THE SCHOOL DISTRICT OF PHILADELPHIA

Correlation and Classification Accuracy between the Star Computer Adaptive Tests and the PSSAs, 2021-22

Key Findings:

- The correlation between Star Reading and the PSSA ELA ranged from .73 to .80 across the 2021-22 Star testing windows (Fall, Winter 1, Winter 2, and Spring). The correlation between Star Math and PSSA Math ranged from .68 to .80.
- Classification accuracy for Star Reading and the PSSA ELA varied by grade band: For grades 3-5, sensitivity (the percentage of students who scored proficient or higher on the PSSA ELA who were correctly identified by Star) ranged from 67% to 86% while specificity (the percentage of students who scored below proficient on the PSSA ELA who were correctly identified by Star) ranged from 87% to 94% across grades and testing windows. For grades 6-8, sensitivity ranged from 54% to 66%, while specificity ranged from 94% to 96% across grades and testing windows.
- The estimated probability of scoring proficient or higher on the PSSA ELA given a score of At/Above Benchmark on Star Reading (highest performance level) varied by grade band: 72% to 85% for grades 3-5 and 87% to 93% for grades 6-8, across grades and testing windows. For grades 6-8, there was a relatively higher probability that students who scored *below* At/Above Benchmark would score proficient on the PSSA.
- For Star Math and PSSA Math, sensitivity ranged from 50% to 88%, while specificity ranged from 92% to 98% across grades and testing windows.
- The estimated probability of scoring proficient or higher on the PSSA Math given a score of At/Above Benchmark on Star Math ranged from 61% to 86% across grades and testing windows.

Roland Reyes, MS
Senior Statistician

Molly Schlesinger, PhD
Senior Research Associate

Office of Research and Evaluation

July 2023

Table of Contents

List of Tables

ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions of our colleagues across the Office of Evaluation, Research, and Accountability who support the administration of the Star Assessments and the PSSAs, as well as our colleagues who process and manage the assessment data. The authors thank Lucas Westmaas, Joanne Zhu, and Celia Fan from the District Performance Office for assistance with the data. The authors also thank Ebru Erdem, Joy Lesnick, and Michael Frisone from the Office of Research and Evaluation for helpful feedback and quality assurance support.

Introduction

In the 2021-22 school year, the School District of Philadelphia (SDP) used Star Assessments as the District-wide universal screener for reading and math for students in grades K-12. The Star Assessments, developed by Renaissance Learning Inc., comprise a set of computer-adaptive tests (CATs) and curriculum-based measures (CBMs). These tests serve several purposes throughout the District, including identifying students who are not performing at grade level and may need additional enrichment, monitoring academic achievement, and tracking students' development of skills aligned to state standards. In addition, and focal to the current analysis, two Star CATs, Star Reading and Star Math, are used to estimate the percentage of students who are likely to score proficient or higher on the end of year state assessment, the Pennsylvania System of School Assessment (PSSA). The purpose of this analysis is to evaluate the relationship between Star Reading, Star Math, and the PSSA.

About the Tests

Star Reading and Star Math CATs are administered to students in grades 3-12 District-wide. Star Reading contains items that measure students' skills in domains including vocabulary, reading comprehension, analyzing literary text, and understanding author's craft. Star Math assesses skills in areas including numbers and operations, algebra, geometry and measurement, and data analysis, statistics, and probability. Both tests are 34 items in length and are computer adaptive, meaning that the difficulty of items administered in a testing session will depend on the students' pattern of correct and incorrect responses. At the conclusion of a testing session, students score in one of four levels depending on their performance: At/Above Benchmark, On Watch, Strategic Intervention, or Intensive Intervention. During the 2021-22 school year, Star Assessments were administered four times per year within designated testing windows, and one use of the results was to estimate the percentage of students who were likely to perform at a proficient level on the PSSA that year.

The PSSA is a standards-based, criterion-referenced test administered near the end of the school year. All Pennsylvania public-school students in grades 3-8 are assessed in English Language Arts (ELA) and Math.[1] The purpose of the PSSA is to measure how well students acquired the knowledge and skills described in the Pennsylvania Anchor Content Standards as defined by the Eligible Content.[2] Students score in one of four performance levels depending on how well they did on the assessment: Advanced, Proficient, Basic, or Below Basic. Scores in the Advanced or Proficient range indicate that a student has met grade level standards.[3]

[1] Students who are English Learners who attended school in the United States for less than 12 months by the end of the current year's PSSA testing window are not required to take the PSSA ELA. Students with significant cognitive disabilities might be required to take the Pennsylvania Alternate System of Assessment (PASA) instead of the PSSA. For more details see: https://www.education.pa.gov/K-12/Assessment%20and%20Accountability/PSSA/Pages/default.aspx

[2] Data Recognition Corporation. (2022). *2022 Pennsylvania System of School Assessment technical report: Mathematics, English Language Arts, and Science.* https://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PSSA/Technical%20Reports/2022%20PSSA%20Technical%20Report.pdf

[3] For more information see: https://www.education.pa.gov/K-12/Assessment%20and%20Accountability/PSSA/Pages/DescriptorsCutScores.aspx

Board Goals and Guardrails

The use of the Star Assessments to forecast performance on the PSSAs is embedded within the District's Goals and Guardrails. In 2020-21, the District's Board of Education established the Goals and Guardrails which outline achievement goals for the next five years.[4] Goals 1 and 3 of this framework aim to increase the percentage of students in grades 3-8 scoring Proficient or Advanced on the PSSA ELA and PSSA Math each year, respectively. Star Reading and Star Math results are used as "Leading Indicators" to track progress toward these goals. For this purpose, the primary output of interest from the Star CATs is the number of students scoring At/Above Benchmark, which serves as a within-year estimate of the percentage of students who will score Proficient or Advanced on the PSSAs administered at the end of the school year.[5]

Purpose and Research Questions

The purpose of this analysis was to examine the relationship between Star Reading, Star Math, and the PSSAs. Two approaches were used. First, correlations between the Star CAT and the PSSA were estimated. Correlations are one way to describe the degree to which performance on the tests was related. Second, classification accuracy between the Star At/Above Benchmark performance level and the PSSA Proficient or Advanced performance levels was evaluated. This second approach provided information on how well the Star At/Above Benchmark performance level predicted a Proficient or Advanced score on the PSSA. To this end, the following research questions were investigated:

1. What were the correlations between the Star Assessments and the PSSAs for ELA and Math?
2. How accurately did the Star At/Above Benchmark performance level identify students who scored Proficient/Advanced or Basic/Below Basic on the PSSA for ELA and Math?
3. Among students who scored At/Above Benchmark on Star, what percentage actually scored Proficient or Advanced on the PSSA for ELA and Math?

[4] For more information see: https://www.philasd.org/schoolboard/goals-and-guardrails/

[5] See: https://www.philasd.org/era/wp-content/uploads/sites/865/2022/08/Spring-2022-Progress-Monitoring-Report-Reading-Goals-1-2.pdf

Method

Participants

The Star tests were administered in four testing windows in the 2021-22 school year, and planned analyses focused on all four windows (Table 1). The PSSAs were administered in the spring of the 2021-22 school year.

Table 1. Testing windows in the 2021-22 school year

Testing Windows	Dates
Star Fall	9/8/2021 – 10/8/2021
Star Winter 1	12/1/2021 – 12/23/2021
Star Winter 2	3/7/2022 – 3/31/2022
PSSA	4/25/2022 – 5/13/2022
Star Spring	5/16/2022 – 6/14/2022

Source: SDP 2021-22 Assessment Calendar
Note: The Fall, Winter 1, and Winter 2 testing windows were extended, and test results completed within those extensions were used in later analyses.

This report draws from the sample of students in the 2021-22 school year who completed the PSSA ELA, PSSA Math, or both. Each subject was analyzed separately. Analyses were separated by subject (Reading or Math), and within each subject there were four separate samples, one each for the Fall, Winter 1, Winter 2, and Spring testing windows. To be included in a sample, students must have completed both the PSSA and the Star CAT for the corresponding testing window. If student A completed the PSSA Math and only completed Star Math in the Fall, then they would be included in the Fall sample only. If student B completed the PSSA Math and completed Star Math in all four testing windows, then they would be included in all four samples (Fall, Winter 1, Winter 2, and Spring) in the analyses for PSSA Math and Star Math.

