HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

2025 No. 051

School Classification Accuracy

Prepared for:	Kentucky Department of Education Office of Assessment and Accountability 300 Sower Boulevard Frankfort, KY 40601	**Prepared under:**	Contract #2400003213
Authors:	Dea Mulolli Hye-Jeong Choi Emily R. Dickinson	**Date:**	May 13, 2025

School Classification Accuracy

Table of Contents

List of Tables

List of Figures

School Classification Accuracy

Introduction

KRS 158.6455 requires the Kentucky Board of Education to create an accountability system to classify schools and districts that comply with the federal Every Student Succeeds Act (ESSA) of 2015. In Spring 2022, the Kentucky Department of Education implemented a new accountability model to meet ESSA requirements. As in previous systems, this new model uses students' state assessment scores to award points to schools for students' academic performance. Initial changes included how these points were weighted and combined with other indicators to derive school-level classifications. In Spring 2023, the model was further updated to include a score for both status and change scores on each indicator. Schools are now assigned an overall accountability score, which is a weighted composite based on status and change scores for each of the following indicators:

State Assessment Results (SAR) in Reading and Mathematics (RM). This component is based on reaching the desired level of knowledge and skills as measured on state-required academic assessments in reading and mathematics. Student performance is aggregated at the school, district, and state levels. Schools are rated based on student performance levels: Novice (0 points), Apprentice (0.5 points), Proficient (1.0 points), and Distinguished (1.25 points). Student performance is generated from the Kentucky Summative Assessment (KSA) and the Alternate KSA.

State Assessment Results (SAR) in Science (SC), Social Studies (SS), and Writing (CW). This component is based on reaching the desired level of knowledge and skills as measured on state-required academic assessments in science, social studies, and writing. Student performance is aggregated at the school, district, and state levels. Schools are rated based on student performance levels: Novice (0 points), Apprentice (0.5 points), Proficient (1.0 points), and Distinguished (1.25 points). Student performance is generated from the Kentucky Summative Assessment (KSA) and the Alternate KSA.

English Learner Progress (ELP). This component is based on an improvement in the English Language Proficiency Exam for English Learners. English learners' progress is included in the calculation using an English learner progress table.[1]

Quality of School Climate and Safety (QSCS). This component is based on measures of the school environment. Students' perception data from surveys provide a measure of the school environment. Survey questions ask students to rate aspects of their school's climate and safety on an agreement scale using questions coded such that Agree or Strongly Agree represent positive perceptions, while Disagree or Strongly Disagree represent negative perceptions. Survey items are assigned scores of 0.00 for Strongly Disagree and 33.33 for Disagree. A score of 66.66 is assigned for Agree, and 100.00 for Strongly Agree. The scores are averaged for each question to get a question score. The question scores are then averaged to create an index.

Postsecondary Readiness (PSR, high school only). This component is based on whether a student has attained the necessary knowledge, skills, and dispositions to successfully transition to the next level of his or her educational career. To demonstrate postsecondary readiness, high

[1] https://education.ky.gov/AA/Acct/Documents/ELProgress_Indicator_Tables.pdf

school students must earn a high school diploma or be classified as a Grade 12 non-graduate and meet the requirements for one type of readiness (Academic or Career).[2]

Graduation Rate (GR, high school only). This component is based on the percentage of students earning a high school diploma compared to the cohort starting in Grade 9. Kentucky uses a 4-year adjusted cohort rate and an extended 5-year adjusted cohort in accountability, which recognizes the persistence of students and educators in completing the requirements for a Kentucky high school diploma. The 4-year and 5-year rates are averaged for accountability reporting.

Table 1 presents the weighting of the accountability indicators by grade span, along with scale ranges for each indicator. State assessment results in reading and mathematics are the indicators assigned the most weight at all grade spans. The English learner progress, quality of school climate, and safety indicators are weighted the same across the grade spans and are assigned the least weight. Postsecondary readiness and graduation rates are only applied to high schools. If data for any indicator is not available, weights are redistributed proportionally to the remaining indicators to align with the weighted indicator values approved by the Kentucky Board of Education.

