An Analysis of Vendor-Provided Benchmark Assessments

Technical Report

Clint Sattler
Supervisor of Research and Evaluation
Knox County Schools
Department of Research, Evaluation, and Assessment

April 2023

Overview

The Knox County Schools (KCS) has used a variety of tests to benchmark students' progress toward proficiency on the state tests (Tennessee Comprehensive Assessment Program; TCAP). Benchmark assessments have served different purposes in KCS, including monitoring progress toward school and district goals, communicating student performance to parents, informing formative processes, and orienting staff and students to grade-level rigor. Knox County has relied on third-party vendors to provide these assessment tools for over a decade. Vendor assessments provide vetted item banks aligned to standards, automated scoring and scaling of results, and standardized reporting (used by students, parents, teachers, and administrators).

Benchmark tests need to have predictive validity to be used for some mid-year resourcing and policy decisions. The KCS department of Research, Evaluation, and Assessment (REA) analyzed the quantitative validity of three benchmark assessments administered since the 2010-2011 school year to students in grades 3 through 8. Attributes of the three benchmark assessments given in the past decade are provided below.

Discovery Education/Thinklink Predictive Assessment (Discovery ED)

The Discovery Education/Thinklink Predictive Assessment was administered online from the 2010-2011 school year (SY1011) to SY1314. KCS administered math and English/Language Arts (ELA) benchmarks to all grade 3-8 students from SY1011 to SY1213. KCS required math and ELA benchmarks for grade 3-5 students in SY1314. All other benchmark tests were considered optional. KCS administered the tests three times per year (Fall, Winter, and Spring benchmarking periods). KCS ceased using the Discovery ED benchmark because the vendor no longer provided the predictive assessment product.

Discovery ED was a multiple-choice fixed-standard assessment. The items on the test were grade-level content and pre-equated to mimic the difficulty profile of the state test. Discovery ED's psychrometric team aligned benchmark content with Tennessee state standards (except SY1314; see the methodology section) as presented in the Tennessee Department of Education's (TDOE's) state test blueprint. KCS provided de-identified data to Discovery ED after each state testing cycle so that Discovery ED could calibrate their proficiency prediction models annually.

Post-benchmark reports were available electronically through a web-based portal. The results were generated via batch processing, so students who tested early in the testing window were required to wait until the testing window closed to see their results. Teacher reports aggregated results by reporting categories that aligned with TCAP reporting categories. Student-level data could be filtered to aggregate results by ethnicity, race, economic status, English learner status, and special education status. Teachers could export

an item analysis for each benchmark to see student responses for each question. The item analysis included the question number, reporting category, state standard, difficulty, and correct answer. Discovery ED would flag items students missed most frequently and hyperlink to resources that could be used in the classroom tied to the questions reporting category.

Discovery ED measured student performance on a single vertical scale. Student growth was calculated by comparing performance one two consecutive benchmark assessments. Growth was measured comparatively to the other students in the district. Estimates of absolute growth could be calculated (offline) by REA staff since all exports included the students vertical scale score.

STAR Renaissance (STAR)

TDOE launched a redesigned intervention process (RTI^2) in SY1415. RTI2 mandated the use of skills-aligned screening assessments. KCS contracted with STAR Renaissance to screen students for intervention and benchmark students' academic progress. KCS administered STAR Renaissance online from SY1415 to SY1819 in three testing windows (Fall, Winter, and Spring in ELA and Math), but schools could also test outside of these windows to monitor students' progress. All students in grades 3-8 took the assessment in SY1415 and SY1516. The district made participation optional in SY1617 (for grades 3-5) when KCS made Aimsweb the elementary RTI^2 screener. KCS required students in grades 6-8 to take the STAR assessment until SY1819. KCS made STAR optional for grades 6-8 in SY1819 (when Aimsweb was designated the RTI^2 screener for all KCS students). KCS abandoned benchmark testing in SY1819 for a variety of reasons. The reasons for ending the benchmark testing program included:

- Skills-focused assessments were deemed more effective RTI^2 screeners.
- Concerns over the time spent assessing students (through benchmarking, screening, and diagnostic testing) versus teaching students.
- Budgetary constraints.

STAR Renaissance was a multiple-choice and multiple-select computerized adaptive assessment. The testing platform adjusted the difficulty of the questions until it could reasonably determine a student's instructional level (grade-level equivalent). Students below or above grade level would see few grade-level items after their initial benchmark test. STAR Renaissance developed an extensive item bank aligned to Tennessee state standards to ensure that students taking the assessment multiple times would not see the same question more than once.

Post-benchmark reports were available immediately after test administration via a web-based portal. Teachers and students did not have to wait for the end of a testing window to

see the results. Teacher reports aggregated results by domains (aligned to TCAP reporting categories), standards (aligned to state standards), and skill areas. Teachers could not obtain an item analysis for the benchmark to protect the integrity of STAR's item pool. Instead, teachers could see student performance indices in each domain, standard, and skill. STAR Renaissance's proprietary algorithms would use the information from the benchmark tests to recommend a sequence and provide resources to address student needs. STAR would also flag students who seemed to rush through their test (based on the amount of time a student was on the testing platform) to alert teachers to suspect test results.

STAR measured student ability on a single vertical scale. STAR calculated student growth by comparing scaled scores across assessments. STAR modeled student growth using student growth percentiles (SGPs). REA could calculate absolute growth from vertical scale scores. Additionally, STAR extracts included student ability estimates (Rasch Scores) and the standard error of measurement (SEM).

Mastery View Predictive Assessment (Mastery View/Case 21)

The differential impacts of the COVID-19 pandemic led to the reimplementation of benchmark testing in SY2122. KCS responded to the shutdown at the end of SY1920 with a goal of "acceleration instead of remediation." The acceleration process required teachers to uncover knowledge gaps that developed during interrupted learning cycles in SY1920 and SY2021. KCS chose Case 21 (now Mastery View Predictive) assessments to help inform acceleration-focused instructional practices. The Mastery View/Case 21 assessment has been administered online from SY2122 to the present. The district administered the test three times in SY2122 (Fall, Winter, and Spring), but logistic issues limited KCS to two benchmarks in SY2223. KCS requires Mastery View benchmarks in grades 2-12 in ELA, and math and grades 3-12 in Science and Social Studies.