For example, a total of 7,915 third grade students completed the PSSA ELA in the 2021-22 school year (Table 2). Of those students, 7,111 (90%) also completed Star Reading in the Fall and were therefore included in the Fall sample. For the Winter 1 sample, we returned to the total 7,915 third grade students who took the PSSA ELA and found that 7,232 (91%) also completed Star Reading in Winter 1; these students were included in the Winter 1 sample. This sampling procedure was repeated for the remaining testing windows and for each grade.

Given that students were required to take the Star assessments, the samples for each Star testing window were representative of the total number of students who completed the PSSA ELA (Table 2). Similar results were found for the Star Math and PSSA Math samples (Table 3). When looking at the demographic characteristics for each sample, they were also generally representative of the District population of PSSA ELA (Table 4) and PSSA Math (Table 5) test takers.

Table 2. Sample size for each Star Reading and PSSA ELA sample, 2021-22

Grade	Number who took PSSA ELA	Among students who took PSSA ELA, number who also completed Star Reading in the designated testing window							
		Fall		Winter 1		Winter 2		Spring	
		n	%	n	%	n	%	n	%
3	7,915	7,111	90%	7,232	91%	7,472	94%	7,476	94%
4	7,867	7,381	94%	7,315	93%	7,581	96%	7,516	96%
5	7,770	7,321	94%	7,178	92%	7,518	97%	7,386	95%
6	7,274	6,748	93%	6,660	92%	6,979	96%	6,849	94%
7	7,291	6,768	93%	6,592	90%	6,864	94%	6,582	90%
8	7,369	6,843	93%	6,663	90%	6,928	94%	6,492	88%
Total	45,486	42,172	93%	41,640	92%	43,342	95%	42,301	93%

Source: Qlik PSSA and Keystones app, accessed March 2, 2023; Qlik Report Library Academic Screeners, accessed August 31, 2022

Note: % is the percentage of students who took the PSSA who also completed Star Reading in the given testing window. Only includes students with both a PSSA ELA score and a Star Reading score for the designated SDP testing window. Scores from a Star Spanish-Language test or from the PASA were excluded.

Table 3. Sample size for each Star Math and PSSA Math sample, 2021-22

Grade	Number who took PSSA Math	Among students who took PSSA Math, number who also completed Star Math in the designated testing window							
		Fall		Winter 1		Winter 2		Spring	
		n	%	n	%	n	%	N	%
3	8,027	7,367	92%	7,398	92%	7,722	96%	7,599	95%
4	7,958	7,402	93%	7,324	92%	7,694	97%	7,555	95%
5	7,835	7,308	93%	7,227	92%	7,501	96%	7,458	95%
6	7,334	6,721	92%	6,657	91%	6,987	95%	6,831	93%
7	7,348	6,722	91%	6,568	89%	6,906	94%	6,664	91%
8	7,338	6,775	92%	6,618	90%	6,844	93%	6,625	90%
Total	45,840	42,295	92%	41,792	91%	43,654	95%	42,732	93%

Source: Qlik PSSA and Keystones app, accessed March 2, 2023; Qlik Report Library Academic Screeners, accessed August 31, 2022

Note: % is the percentage of students who took the PSSA who also completed Star Math in the given testing window. Only includes students with both a PSSA Math score and a Star Math score for the designated SDP testing window. Scores from a Star Spanish-Language test or from the PASA were excluded.

Table 4. Demographic characteristics for all students who took the PSSA ELA and for each sample who took the PSSA ELA and Star Reading in the designated testing window, 2021-22

Demographic Characteristic	Demographic Group	PSSA[a]	Star Testing Window			
			Fall[b]	Winter 1[c]	Winter 2[d]	Spring[e]
Grade Level	3	17%	17%	17%	17%	18%
	4	17%	18%	18%	18%	18%
	5	17%	17%	17%	17%	18%
	6	16%	16%	16%	16%	16%
	7	16%	16%	16%	16%	16%
	8	16%	16%	16%	16%	15%
Race/Ethnicity	Asian	10%	10%	11%	10%	10%
	Black/African American	46%	46%	46%	46%	46%
	Hispanic/Latinx	24%	24%	24%	24%	24%
	Multi-Racial/Other	5%	5%	5%	5%	5%
	White	15%	16%	16%	15%	16%
Gender	Female	49%	49%	49%	49%	49%
	Male	51%	51%	51%	51%	51%
Economic Disadvantage	Economically Disadvantaged	76%	76%	76%	76%	76%
	Not Economically Disadvantaged	24%	24%	24%	24%	24%
English Learner	English Learner	16%	15%	15%	15%	15%
	Not an English Learner	84%	85%	85%	85%	85%
Special Education	Has an IEP	16%	16%	15%	16%	16%
	Does not have an IEP	84%	85%	85%	85%	85%

Source: Qlik PSSA and Keystones app, accessed August 31, 2022 and March 8, 2023; Qlik Report Library Academic Screeners, accessed August 31, 2022

Note: The PSSA column includes all District students who completed the PSSA. Each Star Testing Window column only includes students with both a PSSA ELA score and a Star Reading score for the designated testing window. Scores from a Star Spanish-Language test or from the PASA were excluded. Students who are American Indian/Alaskan Native or Native Hawaiian/Pacific Islander each represented < 1% of the sample and are included in the Multi-Racial/Other category. Non-Binary students represented < 1% of the sample. IEP = Individualized Education Plan. Category *Has an IEP* does not include students with gifted IEPs. Categories may not sum to 100% due to rounding.

[a]n = 45,486; [b]n = 42,172; [c]n = 41,640; [d]n = 43,342; [e]n = 42,301

Table 5. Demographic characteristics for all students who took the PSSA Math and for each sample who took the PSSA Math and Star Math in the designated testing window, 2021-22

Demographic Characteristic	Demographic Group	PSSA[a]	Star Testing Window			
			Fall[b]	Winter 1[c]	Winter 2[d]	Spring[e]
Grade Level	3	18%	17%	18%	18%	18%
	4	17%	18%	18%	18%	18%
	5	17%	17%	17%	17%	18%
	6	16%	16%	16%	16%	16%
	7	16%	16%	16%	16%	16%
	8	16%	16%	16%	16%	16%
Race/Ethnicity	Asian	10%	11%	11%	11%	11%
	Black/African American	45%	45%	45%	46%	46%
	Hispanic/Latinx	25%	24%	24%	24%	24%
	Multi-Racial/Other	4%	5%	5%	5%	5%
	White	15%	16%	16%	15%	16%
Gender	Female	49%	49%	49%	49%	49%
	Male	51%	51%	51%	51%	51%
Economic Disadvantage	Economically Disadvantaged	76%	76%	76%	76%	76%
	Not Economically Disadvantaged	24%	24%	25%	24%	25%
English Learner	English Learner	17%	16%	16%	16%	16%
	Not an English Learner	83%	84%	84%	84%	84%
Special Education	Has an IEP	16%	16%	15%	16%	15%
	Does not have an IEP	84%	85%	85%	85%	85%

Source: Qlik PSSA and Keystones app, accessed August 31, 2022 and March 8, 2023; Qlik Report Library Academic Screeners, accessed August 31, 2022

Note: The PSSA column includes all District students who completed the PSSA. Each Star Testing Window column only includes students with both a PSSA Math score and a Star Math score for the designated testing window. Scores from a Star Spanish-Language test or from the PASA were excluded. Students who are American Indian/Alaskan Native or Native Hawaiian/Pacific Islander each represented < 1% of the sample and are included in the Multi-Racial/Other category. Non-Binary students represented < 1% of the sample. IEP = Individualized Education Plan. Category *Has an IEP* does not include students with gifted IEPs. Categories may not sum to 100% due to rounding.

[a]n = 45,840; [b]n = 42,295; [c]n = 41,792; [d]n = 43,654; [e]n = 42,732

Measures

Star Reading and Star Math

The Star Reading and Star Math tests provide several metrics that describe student performance.[6] For this analysis, we used the Star Unified Scale Score and Star performance levels.