Table 1. Weighting of Accountability Indicators by Grade Span

Indicator	Elementary School	Middle School	High School	Scale
State Assessment Results (SAR) in Reading and Mathematics (RM)	51	46	45	0-125
State Assessment Results (SAR) in Science (SC), Social Studies (SS), and Writing (CW)	40	45	20	0-125
English Learner Progress (ELP)	5	5	5	0-140
Quality of School Climate and Safety (QSCS)	4	4	4	0-100
Postsecondary Readiness (PSR)	NA	NA	20	0-125
Graduation Rate (GR)	NA	NA	6	0-100

Since 2023, indicator scores have been calculated by combining status with change scores. Change scores are a simple subtraction of prior year status scores from current year status scores. Indicator scores are then a simple combination of status scores and change scores. Status and change receive equal weight in determining overall performance. Change and indicator scores are calculated the same way for every indicator. It is important to note that the change score measures the performance of the population of students in the school from year to year; it is not a measure of individual students' change or growth.

Individual schools are classified into one of five performance levels based on their status and change scores. Cut scores identified via a standard-setting process are applied to assign schools to one of five levels (red, orange, yellow, green, blue), with red being the lowest rating and blue being the highest rating. As an example, Figure 1 presents status and change cut scores for high schools. Table 2 presents the cut scores for overall performance for the three grade spans.

[2] https://education.ky.gov/AA/Acct/Pages/Postsecondary-Readiness.aspx

Figure 1. High School Cut Scores for Each Indicator (KDE, 2024)

High School Indicator Status Cut Scores

School Level	Indicators	Very Low	Low	Medium	High	Very High
High School Status	State Assessment Results in Reading/Mathematics	0-38.9	39.0-52.9	53.0-64.9	65.0-76.9	77.0-125
	State Assessment Results in Science/ Social Studies/Writing	0-31.9	32.0-46.9	47.0-54.9	55.0-62.9	63.0-125
	English Learner Progress	0-9.9	10.0-23.9	24.0-30.9	31.0-44.9	45.0-140
	Quality of School Climate and Safety	0-53.9	54.0-58.9	59.0-63.9	64.0-67.9	68.0-100
	Postsecondary Readiness	0-58.9	59.0-75.9	76.0-87.9	88.0-94.9	95.0-125
	Graduation	0-85.9	86.0-91.9	92.0-94.9	95.0-97.9	98.0-100

High School Indicator Change Cut Scores

School Level	Indicators	Declined Significantly	Declined	Maintained	Increased	Increased Significantly
High School Change	State Assessment Results in Reading/Mathematics	-12.1 or less	-12.0 to -4.9	-5.0 to 0.0	0.1 to 6.2	6.3 or more
	State Assessment Results in Science/ Social Studies/Writing	-11.1 or less	-11.0 to -3.6	-3.5 to 0.0	0.1 to 6.9	7.0 or more
	English Learner Progress	-13.1 or less	-13.0 to -4.1	-4.0 to 0.0	0.1 to 9.5	9.6 or more
	Quality of School Climate and Safety	-4.1 or less	-4.0 to -2.1	-2.0 to 0.0	0.1 to 3.9	4.0 or more
	Postsecondary Readiness	-5.1 or less	-5.0 to -2.1	-2.0 to 0.0	0.1 to 11.9	12.0 or more
	Graduation	-5.1 or less	-5.0 to -2.1	-2.0 to 0.0	0.1 to 2.9	3.0 or more

Table 2. Cut Scores for Overall Performance Ratings

School Level	Red	Orange	Yellow	Green	Blue
Elementary Schools	0-37.9	38.0-54.9	55.0-69.9	70.0-82.9	83.0 or more
Middle Schools	0-35.9	36.0-50.9	51.0-63.9	64.0-76.9	77.0 or more
High Schools	0-48.9	49.0-59.9	60.0-70.9	71.0-80.9	81.0 or more

The overall performance rating is a combination of the available weighted indicator scores used to calculate the overall scores. Schools may receive the same overall performance rating even if their performance on indicators varies greatly, given that the rating combines indicators (KDE, 2024). At the high school level, SAR in RM carry a weight of .45 in the overall performance calculation. To illustrate this process, consider a school with a current year's status score of 65 and a positive change score of 5. These components combine to create an indicator score of 70. When multiplied by the .45 weight, this yields a weighted indicator score of 31.5. Figure 2 demonstrates a complete calculation of the overall performance for a high school. In this example, after combining all weighted indicators, the school achieves an overall score of 72.1. According to the performance rating categories in Table 2, this score places the school in the green category, indicating strong performance.