Mastery View/Case 21 is a multiple-choice, multiple-select fixed-standard assessment. English/Language Arts tests also contain a writing prompt not used in scoring. The items on the test are grade-level content and pre-equated to mimic the difficulty profile of the state test. The Mastery View development team uses TDOE test blueprints and input from district-level staff to identify content for the benchmark exam. The district's input ensures that the content on the predictive assessment aligns with KCS pacing guides. Theoretically, this prevents students from encountering content they haven't learned yet.

Mastery View generates reports via batch processing, so students who tested early in the testing window wait until the testing window closes to see their results. Results are available as flat files (.pdfs) distributed through a shared drive. Generally, two people per school have access. The drive administrators distribute class-level reports to the teachers of record.

KCS's largest schools have approximately 150 individual files per subject area that admins must distribute.

Teacher reports show results aggregated by Tennessee state standards and depth of knowledge (DoK, a proxy for item difficulty). Teachers can access an item analysis that shows student-level responses for each question. The item analysis includes the question number, Tennessee state standard, item DoK, and the correct answer. The item analysis also provides the percentage of correct responses in the class, school, and district. Student-level results show the percentage of correct responses, the projected proficiency level on the state exam, and a suggested exam grade that teachers could factor into students' final grades. However, KCS does not factor the benchmark test into any student's final grade.

Mastery View converts the percent correct on the benchmark to a predicted performance level. These raw scores are converted to predicted performance levels through calibration to historic state test data. Mastery View's technical documentation does not suggest they generate a scaled score. Mastery View does not measure student growth, nor can reliable growth measures be determined from the raw data. Raw scores on one benchmark exam may not be comparable to scores in a later benchmark.

Methodology

This analysis studies the accuracy of benchmarks given in the spring at KCS (typically in March or April). REA mapped spring benchmark results to state test results (TCAP) via unique student identifiers (the Tennessee state student identification number). REA excluded results from modified state exams (MSAA and TCAP-Alt) since benchmark vendors don't offer an analogous modified test. The analysis includes data from students in grades 3-8 since these students are required to take the state exam. REA excluded high school results because KCS has not historically required benchmark tests in high school courses. Students in 8th grade who took the Algebra I state End-of-Course exam were excluded from the analysis accordingly.

The number of TCAP/Benchmark data points matched in a content area/grade-level combination varied by year and content area. These discrepancies may impact the validity calculations. Readers can account for the grade tested when comparing the validity data presented in this report using the information in Table 1. Additionally, readers should be cautious when interpreting data generated during SY1314. Tennessee planned to administer the National Partnership for Assessment of Readiness for College and Careers (PARCC) assessment in SY1415. KCS policymakers asked Discovery ED to model the benchmark exam on PARCC standards and performance criteria rather than Tennessee state requirements. The SY1314 information is available in this report, but REA removed SY1314 data from some benchmark comparisons. Additionally, TCAP data was not available for SY1516. TCAP testing was halted that year for technical difficulties when attempting to conduct an online assessment. Therefore, validity calculations are missing for SY1516.

Table 1: Number of Students with TCAP and Benchmark Data by Year

			Grade					
Year	Subject	Vendor	3	4	5	6	7	8
SY1011	ELA	Discovery ED	4193	4214	4194	4027	3895	3767
SY1112	ELA	Discovery ED	4342	4267	4283	4037	3930	3918
SY1213	ELA	Discovery ED	4200	4184	4127	3775	3786	3741
SY1314	ELA	Discovery ED	3435	3594	3579	1680	1936	1251
SY1415	ELA	STAR Renaissance	4521	4378	4379	718	720	479
SY1617	ELA	STAR Renaissance	1178	1295	1123	4081	4017	3926
SY1718	ELA	STAR Renaissance	156	161	141	4268	3964	4001
SY1819	ELA	STAR Renaissance	102	157	74	3175	2965	2749
SY2122	ELA	Mastery View/Case 21	3950	4198	4003	3701	3748	3722
SY1011	Math	Discovery ED	4213	4204	4177	4032	3924	3057
SY1112	Math	Discovery ED	4339	4257	4273	4041	3911	2876
SY1213	Math	Discovery ED	4207	4187	4128	3753	3819	2909
SY1314	Math	Discovery ED	3602	3738	3631	1557	955	567
SY1415	Math	STAR Renaissance	3938	3735	3823	607	717	292
SY1617	Math	STAR Renaissance	1116	1165	1027	4111	4049	2897
SY1718	Math	STAR Renaissance		2		4138	3848	2829
SY1819	Math	STAR Renaissance				3164	3093	2150
SY2122	Math	Mastery View/Case 21	3967	4224	3951	3585	3700	2426
SY1213	Science	Discovery ED	1421	2334	1971	3412	3019	3606
SY1314	Science	Discovery ED	1220	1778	1933	1787	2369	2574
SY2122	Science	Mastery View/Case 21	3921	4208	3934	3651	3689	2596
SY2122	Social Studies	Mastery View/Case 21				3588	3622	3435

REA included two types of predictive validity in this study. Predictive validity measures a benchmark's ability to predict student-level performance on the state assessment. The benchmark tests reviewed in this study estimated student performance on a four-level scale mimicking the TCAP performance levels. Two of the performance levels correspond to "passing" the test therefore we can measure how accurately a benchmark predicted pass/fail performance. REA marked students with matching TCAP/benchmark levels with a 1 and mismatches with a 0. REA determined the predictive validity of benchmark i for each student j as:

$$Predictive\ Validity_i = \frac{1}{N_j} \sum_j Matching\ level_j$$

Additionally, REA used Cramer's V to measure how strongly the predicted performance levels correlated with the performance levels on the state assessment. Benchmark performance levels can be classified as i ∈ (1, 2, 3, 4) and TCAP performance levels can be classified as j ∈ (1, 2, 3, 4). REA determined Cramer's V for benchmark k as:

$$\chi_k^2 = \sum_{i,j} \frac{\left(n_{ij,k} - \frac{n_{i,k} n_{j,k}}{n_k}\right)^2}{\frac{n_{i,k} n_{j,k}}{n_k}}$$

$$V_k = \sqrt{\frac{\chi_k^2 / n_k}{3}}$$

Criterion validity measures the benchmark's accuracy when predicting aggregate (district-level, school-level, or grade-level) performance. Our criterion validity measure only includes the aggregate information for students who took the TCAP and the benchmark assessment. The criterion validity for benchmark i was calculated as:

$$Absolute\ Error_i = \left| \frac{N\ Passing_{TCAP}}{N\ Tested_{TCAP}} - \frac{N\ Passing_i}{N\ Tested_i} \right|$$

$$Criterion\ Validity_i = 100\% - \frac{Absolute\ Error_i}{\left(\frac{N\ Passing_{TCAP}}{N\ Tested_{TCAP}}\right)}$$

Results: District-level Validity

The predictive validity (by year) using four performance levels is in Table 2. REA considers a prediction correct if the TCAP performance level (e.g., Below, Approaching, Met Expectations, Exceeded Expectations) matches the corresponding benchmark performance level (e.g., 1, 2, 3, 4). A student would be marked with a correct four-level prediction if, for example, the benchmark predicted level was a 1 and the student scored in the "Below" category on the state test. REA marked the student with an incorrect prediction if the benchmark assessment predicted any other performance level for the student. Approximately 60% of benchmark assessments accurately place students in their TCAP performance level (e.g., Below, Approaching, On-Track, Mastered). REA reminds readers that SY1314 benchmark tests weren't aligned with the TCAP.