Star Unified Scale Score

The Star Unified Scale Score is reported on the Unified Scale, which expresses student performance on all Star CATs (e.g., Star Reading, Star Math) using a common metric. The Unified Scale is based on an item response theory (IRT) model that accounts for the difficulty of the items administered when scoring student performance. The Unified Scale is also a vertical scale, meaning that scores reported on this scale can be used to compare student performance across grades and to track student growth over time. The Unified Scale is scaled for grades K-12, with scores ranging from a low of 600 to a maximum of about 1400 for Star Reading and Star Math.

Star Performance Levels

The Star performance levels report student performance in one of four categories: At/Above Benchmark, On Watch, Strategic Intervention, and Intensive Intervention. They are based on the student's national percentile rank (NPR), which compares a student's Unified Scale Score against a nationally representative sample of students who are in the same grade and who took the same CAT at roughly the same time. For the 2021-22 school year, the District's Star performance levels (and related cut scores, or how NPR is used to define each performance level) were as follows:

- Star Reading:
    - At/Above Benchmark (≥ 40th national percentile rank [NPR])
    - On Watch (25th to 39th NPR)
    - Strategic Intervention (24th to 10th NPR)
    - Intensive Intervention (< 10th NPR)
- Star Math:
    - At/Above Benchmark (≥ 70th NPR)
    - On Watch (25th to 69th NPR)
    - Strategic Intervention (24th to 10th NPR)
    - Intensive Intervention (< 10th NPR)

[6] For more details on Star metrics used in SDP see: https://www.philasd.org/research/2022/06/09/star-tests-in-the-school-district-of-philadelphia-a-summary-of-metrics-that-describe-achievement-and-growth/

PSSA

To relate Star performance to the PSSA, the PSSA scale scores and PSSA performance levels were used.

PSSA Scale Score

Similar to the Star CATs, scale scores on the PSSA are based on an IRT model that accounts for the difficulty of the items. However, the PSSA scales are unique for each grade and subject, and therefore it is not appropriate to compare student performance across grades or subjects, or to use scale scores to track student academic growth.[7] The range of scores reported on the PSSA scale depends on subject, grade level, and testing year; based on the 2022 Technical Report, PSSA ELA and PSSA Math scale scores generally ranged from a minimum of 600 to a maximum of about 1650 (ELA) and 1550 (Math).

PSSA Performance Levels

The PSSA performance levels report student performance in one of four categories: Advanced, Proficient, Basic, and Below Basic. Performance levels are based on PSSA scale scores, with the cutoff for Proficient set at 1000 across all grades 3-8 in both ELA and Math.[8] Scores in the Proficient or Advanced range indicate that a student has met grade level standards.

Data Analysis

Two main analyses were completed. First, correlations were estimated between Star Unified Scale Scores and PSSA scale scores by subject, grade level, and testing window. Second, classification accuracy metrics describing the relationship between the At/Above Benchmark on Star and Proficient or Advanced on the PSSA were calculated. These metrics are described in further detail below.

Classification Accuracy

The classification accuracy metrics examined in this report include the percentage of correct classifications, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) (Box 1). The percentage of correct classifications summarizes the proportion of students who were correctly classified as scoring Proficient or Advanced on the PSSA by Star. Sensitivity and specificity were used to answer research question 2 (“How accurately did the Star At/Above Benchmark performance level identify students who scored Proficient/Advanced or Basic/Below Basic on the PSSA for ELA or Math?”) and PPV and NPV (predictive values) were used to address research question 3 (“Among students who scored At/Above Benchmark on Star, what percentage actually scored Proficient or Advanced on the PSSA for ELA or Math?”).

[7] Data Recognition Corporation. (2022). *2022 Pennsylvania System of School Assessment technical report: Mathematics, English Language Arts, and Science.* https://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PSSA/Technical%20Reports/2022%20PSSA%20Technical%20Report.pdf

[8] For more information see: https://www.education.pa.gov/K-12/Assessment%20and%20Accountability/PSSA/Pages/DescriptorsCutScores.aspx

Before calculating these metrics, performance levels from each test were collapsed from four to two categories. For Star, the performance levels used in this analysis were At/Above Benchmark and *below* At/Above Benchmark (comprising On Watch, Strategic Intervention, and Intensive Intervention). For PSSA, the groups used in this analysis were Proficient/Advanced and Basic/Below Basic. Results were then organized into a 2x2 table with counts in each cell, and classification accuracy statistics were calculated as described in Box 1.

Box 1: Calculating Classification Accuracy Metrics

		PSSA	
	Performance levels	**Proficient/Advanced**	**Basic/Below Basic**
Star	**At/Above**	A	B
	Below At/Above	C	D

- Sensitivity = $\frac{A}{A+C}$; the proportion of Proficient/Advanced students who were correctly identified by Star
- Specificity = $\frac{D}{B+D}$; the proportion of Basic/Below Basic students who were correctly identified by Star
- Correct Classifications = $\frac{A+D}{A+B+C+D}$; the proportion of correct classifications
- Prevalence= $\frac{A+C}{A+B+C+D}$; the proportion of students who scored Proficient/Advanced on the PSSA
- Positive predictive value = $\frac{A}{A+B} = \frac{sensitivity \ast prevalence}{sensitivity \ast prevalence + (1 - specificity) \ast (1 - prevalence)}$; the proportion of students who scored At/Above on Star who actually scored Proficient/Advanced on the PSSA
- Negative predictive value = $\frac{D}{C+D} = \frac{specificity \ast (1 - prevalence)}{specificity \ast (1 - prevalence) + (1 - sensitivity) \ast (prevalence)}$; the proportion of students who scored *below* At/Above on Star who actually scored Basic/Below Basic on the PSSA

References:

Akobeng, A. K. (2007). Understanding diagnostic tests 1: Sensitivity, specificity, and predictive values. *Acta Paediatrica, 96,* 338-341. https://doi.org/10.1111/j.1651-2227.2006.00180.x

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 2: Predictive values. *The BMJ, 309,* 102. https://doi.org/10.1136%2Fbmj.309.6947.102

Sensitivity and Specificity

Sensitivity and specificity describe the classification accuracy between Star (At/Above Benchmark vs *below* At/Above Benchmark) and the PSSA (Proficient/Advanced vs Basic/Below Basic).[9] Sensitivity is the proportion of Proficient/Advanced students who were correctly identified by Star. High sensitivity means that among students who scored Proficient/Advanced on the PSSA, a large proportion also scored At/Above Benchmark on Star. Specificity is the proportion of students who scored Basic/Below Basic who were correctly identified by Star. High specificity means that among students who scored Basic/Below Basic on the PSSA, a large proportion also scored *below* At/Above Benchmark on Star. When both sensitivity and specificity are high, the proportion of incorrect classifications (e.g., a student is Proficient/Advanced on the PSSA but scored *below* At/Above Benchmark on Star) is low.[10] While it is desirable to have high values for both metrics (indicating high accuracy), sensitivity and specificity are inversely related, meaning that as one metric increases the other will decrease.

PPV and NPV

PPV (positive predictive values) and NPV (negative predictive values), which collectively are known as predictive values, are used to estimate the probability that students will score in a certain performance level on the PSSA given their performance on Star.[11] PPV is the proportion of students who scored At/Above Benchmark on Star who actually scored Proficient/Advanced on the PSSA, and NPV is the proportion of students who scored *below* At/Above Benchmark on Star who actually scored Basic/Below Basic on the PSSA. PPV and NPV are influenced by a test's sensitivity and specificity, respectively, as well as the proportion of students who scored Proficient/Advanced on the PSSA (prevalence).[12]

Results

Star Reading and PSSA English/Language Arts

Descriptive Statistics

The distribution of PSSA performance levels for each sample with a PSSA ELA score and a Star Reading score was similar to performance levels for all District students who completed the PSSA ELA, with differences not exceeding three percentage points in each grade level (Table 6).[13] Overall, between 29% to 44% of students scored Proficient or Advanced on the PSSA ELA, and between 25% to 34% of students scored At/Above Benchmark on Star across all grades and testing windows.