Figure 2. Example Calculation of Overall Performance for High School level (KDE, 2024)

Overall Performance High School Example Calculation

High School	Current Year Status Score	Prior Year Status Score	Change Score (Current Status minus Prior Status)	Indicator Score (Current Status plus Change Score)	Indicator Weight	Weighted Indicator Score (Indicator Score multiplied by Indicator Weight)
State Assessment Results in Reading and Mathematics	65	60	5	70	.45	**31.5**
State Assessment Results in Science, Social Studies and Writing	55	50	5	60	.2	**12**
EL Progress	38.1	34.0	4.1	42.2	.05	**2.1**
Quality of School Climate and Safety	84.4	86.2	-1.8	82.6	.04	**3.3**
Postsecondary Readiness	88	86	2	90	.2	**18**
Graduation Rate	88	90	-2	86	.06	**5.2**
					Overall Score	**72.1**

Because overall school scores and ratings are based on a combination of indicators, there are multiple potential sources of measurement error. First, each indicator has a measurement error. Second, as change scores use both the current year's and the previous year's scores, the measurement error in the previous year can also be included in the overall school score or ratings. It is essential to determine the extent to which school classifications can be expected to be accurate. Choi et al. (2024) investigated the relationship between status and change scores in the KDE accountability system to understand the benefits and implications of introducing change alongside status as part of the accountability system. They found that, for most schools in KY, accounting for change did not impact the overall classification as status score, especially RM, tended to be the strongest predictor of overall school performance. The current study aims to identify and clarify design issues critical for ensuring that the accountability system can accurately and consistently classify schools and districts.

Reliability Issues

This section of the report discusses issues related to the reliability of the overall accountability scores, which incorporate both status and change. In general, the reliability of change scores is a function of standard deviations, each score's (prior year score and current year score) reliability, and the correlation between those two scores (Zimmerman, 2009). The higher the correlation between prior and current scores, the lower the reliability of change scores. Table 3 presents correlations between status scores from the 2023-2024 and 2024-2025 school years. The results show that correlations were quite high except for the ELP indicator: Correlations between academic achievement indicators were higher than .78, correlations for the QSCS indicator were higher than .68, and correlations for the PSR and GR indicators were .75 and .84, respectively. More detailed descriptions and additional limitations are described in the prior year report (Mulolli et al., 2024). The remainder of this section provides a brief overview of the characteristics of each accountability indicator and related limitations for the quantification of error variance in the overall score.

Table 3. Correlation Between Status Scores Across School Years

Level	Indicator	Corr.	N*	Mean	STD	Min	Max
Elementary School	SAS/RM	.93	694	63.64	16.40	13.7	103.9
	SAS/SC/SS/WR	.87	672	61.72	15.44	14.9	109.3
	ELP	.21	167	63.74	9.77	29.4	93.3
	QSCS	.80	694	77.52	5.57	62.7	97.6
Middle School	SAS/RM	.94	315	59.78	13.96	13.9	103.4
	SAS/SC/SS/WR	.90	308	55.26	13.16	9.4	101.6
	ELP	-.01	49	27.35	9.61	10.4	50.9
	QSCS	.81	315	67.75	5.83	53.5	92.0
High School	SAS/RM	.82	227	58.12	13.17	14.7	98.2
	SAS/SC/SS/WR	.79	224	50.43	11.68	10.9	76.3
	ELP	.34	44	29.37	7.80	10.2	41.4
	QSCS	.69	227	63.99	4.55	50.5	83.2
	PSR	.75	222	90.96	10.32	34.1	115.4
	GR	.84	227	94.95	3.15	82.8	100.0

Note. N is the number of schools included in both school years.

