38	234.01	776	2377
	1st 4Sight	363	1359.03	146.35	1046	1725
	2nd 4Sight	363	1389.32	155.92	999	1725
	3rd 4Sight	363	1414.90	154.89	1023	1725

Table 2

Descriptive Statistics for Reading Scores by School Level

Level	***Variable***	***N***	***M***	***SD***	***Minimum***	***Maximum***
Elementary	PSSA	660	1389.11	182.06	826	2046
	1st 4Sight	660	1289.00	121.07	845	1598
	2nd 4Sight	360	1365.84	140.95	902	1607
	3rd 4Sight	360	1373.28	150.24	829	1635
Intermediate	PSSA	1169	1407.16	198.61	722	2120
	1st 4Sight	1169	1294.32	148.15	778	1572
	2nd 4Sight	1169	1314.70	143.72	752	1585
	3rd 4Sight	1169	1342.72	136.75	775	1584
Middle School	PSSA	421	1462.58	221.72	700	2342
	1st 4Sight	421	1368.71	165.78	742	1637
	2nd 4Sight	421	1432.49	153.71	742	1646
	3rd 4Sight	421	1411.40	135.96	827	1618
High School	PSSA	350	1366.21	217.29	700	2031
	1st 4Sight	350	1349.13	150.47	690	1581
	2nd 4Sight	350	1353.01	148.63	752	1601
	3rd 4Sight	350	1361.41	119.13	838	1551

For mathematics, the average PSSA score is 1417 with average variability around the mean (one standard deviation) is 207 points. The average score on the 4Sight tests increased from the first to the second and to the third. For Reading, the average PSSA score is 1406 with average variability around the mean (one standard deviation) is 203 points. The average score on the 4Sight tests increased from the first to the second and to the third in Reading as well. The following are results with respect to correlation and hypotheses testing.

There is a strong positive linear relationship between PSSA scores and each of the three 4Sight scores for both mathematics and reading. The sample correlation coefficient between the mathematics PSSA scores (Table 3 below) and the first, second, and third 4Sight scores are 0.6616, 0.6885 and 0.6910. As for reading PSSA scores, the correlation coefficients are 0.7007, 0.7028, and 0.6957 respectively. All coefficients are significantly different than zero which implies relatively strong positive linear relationship between performance on the 4Sight test and performance on the PSSA mathematics and reading assessments. Based on the above results hypothesis H1 is rejected.

Table 3				
Pearson Correlation Coefficients Between PSSA Scores and 4Sight Tests				
Variable	***N***	***Mathematics***	***N***	***Reading***
1st 4Sight	2780	0.6616	2600	0.7007
2nd 4Sight	2780	0.6885	2300	0.7028
3rd 4Sight	2363	0.6910	2300	0.6957

Tables 4 and 5 show the correlation coefficients between the PSSA scores and the first, second and third 4Sight scores. All coefficients are statistically significant at the 99.99% level. It is observed that the linear positive associations between the PSSA scores and the three 4Sight scores are stronger for mathematics than reading especially for the seventh, eighth, and eleventh grade levels. Based on the very large and significant correlation coefficients, hypothesis H2 is also rejected.

Table 4							
Pearson Correlation Coefficients Between PSSA Mathematics Scores and 4Sight Tests by Grade Level							
Variable	***3rd***	***4th***	***5th***	***6th***	***7th***	***8th***	***11th***
N	312	395	372	417	456	465	363
1st 4Sight	0.6621	0.6608	0.6264	0.7348	0.8328	0.8322	0.8001
2nd 4Sight	0.5711	0.7111	0.5097	0.7727	0.7884	0.7604	0.7668
3rd 4Sight	0.5984	0.6844	0.5474	-	0.7958	0.7627	0.7698

Table 5							
Pearson Correlation Coefficients Between PSSA Reading Scores and 4Sight Tests by Grade Level							
Variable	***3rd***	***4th***	***5th***	***6th***	***7th***	***8th***	***11th***
N	300	360	349	339	481	421	350
1st 4Sight	0.7377	0.7306	0.6851	0.6881	0.7703	0.6981	0.7133
2nd 4Sight	-	0.7178	0.6824	0.6872	0.7610	0.7298	0.7193
3rd 4Sight	-	0.6949	0.6695	0.6719	0.7705	0.7282	0.6845

Tables 6 and 7 show the correlation coefficients between the PSSA scores and the first, second and third 4Sight scores for the different ethnic groups. All coefficients are statistically significant at the 99.99% level except for mixed ethnicity. It is observed that Asian mathematics correlation coefficients between the PSSA scores and the three 4Sight scores are larger than Asian reading correlation coefficients. Based on the very large and significant correlation coefficients, hypothesis H3 is also rejected.