[9] Akobeng, A. K. (2007). Understanding diagnostic tests 1: Sensitivity, specificity, and predictive values. *Acta Paediatrica, 96,* 338-341. https://doi.org/10.1111/j.1651-2227.2006.00180.x. Trevethan, R. (2017). Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health, 5*(307)*,* 1-7. https://doi.org/10.3389/fpubh.2017.00307

[10] See: https://intensiveintervention.org/sites/default/files/Classification_Accuracy_508.pdf

[11] Akobeng, A. K. (2007). Understanding diagnostic tests 1: Sensitivity, specificity, and predictive values. *Acta Paediatrica, 96,* 338-341. https://doi.org/10.1111/j.1651-2227.2006.00180.x. Trevethan, R. (2017). Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health, 5*(307)*,* 1-7. https://doi.org/10.3389/fpubh.2017.00307

[12] Altman, D. G., & Bland, J. M. (1994). Diagnostic tests 2: Predictive values. *The BMJ, 309,* 102. https://doi.org/10.1136%2Fbmj.309.6947.102

[13] See Appendix A for PSSA performance group distributions for the District.

Table 6. Performance level distributions for students with both a PSSA ELA score and a Star Reading score in each Star testing window, 2021-22

		PSSA					Star			
Grade	**n**	**Pro + Adv**	**Adv**	**Pro**	**Basic**	**Below Basic**	**At/Above**	**On Watch**	**Strategic**	**Intensive**
Fall										
3	7,111	30%	5%	25%	35%	35%	28%	13%	16%	43%
4	7,381	29%	10%	19%	33%	38%	27%	12%	18%	43%
5	7,321	33%	6%	27%	33%	34%	26%	13%	20%	40%
6	6,748	38%	12%	26%	45%	17%	25%	12%	20%	42%
7	6,768	44%	12%	32%	47%	9%	28%	13%	19%	40%
8	6,843	42%	10%	32%	37%	22%	26%	15%	20%	39%
Winter 1										
3	7,232	30%	5%	25%	35%	36%	30%	12%	17%	41%
4	7,315	29%	10%	19%	33%	38%	30%	12%	18%	41%
5	7,178	33%	6%	27%	33%	34%	28%	13%	20%	40%
6	6,660	38%	12%	26%	45%	17%	26%	12%	20%	42%
7	6,592	44%	13%	32%	47%	9%	28%	13%	20%	39%
8	6,663	42%	10%	32%	36%	22%	27%	14%	20%	40%
Winter 2										
3	7,472	29%	5%	24%	35%	36%	34%	12%	17%	38%
4	7,581	29%	10%	19%	33%	39%	33%	13%	17%	37%
5	7,518	33%	6%	27%	33%	34%	31%	14%	19%	37%
6	6,979	37%	12%	26%	45%	18%	28%	13%	20%	40%
7	6,864	43%	12%	31%	48%	9%	28%	13%	21%	38%
8	6,928	41%	10%	31%	37%	22%	26%	15%	20%	39%
Spring										
3	7,476	29%	5%	24%	35%	36%	34%	11%	15%	40%
4	7,516	29%	10%	19%	33%	39%	33%	12%	15%	40%
5	7,386	33%	6%	27%	33%	34%	31%	12%	18%	39%
6	6,849	38%	12%	26%	45%	17%	28%	12%	18%	42%
7	6,582	44%	13%	31%	48%	9%	28%	13%	19%	40%
8	6,492	42%	10%	32%	37%	21%	25%	13%	20%	43%

Source: Qlik PSSA and Keystones app, accessed August 31, 2022; Qlik Report Library Academic Screeners, accessed August 31, 2022

Note: Only includes students with both a PSSA ELA score and a Star Reading score for the designated SDP testing window. Star scores are the student's latest and best score. Scores from a Star Spanish-Language test or from the PASA were excluded. For the PSSA performance levels, Adv = Advanced, Pro = Proficient, Basic = Basic, Below Basic = Below Basic. For the Star performance levels, At/Above = At or Above Benchmark, On Watch = On Watch, Strategic = Strategic Intervention, Intensive = Intensive Intervention.

Correlations

Correlations between Star Reading and PSSA ELA scale scores were estimated for all grades and testing windows. Correlations ranged from .73–.80 across grades and testing windows, with correlations increasing throughout the school year for most grades (Table 7).[14]

Table 7. Correlation between Star Reading Unified Scale Scores and PSSA ELA scale scores in each Star testing window, 2021-22

Grade	Fall		Winter 1		Winter 2		Spring	
	n	Correlation	n	Correlation	n	Correlation	n	Correlation
3	7,111	0.73	7,232	0.76	7,472	0.76	7,476	0.76
4	7,381	0.75	7,315	0.78	7,581	0.78	7,516	0.78
5	7,321	0.77	7,178	0.80	7,518	0.80	7,386	0.80
6	6,748	0.76	6,660	0.78	6,979	0.79	6,849	0.78
7	6,768	0.76	6,592	0.77	6,864	0.78	6,582	0.77
8	6,843	0.76	6,663	0.77	6,928	0.78	6,492	0.75

Source: Qlik PSSA and Keystones app, accessed August 31, 2022; Qlik Report Library Academic Screeners, accessed August 31, 2022

Note: Only includes students with a PSSA score and a Star score for the designated testing window. Star scores are the student's latest and best score. Scores from a Star Spanish-Language test or from the PASA were excluded. All correlations were statistically significant at $p < .001$.

Classification Accuracy

Classification accuracy metrics for each Star Reading and PSSA ELA sample are summarized in Table 8.

Star At/Above and PSSA Proficient/Advanced

The Star At/Above and PSSA Proficient/Advanced columns present the same values reported in Table 6, and the column labeled Star-PSSA Difference captures the difference between the two. Results showed that the differences between the two tests ranged from 0% to 16% across grades and testing windows (absolute value). Results varied by grade band, with grades 3-5 having smaller absolute differences than grades 6-8. In particular, the larger negative values for the 6-8 grade band indicated that more students scored Proficient/Advanced on the PSSA versus scoring At/Above Benchmark on Star.

Correct Classifications

The column labeled Correct Classifications reports the percentage of students who a) scored At/Above Benchmark on Star Reading *and* scored Proficient/Advanced on the PSSA and b) scored *below* At/Above Benchmark on Star Reading *and* scored Basic/Below Basic on the PSSA. Results showed that for each grade, the percentage of correct classifications was consistent across testing windows. Again,

[14] Due to an accumulation of scores in the bottom tail of the Star Reading Unified Scale Score distribution, correlations were re-estimated using Spearman's rho. Spearman's rho correlations tended to be higher than the reported Pearson correlations. See Appendix B for details.

values depended on grade band, with grades 6-8 having a lower rate of correct classifications (79% to 84%) compared to grades 3-5 (85% to 87%) across testing windows.

Sensitivity and Specificity

Sensitivity and specificity describe the accuracy by which Star correctly identified students who scored Proficient/Advanced or Basic/Below Basic on the PSSA: sensitivity is the percentage of students, among those who scored Proficient/Advanced on the PSSA, who also scored At/Above Benchmark on Star; specificity is the percentage of students, among those who scored Basic/Below Basic on the PSSA, who also scored *below* At/Above Benchmark on Star. Higher values indicate higher accuracy. Results showed that specificity tended to be higher than sensitivity across all testing windows (87% to 96% for specificity versus 54% to 86% for sensitivity), suggesting that Star was more accurate when identifying students who scored Basic/Below Basic on the PSSA than when identifying students who scored Proficient/Advanced.

Consistent with the two sections above, sensitivity and specificity results showed patterns by grade band, with higher sensitivity rates observed for grades 3-5 than for grades 6-8 across all testing windows. This suggests that Star was more accurate when identifying students who scored Proficient/Advanced among students in the earlier grades than in the later grades. While specificity rates also tended to vary by grade band, they did so to a much lesser degree, and consistently remained as high or higher than sensitivity.