Academic Achievement Indicators

The state assessment results components of the overall accountability score are designed to recognize schools for students reaching the desired level of knowledge and skill as measured on state-required academic assessments in reading, mathematics, science, social studies, and writing. reading and mathematics achievement are combined as one indicator, and science, social studies, and writing achievement are combined as another indicator. Both are based on student academic performance on the KSA and the Alternate KSA, specifically the percentage of students classified at each performance level: Novice, Apprentice, Proficient, and Distinguished (NAPD).

It is well-documented that the amount of student classification error varies across grade/subjects and across performance categories and that overall classification error is relatively small when averaged (Crawford & Dickinson, 2022; Mulolli et al., 2024; 2025). Table 4 illustrates the average distribution error across test content areas for each student classification category in each grade level for the 2023-24 school year, which ranges from 0.25 to 1.99 (%) (Mulolli et al., 2024). Because overall accountability scores rely heavily on students' NAPD classifications, the accuracy of student classifications provides evidence to support the accuracy of school-level scores.

Table 4 demonstrates that although average levels of student misclassification may be quite low overall, they do vary in magnitude across the NAPD categories and across grade levels. Because the state assessment components of the overall score are derived from some combination of the weighted number of students scoring at each NAPD level, the same indicator score may reflect different combinations of these student classifications.

Table 4. Error Distribution for Each Student-Level Proficiency Category (%)

	Novice	Apprentice	Proficient	Distinguished
Grade 3	0.50	0.25	0.69	0.68
Grade 4	0.76	0.96	0.71	0.80
Grade 5	0.48	0.55	0.82	0.87
Grade 6	0.97	0.98	1.26	0.85
Grade 7	0.66	1.91	1.18	0.64
Grade 8	0.48	0.92	1.19	0.69
High School	0.40	0.94	0.92	0.72

Note: Values indicate the average error for each student-level proficiency category for all content areas tested at each grade level.

Table reads: The average difference between students expected to be classified as Novice and students observed to be classified as Novice in grade 3 is 0.50%.

English Learner Progress Indicator

The English Learner Progress (ELP) component of the overall accountability score is designed to recognize schools for non-native English-speaking students making progress toward becoming proficient in English. This indicator is operationalized by comparing a student's World Class Instructional Design and Assessment (WIDA) Assessing Comprehension and Communication in English State-to-State (ACCESS) or Alternate ACCESS performance (i.e., proficiency level) from last year to the current year using a table developed by KDE.[3] Based on this comparison, each tested student is assigned points, and the school indicator is calculated by averaging these points across students.

Across the grade spans, the English learner progress indicator has the second lowest weighting among the accountability indicators and is only included in the accountability calculation for schools serving English learners. In 2024, approximately 24% of Kentucky schools included the English learner progress indicator in their accountability calculation, which is 2% higher than in the previous year. If schools did not serve English learners, the English learner progress indicator weight was distributed proportionally among the remaining indicators. Eligible students who do not participate in testing receive the lowest possible proficiency level rating, which may differentially impact schools that serve high percentages of at-risk students.

The range of points for the English learner progress indicator across all grade levels for the 2023-2024 school year was 0 to 115. For the 2022-2023 school year, the range was 10.40 to 140. We do not have access to the data necessary to calculate the accuracy of Kentucky students' WIDA performance classifications. WIDA (2025) reported that both Cronbach's alpha and marginal classification accuracy were greater than .8 for speaking, listening, and reading. Cronbach's alpha and marginal classification accuracy were much lower for writing (the lower bound was about .6).

One potential concern is if the pattern of missing indicators is systematic rather than random. For example, the overall accountability score of schools not having EL indicator scores was significantly higher than that of schools having EL indicator scores ($t=7.59$, $p<0.001$). The t-test

[3] https://www.education.ky.gov/AA/Acct/Documents/ELProgress_Indicator_Tables.pdf

scores are similar to the previous years (t=6.79, p<0.001 in 2023 and t=6.18, p<0.001 in 2022), which may indicate that the effect of having more EL students on the overall score is consistent.