Table 2: Four-level Predictive Validity

		Subject			
Year	Vendor	ELA	Math	Science	Social Studies
SY1011	Discovery ED	61.9%	59.2%		
SY1112	Discovery ED	61.1%	59.3%		
SY1213	Discovery ED	62.8%	60.2%	55.6%	
SY1314*	Discovery ED	52.9%	51.3%	63.5%	
SY1415	STAR Renaissance	60.9%	58.7%		
SY1617	STAR Renaissance	53.6%	61.1%		
SY1718	STAR Renaissance	53.0%	61.5%		
SY1819	STAR Renaissance	61.5%	63.9%		
SY2122	Mastery View/Case 21	54.8%	64.4%	57.6%	59.2%

Two-level predictive validity measures if a passing/not passing level predicted on a benchmark test is aligned with the outcome on the state test. REA considers a two-level prediction correct if, for example, the benchmark reports a student in the Below category and they score in the Approaching category on the TCAP. Two-level predictive validity (by year) is in Table 3. The results indicate approximately 80% of benchmark assessments accurately predict students' pass/no pass status on the TCAP. REA reminds readers that SY1314 benchmark tests weren't aligned with the TCAP.

Table 3: Two-level Predictive Validity

		Subject			
Year	Vendor	ELA	Math	Science	Social Studies
SY1011	Discovery ED	82.6%	82.4%		
SY1112	Discovery ED	82.1%	82.4%		
SY1213	Discovery ED	84.0%	83.1%	80.3%	
SY1314*	Discovery ED	81.2%	77.7%	85.6%	
SY1415	STAR Renaissance	83.9%	82.2%		
SY1617	STAR Renaissance	77.8%	84.5%		
SY1718	STAR Renaissance	78.0%	84.8%		
SY1819	STAR Renaissance	82.4%	85.7%		
SY2122	Mastery View/Case 21	82.7%	87.8%	83.9%	84.2%

Table 4 contains the Cramer's V effect sizes for the relationship between the four-level performance categories predicted by the benchmark exams and students' performance on the TCAP. The effect sizes indicate a strong relationship between students' performance levels predicted by the benchmark and students' TCAP levels.

Table 4: Cramer's V for Benchmark to TCAP: Four-Level Effect Sizes

		Subject			
Year	Vendor	ELA	Math	Science	Social Studies
SY1011	Discovery ED	0.5039	0.5296		
SY1112	Discovery ED	0.4949	0.5183		
SY1213	Discovery ED	0.5195	0.5290	0.4943	
SY1314*	Discovery ED	0.4865	0.4709	0.4998	
SY1415	STAR Renaissance	0.5533	0.5246		
SY1617	STAR Renaissance	0.5054	0.536		
SY1718	STAR Renaissance	0.519	0.5366		
SY1819	STAR Renaissance	0.5151	0.5478		
SY2122	Mastery View/Case 21	0.4883	0.5744	0.5057	0.5522

Criterion validity measures the accuracy of the benchmark prediction in aggregate. REA reports criterion validity (accuracy) on a pass/fail basis since state accountability relies on the percentage of students proficient on the TCAP. District-level accuracy is in Table 5.

Table 5: Two-level Criterion Validity (Accuracy)

Year	Vendor	Subject	TCAP – % Prof	Benchmark - % Prof	Absolute Error	Accuracy
SY1011	Discovery ED	ELA	55.5%	54.0%	1.5%	97.3%
SY1112	Discovery ED	ELA	58.0%	57.4%	0.6%	98.9%
SY1213	Discovery ED	ELA	58.3%	59.4%	1.1%	98.1%
SY1314*	Discovery ED	ELA	53.9%	60.9%	7.0%	87.0%
SY1415	STAR Renaissance	ELA	48.3%	54.7%	6.4%	86.9%
SY1617	STAR Renaissance	ELA	39.9%	55.4%	15.5%	61.2%
SY1718	STAR Renaissance	ELA	37.3%	54.8%	17.6%	52.8%
SY1819	STAR Renaissance	ELA	41.2%	40.0%	1.3%	96.9%
SY2122	Mastery View/Case 21	ELA	40.9%	32.0%	8.9%	78.3%
SY1011	Discovery ED	Math	46.4%	51.7%	5.3%	88.5%
SY1112	Discovery ED	Math	51.2%	52.1%	0.9%	98.2%
SY1213	Discovery ED	Math	54.0%	53.7%	0.3%	99.4%
SY1314*	Discovery ED	Math	55.9%	71.0%	15.2%	72.8%
SY1415	STAR Renaissance	Math	54.5%	48.3%	6.2%	88.6%
SY1617	STAR Renaissance	Math	38.4%	38.2%	0.2%	99.5%
SY1718	STAR Renaissance	Math	36.6%	37.7%	1.1%	97.0%
SY1819	STAR Renaissance	Math	39.7%	39.8%	0.0%	99.9%
SY2122	Mastery View/Case 21	Math	35.9%	30.6%	5.3%	85.3%
SY1213	Discovery ED	Science	70.4%	56.9%	13.4%	80.9%
SY1314	Discovery ED	Science	74.3%	73.9%	0.4%	99.5%
SY2122	Mastery View/Case 21	Science	45.5%	44.1%	1.4%	97.0%
SY2122	Mastery View/Case 21	Social Studies	55.5%	47.2%	8.3%	85.1%

For comparison purposes, Table 6 shows the four-level predictive validity when REA uses the prior-year TCAP performance level to predict the current-year TCAP performance level. For example, a student would be marked with a correct prediction if their 3rd grade Math TCAP was Approaching Expectations and their 4th grade Math TCAP was also Approaching Expectations. A student would have an incorrect prediction if their 4th grade Math TCAP was any other performance level. Readers should note that missing test data in SY1516 and SY1920 prevent us from making some year-to-year comparisons. N counts (similar to the counts in Table 1) are available in Appendix A.