Sensitivity and specificity rates also tended to change across testing windows. For all grades, sensitivity tended to be lowest in the Fall, increasing each testing window up to Winter 2. This change was more pronounced for grades 3-5 than for grades 6-8, where the latter grade band showed more consistency across testing windows. Given the inverse relationship between sensitivity and specificity, it was not surprising to see specificity rates decrease as sensitivity increased, indicating that although Star became more accurate later in the year when identifying students who scored Proficient/Advanced, the accuracy by which Star correctly identified students who scored Basic/Below Basic slightly decreased.

PPV and NPV

PPV indicates the probability that a student who scored At/Above Benchmark on Star would also score Proficient/Advanced on the PSSA. NPV describes the probability that a student who scored *below* At/Above Benchmark would score Basic/Below Basic on the PSSA.

Results showed that PPV and NPV values varied by grade band. For grades 3-5, the PPVs ranged from 72% to 85%, while the NPVs ranged from 85% to 93% across grades and testing windows. This means that, on average, students in grades 3-5 who scored At/Above Benchmark on Star went on to score Proficient/Advanced on the PSSA between 72% to 85% of the time, while students who scored *below* At/Above Benchmark went on to score Basic/Below Basic between 85% to 93% of the time across grades and testing windows. For grades 6-8, PPVs ranged from 87% to 93%, while NPVs ranged from 74% to 89% across grades and testing windows. This means that, on average, students in grades 6-8 who scored At/Above Benchmark on Star went on to score Proficient/Advanced on the PSSA between 87% to 93% of the time, while students who scored *below* At/Above Benchmark went on to score Basic/Below Basic 74% to 89% of the time.

When comparing grade bands, PPVs tended to be higher for grades 6-8, while NPVs were higher for grades 3-5. Thus, when students in grades 6-8 scored At/Above Benchmark on Star, they had a higher estimated probability of scoring Proficient/Advanced on the PSSA when compared to students in grades 3-5. However, this should be interpreted in light of the NPV: lower NPV means there is a higher probability that a student who scores *below* At/Above Benchmark on Star will go on to score Proficient/Advanced on the PSSA (calculated as [1-NPV]).[15] For example, in Winter 1 the NPV for third grade students was 91% and for eighth grade students, 75%. For third grade students, there was about a 9% probability (1-NPV) that a student who scored *below* At/Above Benchmark on Star went on to score Proficient/Advanced, while for eighth grade students there was a ~25% probability of the same event.

Results also showed that PPV and NPV changed across testing windows depending on grade band. For grades 3-5, PPV tended to decrease from the Fall, where it was the highest, until Winter 2. Conversely, NPV tended to increase within the same time frame for the same grade band. For grades 6-8, PPV and NPV tended to be stable over testing windows.

[15] Akobeng, A. K. (2007). Understanding diagnostic tests 1: Sensitivity, specificity, and predictive values. *Acta Paediatrica, 96,* 338-341. https://doi.org/10.1111/j.1651-2227.2006.00180.x

Table 8. Classification accuracy metrics between Star Reading and PSSA ELA in each Star testing window, 2021-22

Grade	n	Star At/Above	PSSA Pro/Adv	Star – PSSA Difference	Correct Classifications	Sens	Spec	PPV	NPV
Fall									
3	7,111	28%	30%	-2%	86%	73%	92%	79%	89%
4	7,381	27%	29%	-2%	86%	73%	92%	78%	89%
5	7,321	26%	33%	-7%	85%	67%	94%	85%	85%
6	6,748	25%	38%	-13%	81%	58%	95%	89%	79%
7	6,768	28%	44%	-16%	79%	58%	96%	92%	74%
8	6,843	26%	42%	-16%	80%	57%	96%	91%	76%
Winter 1									
3	7,232	30%	30%	0%	87%	78%	90%	77%	91%
4	7,315	30%	29%	1%	86%	78%	90%	75%	91%
5	7,178	28%	33%	-5%	86%	71%	93%	84%	87%
6	6,660	26%	38%	-12%	83%	61%	96%	90%	80%
7	6,592	28%	44%	-16%	80%	59%	96%	92%	75%
8	6,663	27%	42%	-15%	79%	57%	96%	91%	75%
Winter 2									
3	7,472	34%	29%	5%	86%	84%	87%	72%	93%
4	7,581	33%	29%	4%	86%	84%	87%	72%	93%
5	7,518	31%	33%	-2%	86%	76%	91%	81%	89%
6	6,979	28%	37%	-9%	84%	66%	94%	87%	82%
7	6,864	28%	43%	-15%	80%	59%	96%	91%	75%
8	6,928	26%	41%	-15%	80%	58%	96%	91%	77%
Spring									
3	7,476	34%	29%	5%	87%	86%	87%	74%	94%
4	7,516	33%	29%	4%	86%	84%	88%	73%	93%
5	7,386	31%	33%	-2%	87%	76%	92%	82%	89%
6	6,849	28%	38%	-10%	83%	64%	94%	87%	81%
7	6,582	28%	44%	-16%	80%	59%	96%	93%	75%
8	6,492	25%	41%	-16%	79%	54%	96%	91%	75%

Source: Qlik PSSA and Keystones app, accessed August 31, 2022; Qlik Report Library Academic Screeners, accessed August 31, 2022

Note: Only includes students with both a PSSA ELA score and a Star Reading score for the designated SDP testing window. Star scores are the student's latest and best score. Scores from a Star Spanish-Language test or from the PASA were excluded. Star At/Above = Percent scoring At or Above Benchmark on Star. PSSA P/A = Percent scoring Proficient or Advanced on the PSSA. Star – PSSA Difference is the difference between the Star At/Above column and the PSSA P/A column. Correct Classifications = the percentage of students who a) scored At/Above Benchmark on Star and scored P/A on PSSA ELA and b) scored *below* At/Above Benchmark on Star and scored Basic or Below Basic on the PSSA. Sens = Sensitivity. Spec = Specificity. PPV = Positive Predictive Value. NPV = Negative Predictive Value.

Star Math and PSSA Math

Descriptive Statistics

Similar to Star Reading, the PSSA performance level distribution for each PSSA Math and Star Math sample was comparable to that for the District-wide sample of students who completed the PSSA Math, with differences not exceeding two percentage points (Table 9).[16] Overall, between 14% to 22% of students scored Proficient/Advanced on the PSSA Math, and between 10% to 21% of students scored At/Above Benchmark on Star Math across grades and testing windows.

[16] See Appendix A for PSSA performance group distributions for the District.

Table 9. Performance level distributions for students with both a PSSA Math score and a Star Math score in each Star testing window, 2021-22

		PSSA					Star			
Grade	n	Pro + Adv	Adv	Pro	Basic	Below Basic	At/Above	On Watch	Strategic	Intensive
Fall										
3	7,367	22%	8%	13%	21%	57%	14%	26%	16%	44%
4	7,402	18%	6%	12%	23%	60%	11%	23%	21%	46%
5	7,308	15%	4%	11%	29%	56%	11%	26%	19%	44%
6	6,721	16%	6%	10%	23%	62%	10%	29%	21%	41%
7	6,722	16%	7%	10%	22%	62%	13%	31%	22%	34%
8	6,775	14%	5%	9%	18%	68%	13%	36%	22%	29%
Winter 1										
3	7,398	22%	8%	13%	21%	57%	18%	28%	16%	38%
4	7,324	18%	6%	13%	23%	59%	16%	27%	20%	38%
5	7,227	16%	5%	11%	28%	56%	15%	29%	19%	37%
6	6,657	16%	6%	10%	22%	62%	14%	32%	18%	36%
7	6,568	17%	7%	10%	22%	61%	17%	33%	20%	30%
8	6,618	14%	5%	9%	18%	68%	16%	39%	21%	25%
Winter 2										
3	7,722	21%	8%	13%	21%	58%	19%	29%	16%	36%
4	7,694	18%	6%	12%	23%	60%	19%	27%	18%	35%
5	7,501	15%	4%	11%	28%	57%	19%	28%	18%	35%
6	6,987	15%	6%	10%	22%	63%	17%	31%	17%	35%
7	6,906	16%	7%	9%	22%	62%	19%	32%	19%	31%
8	6,844	14%	5%	9%	17%	69%	16%	38%	20%	26%
Spring										
3	7,599	21%	8%	13%	21%	58%	20%	26%	18%	37%
4	7,555	18%	6%	12%	23%	60%	21%	27%	15%	37%
5	7,458	15%	4%	11%	28%	57%	22%	26%	17%	36%
6	6,831	15%	6%	10%	22%	62%	18%	28%	17%	37%
7	6,664	17%	7%	10%	22%	61%	20%	30%	17%	33%
8	6,625	14%	5%	9%	18%	68%	16%	35%	19%	30%

Source: Qlik PSSA and Keystones app, accessed August 31, 2022; Qlik Report Library Academic Screeners, accessed August 31, 2022

Note: Only includes students with both a PSSA Math score and a Star Math score for the designated testing window. Star scores are the student's latest and best score. Scores from a Star Spanish-Language test or from the PASA were excluded. For the PSSA performance levels, Adv = Advanced, Pro = Proficient, Basic = Basic, Below Basic = Below Basic. For the Star performance levels, At or Above = At or Above Benchmark, On Watch = On Watch, Strategic = Strategic Intervention, Intensive = Intensive Intervention.