Quality of School Climate and Safety

Scholars defined school climate as the quality and character of school life based on patterns of people's experiences of school life and reflects norms, goals, values, interpersonal relationships, teaching and learning practices, and organizational structures (Cohen et al., 2009). The quality of school climate and safety component of the overall accountability score is designed to recognize schools for providing a safe and engaging school environment. It is measured via the Kentucky Quality of School Climate and Safety (QSCS) survey. The QSCS measures student perceptions of the school environment. Item-level scores are averaged to create a score for each student. Student scores are then averaged to create the school-level indicator score. The range of points for the QSCS indicator across all grade levels for the 2023-2024 school year was 40.80 to 100.00. For the 2022-2023 school year, the range was 59.50 to 100.00.

The QSCS has demonstrated high levels of internal consistency reliability, ranging from .90 to .94, and was found to measure climate and safety perceptions similarly for different student groups (Lee et al., 2020; Dickinson & Thacker, 2022). It is important to note that the weighting of the accountability model is designed such that the quality of school climate and safety indicators has much less influence on schools' overall scores relative to other academic indicators.

Postsecondary Readiness

The Postsecondary Readiness (PSR) component of the overall accountability score is only applicable to high schools and is designed to recognize schools for preparing students to demonstrate readiness for postsecondary success. Postsecondary readiness is an accountability indicator that relies on several different assessment instruments that may be used in various combinations within a given school. A student demonstrates postsecondary readiness by meeting a college readiness benchmark score on a college admissions examination or college placement examination, earning a "C" or higher in 3 hours of KDE-approved dual credit, meeting approved benchmarks on an Advanced Placement (AP), International Baccalaureate (IB), or Cambridge Advanced International (CAI), or another approved, nationally recognized examination, earning an approved industry certification, scoring at or above the benchmark on the Career and Technical Education (CTE) End-of-Program (EOP) assessment for articulated credit, or completing a KDE/Cabinet approved apprenticeship program.

The range of postsecondary readiness points across high schools for the 2023-2024 school year was 20.30 to 125.0, with a mean of 93.44 and a standard deviation of 14.22. The range of postsecondary readiness points across high schools for the 2022-2023 school year was 52.0 to 125.0, with a mean of 96.16 and a standard deviation of 14.87. As the percentage of students meeting benchmarks will be, in part, a function of the reliability of the particular tests used, then the level of classification error at the school level will depend on how many students were assessed with each particular test and where their scores are on the score scale in relation to the cut score.

Graduation Rates

The Graduation (GR) component of the overall accountability score is designed to recognize schools for students completing graduation requirements, and it is also only applicable for high schools. Schools and districts report graduation rates. The range of graduation points among

high schools for the 2023-2024 school year was 81.60 to 100, with a mean of 95.64 and a standard deviation of 3.62. The range of graduation points among high schools for the 2022-2023 school year was 79.30 to 100, with a mean of 94.51 and a standard deviation of 4.09.

School Classification Accuracy Calculation

School classification accuracy can be defined as the precision with which schools are placed into performance categories within an accountability system. As the classification is based on composite scores, school classification accuracy calculations require several statistics, including reliability and standard measurement error for each indicator and correlations between indicators. The formula for the reliability of composite scores is a function of the reliability coefficients of the components of the composites, and the dispersions, intercorrelations, and respective weights of those component scores (Mosier, 1943). Having multiple classification categories adds more complications. Further, in Kentucky's accountability model, each indicator is combined, thus contributing to a school's overall score. As such, the complexity of Kentucky's accountability model does not allow for a straightforward quantification of reliability or for a calculation of error variance for the composite scores.

However, as a demonstration, we roughly calculated school classification accuracy for all those schools that had scores for all indicators for the 2023-2024 school year. We first estimated reliability and standard of measurement for each indicator based on historical data. We consulted multiple technical resources, including the KDE technical report (Pearson, 2024), the Human Resources Research Organization's (HumRRO's) QSCS report (Dickinson & Thacker, 2022), ACT technical documentation (ACT, 2024), and the WIDA assessment framework (WIDA, 2025). Then we computed the correlations between indicators for those schools. Note that we used the ACT as a proxy for postsecondary success, given that the ACT is the test of choice from KDE.[4] Table 5 presents all the necessary statistics for our classification accuracy calculation, including reliability, Standard Error of Measurement (SEM), and weight for each indicator. The table also shows the correlations between indicators. It should be noted that for this demonstration, we utilized performance scores rather than performance ratings.