Table 6: TCAP Four-level Predictive Validity

TCAP Test Year		Subject			
Result Year	Basis Year	ELA	Math	Science	Social Studies
SY1112	SY1011	63.4%	58.6%	59.6%	73.8%
SY1213	SY1112	64.0%	58.2%	61.6%	68.1%
SY1314	SY1213	64.4%	55.8%		74.6%
SY1415	SY1314	64.6%	57.4%		
SY1718	SY1617	61.6%	62.2%		
SY1819	SY1718	62.0%	61.9%		55.1%
SY2122	SY2021	60.6%	61.7%		57.3%

Table 7 shows the two-level predictive validity when a REA used prior-year TCAP pass/fail status level is used to predict the current-year TCAP pass/fail status. For example, a student would be marked with a correct prediction if their 3rd grade Math TCAP was Met Expectations and their 4th grade Math TCAP was Exceeding Expectations. REA marked students with an incorrect prediction if their 4th grade Math TCAP performance level was Below Expectations or Approaching Expectations.

Table 7: TCAP Four-level Predictive Validity

TCAP Test Year		Subject			
Result Year	Basis Year	ELA	Math	Science	Social Studies
SY1112	SY1011	82.8%	81.2%	82.4%	89.6%
SY1213	SY1112	83.6%	80.8%	84.3%	90.9%
SY1314	SY1213	84.2%	80.6%		90.4%
SY1415	SY1314	84.4%	81.0%		
SY1718	SY1617	81.4%	84.4%		
SY1819	SY1718	82.7%	84.6%		81.5%
SY2122	SY2021	83.4%	86.0%		81.6%

The criterion validity (accuracy) for using students' aggregate prior TCAP performance levels to predict aggregate current-year performance levels is in Table 8.

Table 8: TCAP Two-level Criterion Validity (Accuracy)

TCAP Test Year		Subject	Current TCAP - % Prof	Prior TCAP - %Prof	Absolute Error	Accuracy
Result Year	Basis Year					
SY1112	SY1011	ELA	59.0%	54.7%	4.3%	92.7%
SY1213	SY1112	ELA	58.6%	59.2%	0.6%	99.0%
SY1314	SY1213	ELA	56.1%	57.6%	1.4%	97.5%
SY1415	SY1314	ELA	55.4%	53.5%	1.8%	96.7%
SY1718	SY1617	ELA	38.8%	41.1%	2.3%	94.1%
SY1819	SY1718	ELA	39.0%	40.3%	1.3%	96.7%
SY2122	SY2021	ELA	39.0%	34.3%	4.7%	87.9%
SY1112	SY1011	Math	49.6%	46.7%	2.9%	94.1%
SY1213	SY1112	Math	51.7%	52.5%	0.7%	98.6%
SY1314	SY1213	Math	51.6%	52.6%	1.1%	97.9%
SY1415	SY1314	Math	54.5%	52.1%	2.4%	95.6%
SY1718	SY1617	Math	38.8%	40.3%	1.5%	96.1%
SY1819	SY1718	Math	41.1%	38.1%	3.0%	92.7%
SY2122	SY2021	Math	33.8%	31.0%	2.8%	91.7%
SY1112	SY1011	Science	64.1%	57.4%	6.8%	89.5%
SY1213	SY1112	Science	68.7%	66.8%	1.9%	97.3%
SY1112	SY1011	Social Studies	85.6%	83.6%	1.9%	97.7%
SY1213	SY1112	Social Studies	88.9%	87.9%	1.1%	98.8%
SY1314	SY1213	Social Studies	87.0%	86.6%	0.4%	99.5%
SY1819	SY1718	Social Studies	49.5%	49.4%	0.1%	99.8%
SY2122	SY2021	Social Studies	51.6%	47.4%	4.2%	91.9%

Results: School-level Absolute Errors

Schools use grade/content level benchmark data to help inform grade-level support structures (instructional coach deployment, Professional Learning Community (PLC) focus, etc.) and mid-year instructional shifts (pacing, lesson spirals, etc.). The accuracy of grade-level proficiency benchmarks may become increasingly important to resourcing decisions as Tennessee enacts a third-grade retention law. Table 9 shows the median absolute errors of school-specific/grade-level/subject-specific predictions. For example, the median difference between the predicted percentage of proficient students and the actual percentage of proficient students in SY1011 third-grade ELA was 3.3 percentage points. REA excluded cells with n counts less than ten from the analysis.

Table 9: Median Benchmark Absolute Errors: School Level (in Percentage Points)

			Grade					
Year	Subject	Vendor	3	4	5	6	7	8
SY1011	ELA	Discovery ED	3.3	5.5	4.1	3.1	4.4	4.3
SY1112	ELA	Discovery ED	8.7	4.7	4.9	5.5	2.1	3.7
SY1213	ELA	Discovery ED	4.4	5.5	7.3	3.5	3.4	3.4
SY1314*	ELA	Discovery ED	12	12	8.7	9.8	5.6	9.8
SY1415	ELA	STAR Renaissance	10	7.7	6.5	1.6	5.3	6.3
SY1617	ELA	STAR Renaissance	15	10	25	22	11	11
SY1718	ELA	STAR Renaissance	15	21	16	19	14	16
SY1819	ELA	STAR Renaissance	4.9	2.5	1.4	3.5	1.5	4.7
SY2122	ELA	Mastery View/Case 21	6.5	12	10	10	8.5	3.9
SY1011	Math	Discovery ED	9.4	6.3	8.8	4.2	4.9	5
SY1112	Math	Discovery ED	9.1	5.1	4.6	3.6	5.8	5.6
SY1213	Math	Discovery ED	6.4	5	6.7	4.8	6.9	4.1
SY1314	Math	Discovery ED	14	23	12	10	12	11
SY1415	Math	STAR Renaissance	6.3	7	10	8	13	8.4
SY1617	Math	STAR Renaissance	15	5.4	10	8.7	5.6	5.2
SY1718	Math	STAR Renaissance				5.6	11	2.5
SY1819	Math	STAR Renaissance				4.7	2.4	5.6
SY2122	Math	Mastery View/Case 21	4	4	4.3	3	5.5	7.4
SY1213	Science	Discovery ED	6.5	12	12	17	18	13
SY1314	Science	Discovery ED	5.4	6.8	7.9	3.6	1.4	3.2
SY2122	Science	Mastery View/Case 21	4.7	7.2	4.9	4.5	5.6	5.1
SY2122	Social Studies	Mastery View/Case 21				4.2	10	10

Each school has a target for increasing its percentage of proficient students on the state test. For K-8 schools, the median target is approximately four percentage points. The median absolute errors indicate that benchmark tools aren't accurate enough to determine if schools meet grade-level targets. Figure 1 shows the distribution of school-level errors by benchmark vendor. The data suggest that errors at the school/grade/content level vary between vendors and content areas.