Correlations

Correlations between Star Math and PSSA Math performance were estimated for all grades and testing windows. Correlations ranged from .68–.80 and tended to be stable throughout the school year (Table 10).

Table 10. Correlation between Star Math Unified Scale scores and PSSA Math scale scores in each Star testing window, 2021-22

Grade	Fall		Winter 1		Winter 2		Spring	
	n	Correlation	n	Correlation	n	Correlation	n	Correlation
3	7,367	0.79	7,398	0.80	7,722	0.79	7,599	0.80
4	7,402	0.77	7,324	0.78	7,694	0.79	7,555	0.79
5	7,308	0.73	7,227	0.73	7,501	0.73	7,458	0.73
6	6,721	0.76	6,657	0.75	6,987	0.76	6,831	0.76
7	6,722	0.71	6,568	0.70	6,906	0.69	6,664	0.71
8	6,775	0.69	6,618	0.68	6,844	0.68	6,625	0.68

Source: Qlik PSSA and Keystones app, accessed August 31, 2022; Qlik Report Library Academic Screeners, accessed August 31, 2022

Note: Only includes students with a PSSA score and a Star score for the designated testing window. Star scores are the student's latest and best score within a given window. Scores from a Star Spanish-Language test or from the PASA were excluded. All results are statistically significant at *p* < .001.

Classification Accuracy

The following classification accuracy metrics for Star Math and PSSA Math are reported in Table 11.

Star At/Above and PSSA Proficient/Advanced

The Star At/Above and PSSA Proficient/Advanced columns present the same values reported in Table 9, and the column labeled Star-PSSA Difference captures the difference between the two. Across grades and testing windows, results showed that the difference between the two tests ranged from 0% to 8% (absolute value). The differences tended to be largest in the Fall testing window, decreasing somewhat throughout the year for all grades except grade 5, which decreased from 4% in the Fall to 0% in Winter 1, then increased to 4% in Winter 2 and 7% in the Spring.

Correct Classifications

The column labeled Correct Classifications describes the percentage of students who a) scored At/Above on Star Math *and* scored Proficient or Advanced on the PSSA and b) scored *below* At/Above on Star Math *and* scored Basic/Below Basic on the PSSA. The percentage of students who were correctly classified ranged from 89% to 91% across grades and testing windows.

Sensitivity and Specificity

As noted earlier, sensitivity and specificity describe the accuracy by which Star correctly identified students who scored Proficient/Advanced or Basic/Below Basic on the PSSA: sensitivity is the percentage of students, among those who scored Proficient/Advanced on the PSSA, who also scored At/Above Benchmark on Star; specificity is the percentage of students, among those who scored

Basic/Below Basic on the PSSA, who also scored *below* At/Above Benchmark on Star. Higher values indicate higher accuracy. Similar to results for Star Reading and PSSA ELA, specificity was higher than sensitivity across all grades and testing windows, meaning that Star Math was more accurate when identifying students who scored Basic/Below Basic on the PSSA than when identifying students who scored Proficient/Advanced (90% to 98% specificity compared to 50% to 88% sensitivity). Sensitivity was lowest in the Fall, increasing throughout the year with rates highest in the Winter 2 and Spring testing windows. Increases in sensitivity were accompanied by small decreases in specificity, suggesting that as the school year progressed, Star became somewhat less accurate when identifying students who scored Basic/Below Basic on the PSSA. Despite this slight decrease, specificity was consistently higher than sensitivity.

PPV and NPV

PPV indicates the probability that a student who scored At/Above Benchmark on Star would also score Proficient/Advanced on the PSSA. NPV describes the probability that a student who scored *below* At/Above Benchmark would score Basic/Below Basic on the PSSA.

Results showed that NPV was higher than PPV across all grades and testing windows: PPV ranged from 61% to 86%, while NPV ranged from 89% to 97%. This means that the probability a student would score Proficient/Advanced on the PSSA given a score of At/Above Benchmark on Star ranged from 61% to 86%, on average, across grades and testing windows; conversely, the probability that a student would score Basic or Below Basic on the PSSA given a score *below* At/Above Benchmark on Star ranged from 89% to 97%, on average, across grades and testing windows. For example, for fifth grade students in Winter 1, students who scored At/Above Benchmark on Star had an estimated 71% probability of scoring Proficient/Advanced on the PSSA, while students who scored *below* At/Above Benchmark had an estimated 94% probability of scoring Basic or Below Basic on the PSSA, on average. The high NPV suggests that the probability a student would score Proficient/Advanced on the PSSA given a score *below* At/Above Benchmark was very low.

Results also showed that PPV decreased from the Fall, where it was the highest, whereas NPV increased slightly across all testing windows.

Table 11. Classification accuracy metrics between Star Math and PSSA Math in each Star testing window, 2021-22

Grade	n	Star At/Above	PSSA Pro/Adv	Star – PSSA Difference	Correct Classifications	Sens	Spec	PPV	NPV
Fall									
3	7,367	14%	22%	-8%	89%	56%	98%	86%	89%
4	7,402	11%	18%	-7%	89%	50%	98%	82%	90%
5	7,308	11%	15%	-4%	91%	56%	97%	77%	92%
6	6,721	10%	15%	-5%	91%	52%	98%	83%	92%
7	6,722	13%	16%	-3%	91%	64%	97%	79%	93%
8	6,775	13%	14%	-1%	91%	64%	96%	71%	94%
Winter 1									
3	7,398	18%	22%	-4%	89%	66%	95%	79%	91%
4	7,324	16%	18%	-2%	89%	64%	95%	75%	92%
5	7,227	15%	15%	0%	91%	69%	95%	71%	94%
6	6,657	14%	16%	-2%	91%	67%	96%	75%	94%
7	6,568	17%	17%	0%	91%	74%	94%	72%	95%
8	6,618	15%	14%	1%	91%	71%	94%	66%	95%
Winter 2									
3	7,722	19%	21%	-2%	90%	73%	95%	80%	93%
4	7,694	19%	18%	1%	90%	77%	93%	70%	95%
5	7,501	19%	15%	4%	91%	84%	92%	64%	97%
6	6,987	17%	15%	2%	91%	78%	94%	69%	96%
7	6,906	19%	16%	3%	91%	81%	93%	68%	96%
8	6,844	16%	14%	2%	90%	74%	93%	64%	96%
Spring									
3	7,599	20%	21%	-1%	90%	74%	95%	79%	93%
4	7,555	21%	18%	3%	90%	80%	92%	68%	96%
5	7,458	22%	15%	7%	90%	88%	90%	61%	98%
6	6,831	18%	15%	3%	91%	81%	93%	67%	96%
7	6,664	20%	16%	4%	91%	84%	92%	67%	97%
8	6,625	16%	14%	2%	91%	75%	94%	65%	96%

Source: Qlik PSSA and Keystones app, accessed August 31, 2022; Qlik Report Library Academic Screeners, accessed August 31, 2022

Note: Only includes students with both a PSSA Math score and a Star Math score for the designated SDP testing window. Star scores are the student's latest and best score. Scores from a Star Spanish-Language test or from the PASA were excluded. Star At/Above = Percent scoring At or Above Benchmark on Star. PSSA P/A = Percent scoring Proficient or Advanced on the PSSA. Star – PSSA Difference is the difference between the Star At/Above column and the PSSA P/A column. Correct Classifications = the percentage of students who a) scored At/Above Benchmark on Star and scored P/A on PSSA ELA and b) scored *below* At/Above Benchmark on Star and scored Basic or Below Basic on the PSSA. Sens = Sensitivity. Spec = Specificity. PPV = Positive Predictive Value. NPV = Negative Predictive Value.