The reliability of non-academic indicators tended to be higher than that of academic indicators, but the weights for the academic indicators are higher than those of the non-academic indicators. As expected, correlations between academic indicators were very high for all grade spans, ranging above .73. These academic indicators were somewhat correlated with QSCS in elementary and middle schools, ranging from .54 to .64. In high school, the academic indicators were somewhat related to graduation rates, near .56.

Using those statistics and Mosier's (1943) formula, we computed the composite score reliability; the composite score reliabilities were 0.936 for elementary school, 0.931 for middle school, and 0.918 for high school. These can be interpreted as demonstrating very high reliability. Next, using the composite score reliability, we calculated the probability of classification accuracy for each school across all performance rating categories.

[4] https://www.education.ky.gov/AA/Assessments/Pages/ACT.aspx

Table 5. Hypothetical Statistics for Classification Accuracy Calculation

Elementary School (N=187)

Indicator	Reliability	SEM	Weight	Correlation SAS/RM	Correlation SAS/SC/SS/WR	Correlation ELP	Correlation QSCS	Correlation PSR
SAS/RM	.88	6.40	.51				N/A	N/A
SAS/SC/SS/WR	.75	6.08	.40	.87			N/A	N/A
ELP	.70	8.76	.05	.26	.22		N/A	N/A
QSCS	.90	3.16	.04	.64	.60	.15	N/A	N/A

Middle School (N=55)

Indicator	Reliability	SEM	Weight	Correlation SAS/RM	Correlation SAS/SC/SS/WR	Correlation ELP	Correlation QSCS	Correlation PSR
SAS/RM	.85	6.20	.46				N/A	N/A
SAS/SC/SS/WR	.86	6.20	.45	.92			N/A	N/A
ELP	.70	8.76	.05	-.04	-.03		N/A	N/A
QSCS	.90	3.12	.04	.54	.57	.13	N/A	N/A

High School (N=53)

Indicator	Reliability	SEM	Weight	Correlation SAS/RM	Correlation SAS/SC/SS/WR	Correlation ELP	Correlation QSCS	Correlation PSR
SAS/RM	.85	6.20	.45					
SAS/SC/SS/WR	.82	6.70	.20	.73				
ELP	.70	8.76	.05	.02	-.18			
QSCS	.90	3.16	.04	.30	.36	.02		
PSR	.97	1.00	.20	.25	.11	.10	.00	
GR	.70	4.38	.06	.56	.53	.00	.31	.34

Table 6 presents a summary of classification accuracy. The overall classification accuracy means were higher than .78 and medians were higher than .81, suggesting generally robust classification. As the classification accuracy patterns are similar across grade span, we will discuss elementary school results here.

Table 6. Summary Statistics for Classification Accuracy

School Level	N	Mean	Median	Minimum	Maximum
Elementary	187	0.79	0.82	0.50	1.000
Middle	55	0.79	0.81	0.51	1.000
High	53	0.78	0.82	0.39	1.000

Figure 3 illustrates the classification accuracy for each school in our sample, with schools represented along the x-axis and their corresponding classification accuracy probabilities on the y-axis. A horizontal red reference line at 0.75 indicates the arbitrary threshold for classification accuracy.