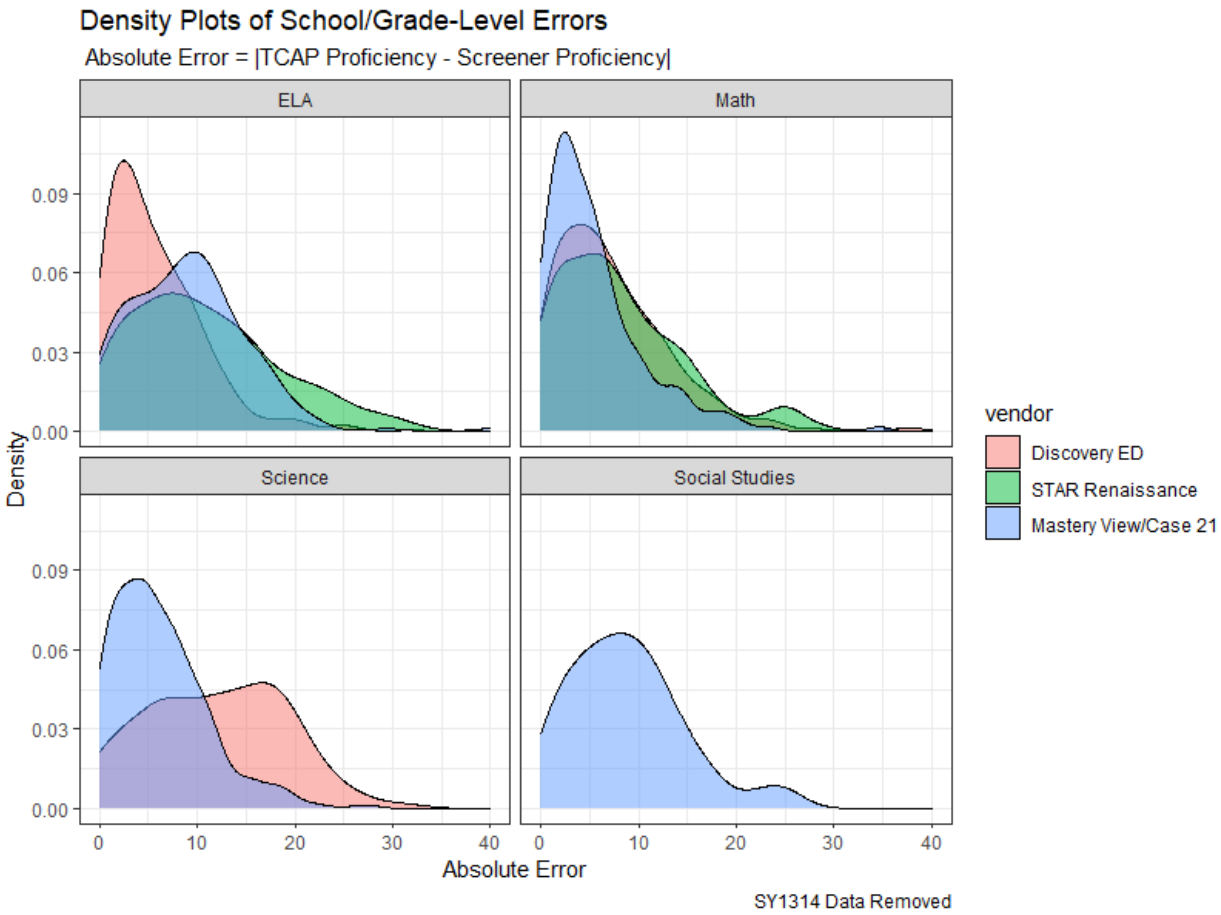

Figure 1: Density Plot of School/Grade-Level Errors

Figure 2 provides information about how the grade-level absolute errors vary by sample size. The errors appear to be somewhat correlated with the number of students tested. However, Figure 2 also shows that school/grade combinations with high n count are not immune to large prediction errors. Figure 3 shows school-level errors after aggregating all grade-level results from a building. You can see that using school-level aggregate data removes some of the data points with the largest errors, especially in ELA. Appendix B contains a table of median absolute errors by year and content area (for comparison with Table 9). The decrease in error using school-level (rather than grade-level) data is more apparent at higher n counts.