Summary

This analysis aimed to describe the relationship between Star and PSSA performance, with attention on correlations between scale scores and classification accuracy between the Star At/Above Benchmark performance level and the PSSA Proficient or Advanced performance levels. Analyses were completed for each testing window, providing insights into the relationship between Star and the PSSA throughout the school year. This section summarizes results in light of the research questions.

Correlations

Results showed that Star Reading correlated with the PSSA ELA between .73 to .80, and Star Math correlated with PSSA Math between .68 to .80 across all grade and testing window combinations. For Star Reading, the correlations were somewhat weaker in the Fall testing window for grades 3-5, but increased in Winter 1 and then stabilized through the rest of the year. For grades 6-8, the correlations were stable across all four testing windows. For Star Math, results showed that the correlations were steady throughout all four testing windows.

Correlational results for grades 3-5 were similar to results from extant research. A recent study examining correlations between Star Math administered in the fall and the PSSA Math found correlations of .76, .79, and .78 for a small sample of 3rd, 4th, and 5th graders, respectively.[17] These are similar to the correlations found in this analysis for Star Math administered in the fall for grades 3 and 4, but are higher for grade 5. Independent studies relating Star Reading with the PSSA ELA were not found, however, research relating Star Reading with the Texas state test, the State of Texas Assessments of Academic Readiness, found correlations ranging from .60 to .70 across fall, winter, and spring testing windows for students in grades 3-5.[18]

Unfortunately, external analyses relating Star assessments to state tests for grades 6-8 are sparse. Renaissance Learning completed a linking study in 2016 analyzing Star and PSSA performance among students in grades 3-8 and found correlations ranging from .79 to .85 for Star Reading/PSSA ELA and .80 to .85 for Star Math/PSSA Math for students in seven school districts in Pennsylvania.[19] Note that the samples in the Renaissance study were quite different from the student population of PSSA test-takers in SDP, which can affect the comparability of results.

Lastly, SDP analyzed the relationship between an alternative assessment tool, the Aimsweb Reading-Curriculum Based Measurement, and the PSSA ELA among a sample of 3rd grade students who

[17] Turner, M. I., Van Norman, E. R., & Hojnoski, R. L. (2022). An independent evaluation of the diagnostic accuracy of a computer adaptive test to predict proficiency on an end of year high-stakes assessment. *Journal of Psychoeducational Assessment, 40,* 911-916. https://doi.org/10.1177/07342829221110676

[18] Ochs, S., Keller-Margulis, M. A., McQuillin, S., & Jones, J. (2016). The validity and diagnostic accuracy of a computer adaptive test of reading. *Research and Practice in the Schools, 4,* 29-41. https://tasp.memberclicks.net/assets/docs/tasp-journal/vol%204%20issue%201%20complete%20issue.pdf

[19] Renaissance Learning Inc. (2016). *Pathway to proficiency: Linking the Star Reading and Star Math scales with performance levels on the Pennsylvania System of School Assessment (PSSA).* https://renaissance.widen.net/view/pdf/ik8tu357dq/R53794.pdf?u=zceria

completed Aimsweb in the spring of 2018.[20] Results showed a showed a correlation of .72 between the two assessments.

Taken together, correlational results from this analysis are similar to results from extant research, suggesting that Star Reading and Star Math performance is related to performance on the respective PSSA assessment. Correlations for grades 3-5 tended to be on par with external research. External research for grades 6-8 was limited, but correlations ranging from .76 to .79 for Star Reading and .68 to .76 were indicative of moderate to strong relationships with the PSSA. Notably, within SDP, the relationship between Star Math and PSSA Math tended to be lower than that found between Star Reading and PSSA ELA.

Classification Accuracy

For Star Reading and the PSSA ELA, there were notable differences in classification accuracy by grade band (grades 3-5 and 6-8). For grades 3-5, the percentage of correct classifications was consistently between 85% to 87% across grades and testing windows. Sensitivity and specificity results suggested that Star more accurately identified students who scored Basic/Below Basic on the PSSA than students who scored Proficient/Advanced on the PSSA. Specifically, 67% to 86% of students who scored Proficient/Advanced on the PSSA scored At/Above Benchmark on Star, while 87% to 94% of students who scored Basic/Below Basic on the PSSA scored *below* At/Above Benchmark across grades and testing windows.

For grades 6-8, the percentage of correct classifications for Star Reading and the PSSA ELA was lower than for grades 3-5, ranging from 79% to 83% across grades and testing windows. Sensitivity and specificity results similarly suggested that Star was more accurate when identifying students who scored Basic/Below Basic than when identifying students who scored Proficient/Advanced. Specifically, between 54% to 66% of students who scored Proficient/Advanced on the PSSA scored At/Above Benchmark on Star, while 94% to 96% of students who scored Basic/Below Basic on the PSSA scored *below* At/Above Benchmark across grades and testing windows. As such, a sizeable portion of students in grades 6-8 who scored Proficient/Advanced on the PSSA ELA scored *below* At/Above Benchmark on Star Reading. It was also observed that grades 6-8 had large differences between the percentage of students who scored At/Above Benchmark on the Star assessment and the percentage of students who scored Proficient or Advanced on the PSSA across all testing windows (see Table 8).

For Star Math and PSSA Math, the percentage of correct classifications was consistent across grades and testing windows, ranging from 89% to 91%. Results for sensitivity and specificity suggested that Star Math tended to be more accurate when identifying students who scored Basic/Below Basic than when identifying students who scored Proficient/Advanced on the PSSA. Specifically, sensitivity ranged from 50% to 88%, while specificity ranged from 92% to 98% across grades and testing windows.

[20] Reitano, A. (2018). *Spring 2018 third grade reading performance: AIMSweb and PSSA assessments.* https://www.philasd.org/research/wp-content/uploads/sites/90/2018/12/Thid-Grade-AIMSWeb-PSSA-Performance-Issue-Brief-Dec-2018.pdf

There is currently no consensus on minimum values for sensitivity and specificity. Researchers offer a range of suggestions, such as a minimum 70% for each metric, with values approaching 80% more desirable; prioritizing the most applicable metric so that the minimum value is 90% or higher; or maximizing both sensitivity and specificity.[21] These differences occur partly because determining optimal values is dependent on the purpose(s) of the assessment, the student population, the consequences of screening decisions, and also because sensitivity and specificity are related so that changes to one metric will affect the other.[22] Of note is that for Star Reading in grades 6-8, the sensitivity metric was lower than 70% across all testing windows, suggesting that the Star Reading At/Above Benchmark threshold may not be adequately identifying students who score Proficient or Advanced on the PSSA ELA. It is important to emphasize, however, that by no means is this standard definitive, and more analyses are needed to determine a suitable standard for SDP.

External research reporting classification accuracy metrics provides further context. The sensitivity and specificity values reported here were somewhat different than those reported by the Renaissance linking study. Specifically, sensitivity was higher than specificity in the Renaissance study, while the reverse was found for SDP. There are two differences between these studies that might account for this. First, as noted earlier, the sample used in the Renaissance study was quite different from the sample of students in SDP, where the sample in the Renaissance study had a much higher proportion of students scoring Proficient or Advanced on the PSSA (i.e., prevalence). Second, the Star performance level thresholds employed in the Renaissance study were somewhat different than those used by the District.[23] When results from this study were compared to the 2018 SDP analysis of third grade students who completed Aimsweb in the spring, sensitivity and specificity were higher for Star Reading.