Table 6												
Pearson Correlation Coefficients Between PSSA Mathematics Scores and 4Sight Tests by Ethnicity												
Variable	***N***	***Asian***	***N***	***Black***	***N***	***Hispanic***	***N***	***Mixed***	***N***	***Native***	***N***	***White***
1st 4Sight	51	0.5571	251	0.5853	273	0.6225	5	-0.3114	9	0.8467	2191	0.6649
2nd 4Sight	51	0.6580	251	0.6847	273	0.5843	5	0.1070	9	0.9248	2191	0.6944
3rd 4Sight	46	0.6732	211	0.6289	237	0.6520	4	-0.0337	8	0.9345	1857	0.6948
*All correlation coefficients are significant except for mixed ethnicity and that could be due to the small number of students in that group.												

Table 7												
Pearson Correlation Coefficients Between PSSA Reading Scores and 4Sight Tests by Ethnicity												
Variable	***N***	***Asian***	***N***	***Black***	***N***	***Hispanic***	***N***	***Mixed***	***N***	***Native***	***N***	***White***
1st 4Sight	56	0.5659	246	0.7026	255	0.6252	6	-0.2497	9	0.9473	2021	0.7093
2nd 4Sight	46	0.4906	220	0.6885	215	0.6473	6	0.3633	8	0.8538	1799	0.7122
3rd 4Sight	46	0.5830	220	0.6773	215	0.6637	6	0.0260	8	0.6979	1799	0.7060
*All correlation coefficients are significant except for mixed ethnicity and that could be due to the small number of students in that group.												

Table 8 shows the correlation coefficients between the PSSA scores and the first, second and third 4Sight scores for different socio-economic groups. All coefficients are statistically significant at the 99.99% level for both groups and subjects. It is observed that the disadvantaged student group reading correlation coefficients between the PSSA score and the three 4Sight scores are significantly larger than mathematics correlation coefficients. Based on the very large and significant correlation coefficients, hypothesis H4 is also rejected.

Table 8

Pearson Correlation Coefficients Between PSSA Scores and 4Sight Tests by Socio-Economic Status

	Mathematics				***Reading***			
Variable	***N***	***Yes***	***N***	***No***	***N***	***Yes***	***N***	***No***
1^{s} 4Sight	891	0.6245	1889	0.6467	839	0.7150	1761	0.6748
2nd 4Sight	891	0.6719	1889	0.6795	737	0.7223	1563	0.6775
3rd 4Sight	762	0.6628	1601	0.6873	737	0.7122	1563	0.6728

Table 9 shows the correlation coefficients between the PSSA score and the first, second and third 4Sight scores by gender. All coefficients are statistically significant at the 99.99% level for both genders and subjects. The correlation coefficients for males and females are identical. Based on the very large and significant correlation coefficients, hypothesis H5 is also rejected.

Table 9

Pearson Correlation Coefficients between PSSA Scores and 4Sight tests by Gender

	Mathematics				**Reading**			
Variable	***N***	***Male***	***N***	***Female***	***N***	***Male***	***N***	***Female***
1st 4Sight	1452	0.6749	1328	0.6241	1362	0.7056	1238	0.6885
2nd 4Sight	1452	0.6934	1328	0.6833	1219	0.7140	1081	0.6812
3rd 4Sight	1234	0.6991	1129	0.6817	1219	0.6956	1081	0.6925

Hypothesis 6R states that there is no difference between the average Reading PSSA scores for males and females. A t-test of the difference between two means was conducted and H6R was rejected because females outscored males on average by 42.3 points. The hypothesis was rejected at the 99.99% level. Conversely, hypothesis H6M explains that there is no difference between the average PSSA mathematics score for males and females. A t-test of the difference between the two means was conducted and H6M was accepted (failed to reject). Females and males on average performed equally in mathematics. The hypothesis was not rejected at the 99.99% level.

Hypothesis H7 states that there is no difference between the average PSSA scores in mathematics or reading for the socio-economic groups. Disadvantage students scored on average 86 points less than advantaged group in reading and 83.5 points less in mathematics. Both differences in averages are significantly different from zero. H7 was rejected (p-value < 0.0001).

Hypothesis H8 states that there is no difference between the average PSSA scores in mathematics or reading for the number of 4Sight exams taken. In reading, the average PSSA score increased by 68 points for those students who took the three 4Sight exams than those who only took one 4Sight exam. The opposite is true for mathematics. Students who took the first and second 4Sight scored 52.3 points more than students who took all three exams. Both differences in averages are significantly different from zero. H8 was rejected (p-value = 0.0001).