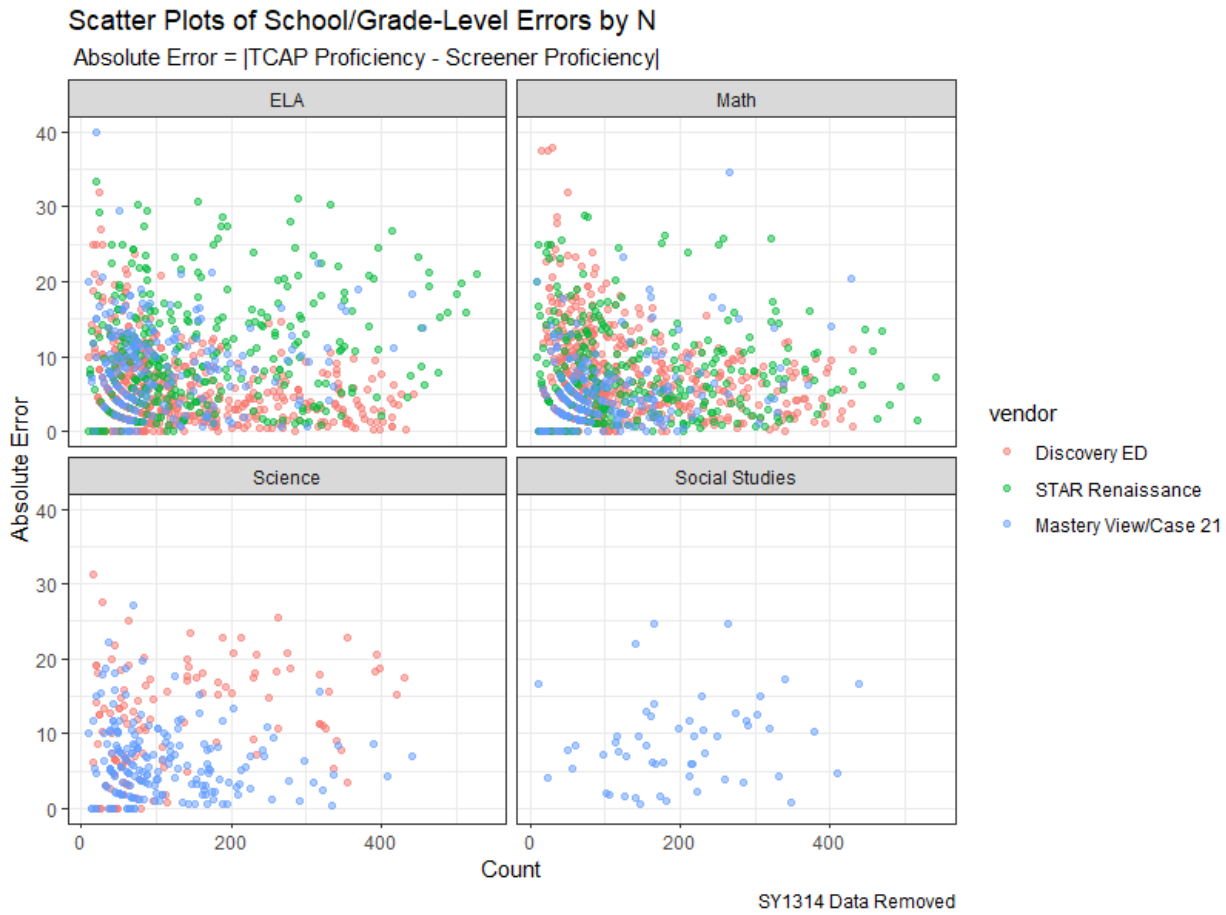

Figure 2: Scatter Plots of School/Grade-Level Benchmark Prediction Errors by Sample Size

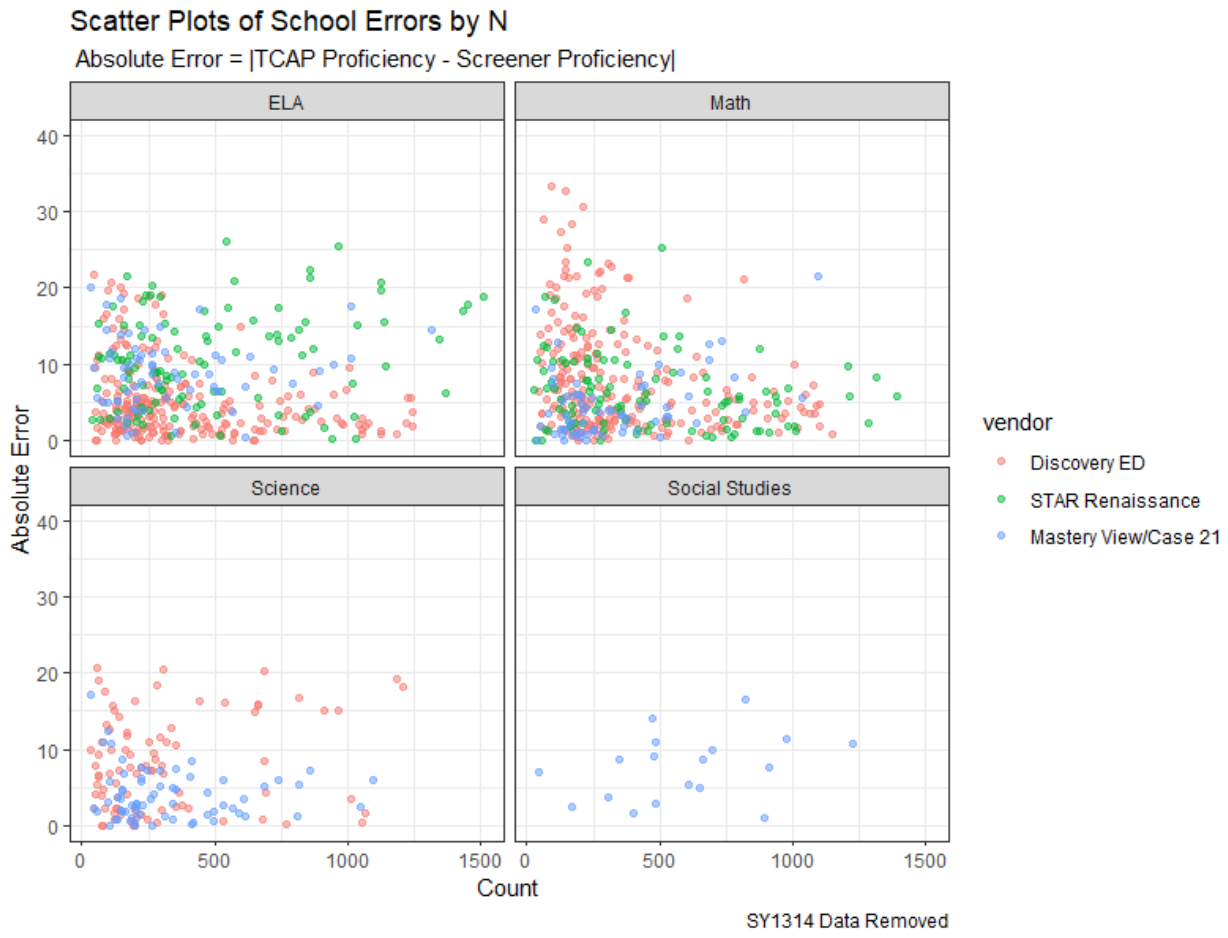

Figure 3: Scatter Plots of School-Level (not grade-level) Benchmark Prediction Errors by Sample Size

Table 10 (for comparison with Table 9) shows the median absolute errors of school-specific/grade-level/subject-specific predictions when using previous TCAP scores (instead of benchmarks). REA excluded cells with n counts less than ten from the analysis. The median absolute error using previous TCAP to predict current TCAP is practically the same as the error using a commercially available benchmark assessment. Figure 4 shows that using previous-year TCAP data aggregated by grade decreases (slightly) the prediction errors compared to benchmark data (Figure 3). The decrease in absolute error using previous TCAP data (rather than benchmark data) at the school level is more apparent at higher n counts.