PPV and NPV

Another major objective of this analysis was to estimate the probability that a student who scored At/Above Benchmark on Star would score Proficient or Advanced on the respective PSSA test. The predictive values were used for this purpose.

For Star Reading and the PSSA ELA, PPV varied by grade band. For grades 3-5, PPV ranged from 72% to 85% across grades and testing windows. For grades 6-8, PPV ranged from 87% to 93%. As noted earlier, these results should be considered in light of the NPV. For grades 3-5, NPV ranged from 85% to 94%, while for grades 6-8, NPV ranged from 74% to 82%. This suggests that when the Star At/Above

[21] See https://intensiveintervention.org/sites/default/files/NCII_AcademicScreening_RatingRubric_2020-06-30.pdf. Also see page 18 in Klingbeil, D. A., Van Norman, E. R., Nelson, P. M., & Birr, C. (2018). Evaluating screening procedures across changes to the statewide achievement test. *Assessment for Effective Intervention, 44,* 17-34. https://doi.org/10.1177/1534508417747390

[22] Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45,* 117-135. https://doi.org/10.1016/j.jsp.2006.05.005; Kettler, R. J., Glover, T. A., Albers, C. A., Feeney-Kettler, K. A. (2014). An introduction to universal screening in educational settings. In R.J. Kettler, T. A. Glover, T., C. A. Albers, & K. A. Feeny-Kettler (Eds.), *Universal screening in educational settings: Evidence-Based decision making for schools* (pp. 3-16). American Psychological Association; Trevethan, R. (2017). Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health, 5*(307), 1-7. https://doi.org/10.3389/fpubh.2017.00307

[23] Approximate ranges of cut scores were the 35th-45th national percentile rank (NPR) for Star Reading and the 70th-80th NPR for Star Math, depending on grade level.

Benchmark performance level is used to predict performance on the PSSA, there is a larger probability that students in grades 6-8 who score *below* At/Above Benchmark also score Proficient/Advanced on the PSSA when compared to grades 3-5.

For Star Math and PSSA Math, results showed that PPV ranged from 61% to 86% while NPV ranged from 89% to 97% across grades and testing windows. NPV was consistently higher than PPV, suggesting that if the Star At/Above Benchmark performance level is used to predict performance on the PSSA, there is a relatively low probability that students who score *below* At/Above Benchmark also score Proficient/Advanced on the PSSA.

Limitations

The findings from this analysis should be considered in light of its limitations. First, classification accuracy metrics were calculated and interpreted with a focus on the identification of students who scored Proficient or higher on the PSSA. This is consistent with one purpose of the Star CATs at SDP, which is to estimate the percentage of students needed to score At/Above Benchmark on Star in order to meet Goals 1 and 3 of the Board's Goals and Guardrails (i.e., the role of Star as a 'Leading Indicator'). However, another purpose of the Star CATs is academic screening, and classification accuracy analyses focusing on that purpose would emphasize the accuracy that the At/Above Benchmark Performance Level identifies students who are at risk of *not* meeting grade level standards (i.e., Proficient or higher on the PSSA). For this purpose, results show that the At/Above Benchmark Performance Level is quite accurate. Second, this analysis only used data from the 2021-22 school year because it was the first year that District-wide Star data and PSSA data were available, and additional analyses year-over-year are needed to determine the stability of results.

Conclusion

Overall, results showed that performance on Star Reading and Star Math was related to performance on the respective PSSA test. When using the Star At/Above Benchmark cut score (≥ 40th NPR for Star Reading and ≥ 70th NPR for Star Math) to predict performance on the PSSA (Proficient/Advanced or Basic/Below Basic), Star Reading correctly identified students between 79% to 87% of the time, while Star Math correctly identified students between 89% to 91% of the time across grades and testing windows; Star Reading showed differences in classification accuracy by grade band. Lastly, classification accuracy was found to be higher when identifying students who scored Basic/Below Basic versus when identifying students who scored Proficient/Advanced on the PSSA.

This analysis provided a first look at the relationship between the Star CATs and the PSSAs in the first year that all SDP students were administered both tests. As noted above, additional analyses year-over-year will be needed to determine the stability of results. Any modifications to the cut scores that define the Star At/Above Benchmark performance level will also require a thorough consideration of the purposes of the Star assessment, the focal population, and the consequences of changing cut scores, among other factors.[24]

[24] Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45,* 117-135. https://doi.org/10.1016/j.jsp.2006.05.005; Trevethan, R. (2017). Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health, 5*(307)*,* 1-7. https://doi.org/10.3389/fpubh.2017.00307

Appendix A: PSSA Performance Level Distributions

Table A1. PSSA ELA performance level distributions District-wide, 2021-22

Grade	N	Advanced	Proficient	Basic	Below Basic
3	7,915	5%	24%	34%	38%
4	7,867	9%	19%	32%	40%
5	7,770	6%	27%	33%	35%
6	7,274	11%	26%	45%	19%
7	7,291	12%	31%	48%	9%
8	7,369	10%	31%	37%	23%

Source: Qlik PSSA and Keystones app, accessed March 2, 2023; PASA excluded

Note: Rows may not sum to 100% due to rounding.

Table A2. PSSA Math performance level distributions District-wide, 2021-22

Grade	N	Advanced	Proficient	Basic	Below Basic
3	8,027	8%	13%	21%	59%
4	7,958	5%	12%	22%	61%
5	7,835	4%	10%	28%	57%
6	7,334	6%	9%	22%	64%
7	7,348	6%	9%	21%	63%
8	7,338	5%	9%	17%	70%

Source: Qlik PSSA and Keystones app, accessed March 2, 2023; PASA excluded

Note: Rows may not sum to 100% due to rounding.

Appendix B: Spearman's rho Correlations Between Star Reading Unified Scale Scores and PSSA ELA Scale Scores

Scatterplots between Star Reading Unified scale scores and PSSA ELA scale scores showed a number of outliers grouped toward the bottom end of the Unified Scale. This occurred because students who score less than half correct on Star Reading are automatically assigned the minimum score of 615 (Star Reading Unified Scale range = 600 to 1400). Because outliers can influence the Pearson correlations reported in Table 7, the robustness of results was checked by re-estimating the correlations using Spearman's rho.

Spearman's rho is an alternative measure of the relationship between Star and PSSA scores that relates the ranks of the scores rather than the scores themselves. Because it operates on ranks, it may be a more robust measure of the relationship when outliers are present.[25] However, unlike Pearson correlations, which describe the strength of the relationship between two variables using a straight line, Spearman's rho describes the monotonic relationship between those variables, meaning that the relationship can be positive (as one variable increases so does the other) or negative (as one variable increases the other decreases), but it does not have to be constant (i.e., linear). Spearman's rho ranges from -1 to 1, with larger absolute values indicating a stronger relationship. Results showed that the Spearman correlations were all larger than the Pearson correlations reported earlier, providing further support for the relationship between the two tests.

Table B1. Spearman's rho Correlations between Star Reading Unified Scale Scores and PSSA ELA scale scores in each Star testing window, 2021-22

Grade	Fall		Winter 1		Winter 2		Spring	
	n	Correlation	n	Correlation	n	Correlation	n	Correlation
3	7,111	0.81	7,232	0.83	7,472	0.83	7,476	0.83
4	7,381	0.83	7,315	0.84	7,581	0.85	7,516	0.85
5	7,321	0.84	7,178	0.85	7,518	0.86	7,386	0.85
6	6,748	0.82	6,660	0.83	6,979	0.84	6,849	0.83
7	6,768	0.83	6,592	0.84	6,864	0.83	6,582	0.84
8	6,843	0.82	6,663	0.82	6,928	0.83	6,492	0.81

Source: Qlik PSSA and Keystones app, accessed August 31, 2022; Qlik Report Library Academic Screeners, accessed August 31, 2022; Qlik Report Library Total Student Enrollment Yearly, accessed September 22, 2022

Note: Only includes students with a PSSA score and a Star score for the designated testing window. Star scores are the student's latest and best score within a given window. Scores from a Star Spanish-Language test or from the PASA were excluded. All results are statistically significant at *p* < .001.

[25] Howell, D. C. (2013). *Fundamental statistics for the behavioral sciences* (8th ed.). Cengage Learning.