Hypothesis H9 states that the average score for all grades are equal. Analysis of Variance for the equality of several means (averages) was done for both PSSA mathematics and reading. The following conclusions were reached. F-value = 25.85, p-value <0.0001 for mathematics and F-value = 13.11, p-value <0.0001 for reading which are statistically significant. The average scores of different grades are not equal. Finally, significant differences in mathematics and reading average scores were confirmed using Tukey's Studentized Range (HSD) Test. H9 is rejected at the 99.99% confidence level.

Hypothesis H10 states that students of different ethnic backgrounds on average performed the same on the PSSA. Analysis of Variance for the equality of several means (averages) was done for both PSSA mathematics and reading scores. The following conclusions were reached. F-value = 10.35, p-value <0.0001 for mathematics and F-value = 13.11, p-value <0.0001 for reading). In mathematics, Asians on average outperformed Blacks by 97 points and 110.7 points over Latino students. Also, White students performed on average better than Black and Latino by 61.9 and 75.7 points respectively. These differences in mathematics average scores were confirmed using Tukey's Studentized Range (HSD) Test. H9 is rejected at the 99.99% confidence level. Finally, the authors failed to reject hypothesis H11 which states that there is no difference between students' performance in elementary, intermediate, middle, and high school. The F-value was extremely low (0.45) indicating that the average PSSA scores for elementary, intermediate, middle, and high school are equal.

4. Conclusions

The statistical analysis of the PSSA and 4Sight data showed an important link between the two. Correlation analysis reported a significant positive linear relationship between students' performance in the PSSA exam and each of the 4Sight exams. Also the authors found significant correlations between PSSA performance by gender, ethnicity, socio-economic status, and grade levels. Additional analysis of data avowed that there are significant differences in the average scores and student grade level, ethnicity, and socio-economic status. White and Asian students outperformed their Black and Latino counterparts in both reading and mathematics. Finally, the socio-economic disadvantaged students did worse than the advantaged pupils.

Although Strauss and Turner's (2009) research explained that the 4Sight is not fully aligned with the Pennsylvania standards, the result of this study shows significant gains with respect to test scores. Similarly, Strauss and Turner indicated that the 4Sight Assessment had little or no effect on student performance on the PSSA exam. However, this study shows significant gains in student scores and the results clearly support the use of the 4Sight Benchmark Assessment. Although the reading and mathematics scores show an increase in student achievement by the use of a formative benchmark assessment, the analysis, interpretation, and use of data to make instructional changes is needed for student growth.

The results of this study supports that it may not be necessary for Granite Rock to administer all three 4Sight exams in mathematics. In mathematics, more does not mean better because the average PSSA score dropped for those who took all three 4Sight exams. This same result was echoed by (Castanga 2008) who found the last 4Sight assessment in mathematics was not as highly correlated as the others and recommended that the last 4Sight test be dropped.

Although Castanga (2008) found the same to be true for the reading scores, the researchers of this study found that not to be the case. The research supports and it is highly recommended to continue to administer all three exams in reading as more assessing leads to an increase in the average student score.

Not only will 4Sight assist school stakeholders in pinpointing academic achievement, it may also help determine a course of action or focus for staff professional development efforts to ensure student success and assist with determining the spending practices to continue formative benchmark assessments.

Regression analysis revealed that the inclusion of all three 4Sight exams explained about 59.3% of the variability in the PSSA reading scores and 54.5% of the variability in the mathematics PSSA scores. Further analysis is needed with new data to confirm the findings of this study.

References

Black, P., and William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80(1)*:pp. 139-148.

Castanga, D. R. (2008). A quantitative case study analysis of the 4Sight bench- mark assessment. Unpublished doctoral dissertation, University of Pittsburgh.

Gronlund, N. E., Waugh, C. K. (2009). *Assessment of student achievement*. Upper Saddle River, New Jersey: Pearson.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). *Measuring how benchmark assessments affect student achievement*. (Issues & Answers Report REL 2007-No. 039). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands.

PSEA Education Services Division. (2007). Professional learning exchange. Retrieved June 4, 2011, from http://www.psea.org/uploadedFiles/Publications/Professional_Publications/Advisories/4SightBenchmark.pdf

Strauss, R. P., & Turner, A. C. (2009). Testing 1,2,3: An analysis of 4Sight in Pennsylvania. Paper presented at the annual conference of the American Education Finance Association.

Success for All Foundation (2010). Overview of 4Sight. Success for All Foundation. Retrieved May 14, 2011, from http://www.successforall.net/Programs/4sight.html

Success for All Foundation (2008). 4Sight reading and mathematics benchmarks 2007-2008 technical report for Pennsylvania. Retrieved May 15, 2011 from: http://www1.whsd.net/courses/H0120/4Sight/technical%20manual.pdf

Wright, R. J. (2008). *Educational assessment: Tests and measurement in the age of accountability*. Thousand Oaks, CA: Sage Publications.