Figure 3. Classification Accuracy Probability for Each School

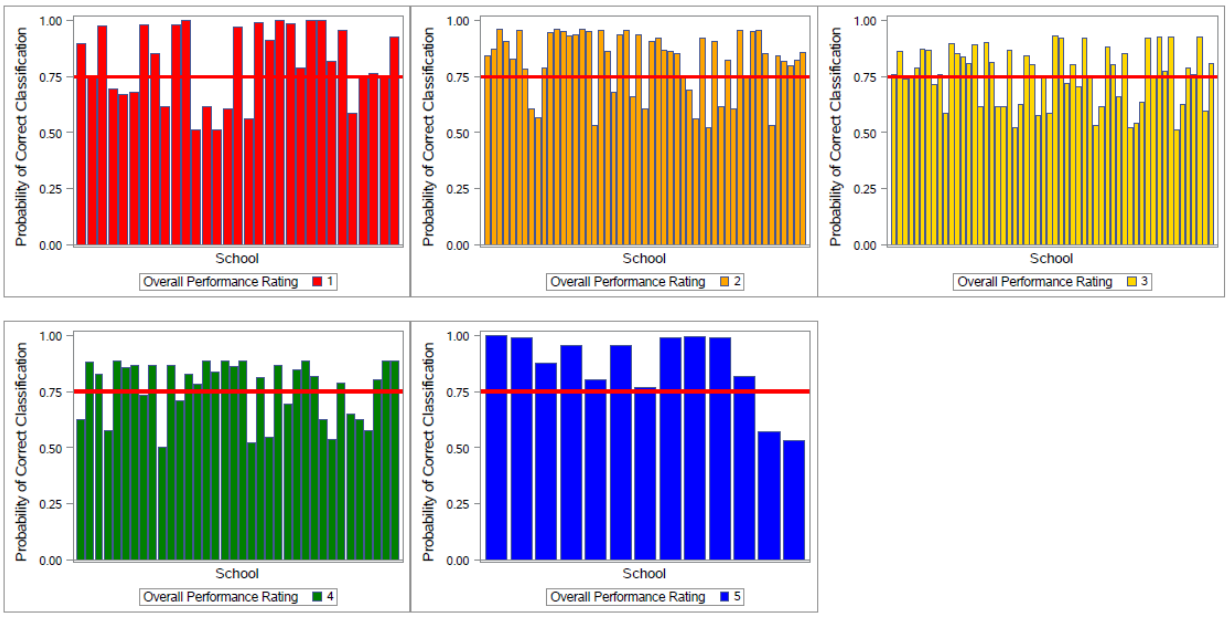
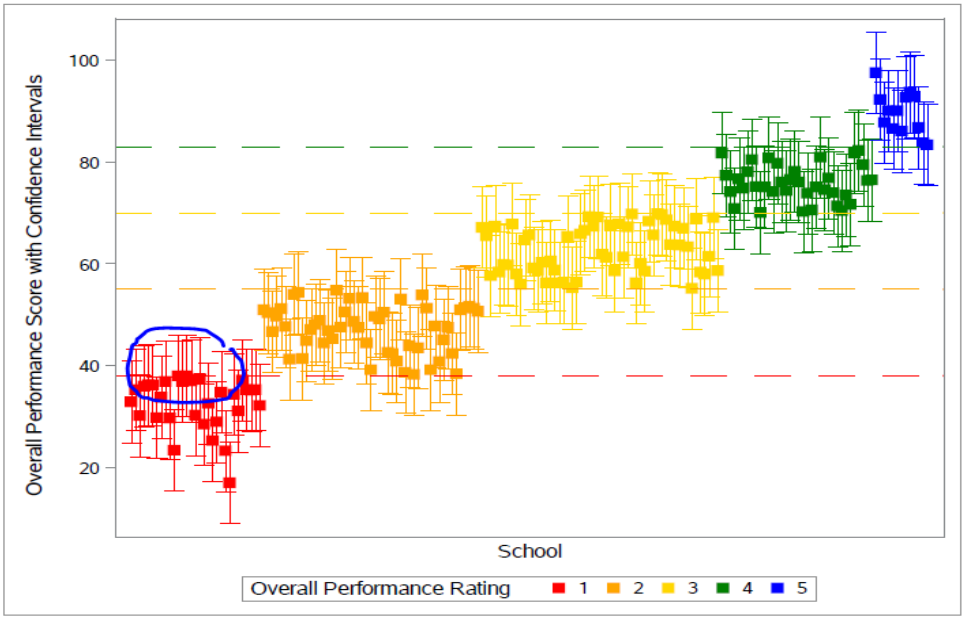


Figure 4 displays the 95% confidence intervals for each school's performance score, with horizontal lines demarcating the cut scores between rating categories. Notably, when the overall performance score was near the cut point, the confidence intervals could cover two adjacent rating categories. For example, see the school in the blue circle in the figure: When the overall performance score was 38, the confidence intervals covered 29.87 to 45.93. This school could be classified as either level 1 (probability=.51) or level 2 (probability=.49). These results confirmed that schools with overall performance scores near category cut points demonstrated lower classification accuracy, with confidence intervals frequently spanning adjacent rating categories.