Table 10: Median Pervious TCAP Absolute Errors: School Level (in Percentage Points)

TCAP Test Year		Subject	Grade				
Result Year	Basis Year		4	5	6	7	8
SY1112	SY1011	ELA	4.6	13.6	5.4	5.5	4.1
SY1213	SY1112	ELA	3.6	5.9	3.6	9.6	3.1
SY1314	SY1213	ELA	3.2	5.9	2.8	6.3	1.9
SY1415	SY1314	ELA	4.4	6.5	6.1	2.3	2.2
SY1718	SY1617	ELA	5.1	5.1	4.0	4.8	10.1
SY1819	SY1718	ELA	4.8	6.7	2.6	4.7	5.8
SY2122	SY2021	ELA	9.5	5.3	6.6	5.7	3.8
SY1112	SY1011	Math	8.2	13.6	6.8	3.8	7.8
SY1213	SY1112	Math	9.1	7.5	9.4	2.8	6.4
SY1314	SY1213	Math	11.6	9.1	4.1	6.4	7.9
SY1415	SY1314	Math	7.4	13.2	5.2	5.5	8.1
SY1718	SY1617	Math	5.4	5.6	4.0	10.4	4.3
SY1819	SY1718	Math	7.4	6.1	5.2	6.1	7.9
SY2122	SY2021	Math	5.7	4.8	3.7	3.3	5.0
SY1112	SY1011	Science	6.7	15.5	12.2	6.3	11.4
SY1213	SY1112	Science	9.8	6.5	7.9	4.1	4.0
SY1112	SY1011	Social Studies	3.8	3.3	2.1	2.7	3.1
SY1213	SY1112	Social Studies	4.5	3.8	2.2	4.3	1.8
SY1314	SY1213	Social Studies	3.2	3.1	2.0	1.8	3.0
SY1819	SY1718	Social Studies			6.9	7.6	7.0
SY2122	SY2021	Social Studies				8.0	7.0

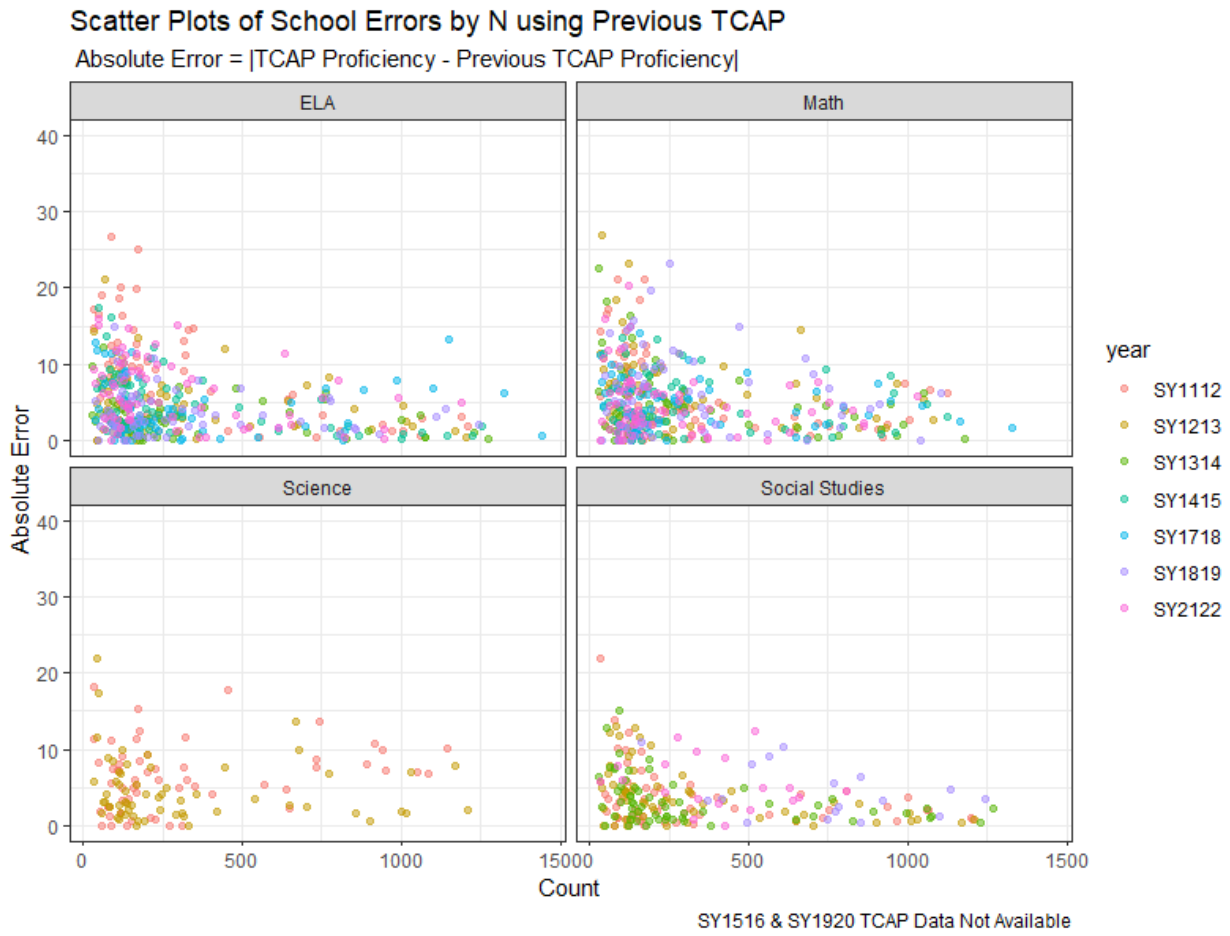

Figure 4: Scatter Plots of School-Level (not grade-level) Previous TCAP Prediction Errors by Sample Size

Conclusions & Considerations

Herman (2005) makes recommendations to ensure quality benchmark exams. Herman recommends evaluating the technical qualities of a benchmark and holding the benchmark test accountable to its purposes. REA analyzed the predictive accuracy of the benchmark assessments administered since SY1011.

The predictive accuracy of the benchmark tools is similar across vendors. Evidence presented in this study suggests that the predictive accuracy of benchmark exams is no better than using prior-year TCAP results to predict current-year outcomes. However, historic state test data does not exist for the students in early grades, students new to the district, and performance in subjects not tested under the TCAP umbrella. The TCAP assessment also does not provide an item analysis or scaled standards-level scores that teachers can use in formative practices.

The author acknowledges that predictive and criterion validity are not the only facets to evaluate the effectiveness of a benchmark assessment. KCS decision-makers should clearly define what is most important when selecting a benchmark assessment vendor. Most vendor-provided benchmarks vary in quality of content, ease of reporting, the information available to teachers, and item types.

Theoretically, the key to a quality benchmark process may not be related to predictive validity. Literature suggests that the value of benchmarks is their ability to aid teacher diagnostic processes (Herman 2005). Benchmark assessments can aid diagnostic processes if they use a variety of design elements (multiple choice, short answer, essay) to help illustrate student thought processes. Oláh (2010) suggests that multiple-choice benchmarks don't help teachers understand student thinking. Teachers tended “to interpret student errors as procedural missteps” when analyzing multiple-choice item analyses. These interpretations were “paralleled by a trend toward procedural instructional response.” Furthermore, Bancroft (2010) suggests that teachers generally find benchmark testing interrupts more valuable classroom instructions.

These findings do not mean that KCS should abandon benchmark testing. Research by Baenen (2006) suggests that frequent use of formative data is an important component of addressing the needs of under-represented students. Baenen notes the necessity of ongoing support and training to realize the benefits of formative testing. KCS can model how to combine benchmark test information with other student data (student work, traditional assessment, screening data, behavioral data, etc.) to inform instruction: Especially since benchmark tests are generally too short to provide a complete picture of student performance (Bancroft 2010).

REA's findings from this analysis highlight some pitfalls of using benchmark test data. Mid-term grade-specific school-level results should likely not be used as a precise proxy for end-of-the-year summative results. The errors in the predictions are generally larger than school improvement targets. Aggregating data to larger n counts may mitigate this problem without entirely solving it. Based on these findings, REA suggests that schools using benchmark data

for progress monitoring purposes aggregate data to the highest n count possible. REA encourages KCS staff to be cautious in using disaggregated benchmark data for any high-stakes decisions or supervisory conversations.

The results indicate that using previous TCAP performance levels to predict current TCAP performance levels is as good (or better than) benchmark assessments. The results suggest that student-level TCAP performance does not change dramatically from one test administration to the next. This finding has implications when interpreting growth data generated through the state assessment. Surface-level evidence suggests that growth, as measured by the Tennessee Value-Added Assessment System (TVAAS), may generally occur in increments that do not impact performance levels. The finding can have significant implications for goal setting using TVAAS and changes in TCAP proficiency. Additionally, if student growth occurs slowly, KCS can re-evaluate the frequency at which it administers benchmark tests. Decreasing test frequency could increase the instructional time without losing finer-grained student performance data.

The results of this study may lead to other research questions that REA may address in the future. Example research questions include:

- How do KCS teachers use the output from benchmark tests to change instruction?
- How do student-level benchmark results vary by benchmarking period (Fall, Winter, Spring)?
- How do typical student growth patterns impact student proficiency over time?
- Can student performance on current-year tests be better predicted using previous TCAP performance and other data (such as Aismweb+ scores, attendance patterns, discipline data, etc.)?

References

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

Baenen, N., Ives, S., Lynn, A., Warren, T., Gilewicz, E., & Yaman, K. (2006). Effective Practices for At-Risk Elementary and Middle School Students, 2002-03 through 2004-05. E&R Report No. 06.03. *Wake County Public School System*.

Bancroft, K. (2010). Implementing the mandate: The limitations of benchmark tests. Educational Assessment, Evaluation and Accountability, 22, 53-72.

Oláh, L. N., Lawrence, N. R., & Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody journal of education*, *85*(2), 226-245.

Appendix A: Number of students with TCAP data in consecutive years.

*Students listed in Grade 3 were retained from the previous year

TCAP Test Year		Subject	Student Grade-Level at Second Consecutive TCAP					
Result Year	Basis Year		3*	4	5	6	7	8
SY1112	SY1011	ELA	8	4020	4030	3933	3907	3903
SY1213	SY1112	ELA	12	3955	3903	3749	3780	3776
SY1314	SY1213	ELA	13	4147	4170	4020	3982	4028
SY1415	SY1314	ELA	15	4102	4139	4055	4055	3982
SY1718	SY1617	ELA	3	4108	4388	4171	4019	4028
SY1819	SY1718	ELA	5	4169	4216	4286	4169	3971
SY2122	SY2021	ELA	7	4043	3813	3664	3870	3891
SY1112	SY1011	Math	8	4026	4036	3937	3905	3182
SY1213	SY1112	Math	13	3960	3909	3743	3779	2925
SY1314	SY1213	Math	13	4150	4178	4013	3977	3078
SY1415	SY1314	Math	15	4107	4149	4055	4035	2915
SY1718	SY1617	Math	4	4149	4396	4196	3983	3017
SY1819	SY1718	Math	6	4174	4192	4260	4166	2947
SY2122	SY2021	Math	7	4055	3860	3704	3870	2808
SY1112	SY1011	Science	8	4025	4036	3933	3914	3343
SY1213	SY1112	Science	13	3955	3907	3752	3775	3789
SY1112	SY1011	Social Studies	8	4020	4028	3928	3911	3893
SY1213	SY1112	Social Studies	13	3951	3907	3727	3762	3770
SY1314	SY1213	Social Studies	13	4136	4167	4018	3951	3999
SY1819	SY1718	Social Studies				4286	4133	3943
SY2122	SY2021	Social Studies					3807	3764

Appendix B: School-level median absolute errors.

*SY1314 Discovery ED tests were intentionally designed to align to common core standards rather than Tennessee state standards.

Year	Vendor	Subject			
		ELA	Math	Science	Social Studies
SY1011	Discovery ED	2.2	6.1		
SY1112	Discovery ED	2.9	4.2		
SY1213	Discovery ED	3.1	3.0	11.3	
SY1314*	Discovery ED	10.5	16.1	2.3	
SY1415	STAR Renaissance	7.5	6.8		
SY1617	STAR Renaissance	15.5	5.1		
SY1718	STAR Renaissance	17.1	2.8		
SY1819	STAR Renaissance	1.9	1.9		
SY2122	Mastery View/Case 21	8.7	3.7	2.9	8.2