Figure 4. School's Overall Performance Score with Confidence Interval



Validity Issues

This section of the report will discuss issues related to the validity of overall accountability scores. Of particular interest are the relations between the component scores and the overall scores, and the associated issues related to the interpretability of overall scores to stakeholders.

Schools are classified based on their overall accountability score. Table 7 presents the range, mean, and standard deviation of overall accountability scores for each school level (elementary, middle, and high schools). Table 8 presents the same descriptive statistics for each performance level within each school level. As shown in Table 8, Level 5 has the largest score range at the elementary school level (32.9 points). In contrast, Level 4 has a smaller range at the elementary school level (12.7points), while Level 3 has the smallest range at the middle school level (12.7 points). At the high school level, Level 1 has the largest score range (26.0 points), whereas Level 4 has the smallest score range (9.6 points). Those results are consistent with the previous year's results.

Table 7. Descriptive Statistics for Overall Accountability Scores

School Level	Min	Max	Mean	STD
Elementary (N=696)	14.2	115.9	64.2	17.3
Middle (N=316)	5.7	103.2	59.1	14.7
High (N=227)	22.9	97.5	66.5	11.9

Table 8. Overall Accountability Score Associated with Each Accountability Classification

School Level	Performance Rating	N	Min	Max	Range	Mean	STD
Elementary	1	58	14.2	37.9	23.7	30.8	5.9
Elementary	2	138	38.2	54.9	16.7	48.0	4.7
Elementary	3	233	55.0	69.9	14.9	62.8	4.4
Elementary	4	172	70.0	82.7	12.7	75.6	3.7
Elementary	5	95	83.0	115.9	32.9	91.1	6.8
Middle	1	23	5.7	35.4	29.7	29.5	6.5
Middle	2	63	36.4	50.8	14.4	45.3	4.3
Middle	3	112	51.2	63.9	12.7	57.7	3.8
Middle	4	90	64.0	76.9	12.9	70.4	3.7
Middle	5	28	77.1	103.2	26.1	84.3	6.7
High	1	17	22.9	48.9	26.0	41.7	7.8
High	2	44	49.1	59.9	10.8	55.7	2.9
High	3	88	60.1	70.9	10.8	66.0	2.9
High	4	53	71.1	80.7	9.6	75.0	2.7
High	5	25	81.2	97.5	16.3	86.0	4.1

Note. Min=Minimum; Max=Maximum; STD=Standard Deviation

Because the overall accountability score combines multiple indicator scores, a key piece of validity evidence is documenting how well each component differentiates between performance levels. One way that can be done is by analyzing the distribution of the component scores within each level and the extent to which there is an overlap in component scores across the levels. This analysis consists of calculating descriptive statistics for each performance level within each grade span (e.g., elementary schools classified as level 1, elementary schools classified as level 2, etc.).

Tables 9-11 present the number of schools at each level within each grade span, along with the minimum, maximum, and mean number of points scored, the range of points scored, and the standard deviation of points scored for each accountability component. For example, Table 9 shows 58 elementary schools classified as level 1 on the reading and mathematics assessment results component, which is higher than the 32 schools last year. Among those schools for this year, the lowest score on this accountability component was 12.3, and the highest was 45.5, with a mean of 28.7.

The range of points for the state assessment results in reading and mathematics indicators across all grade levels for the 2023-2024 school year was 4.00 to 125.0. The lower bound is lower than in the 2022-2023 school year, which was 10.3, while the upper bound is the same. The range of points for the state assessment results in science, social studies, and writing indicators across all grade levels for the 2023-2024 school year was 0 to 125.0, where the lower bound is also smaller than in the previous year, which was 3.3, while the upper bound is the same.

Tables 9-11 also include Cohen's d. One straightforward way to compare group score distributions is to calculate a standardized mean difference score (Cohen's d) of adjacent

categories. Cohen's *d* is interpreted as the difference in means presented in standardized units and can be evaluated using the following benchmarks (Cohen, 1988):

- Less than 0.2 = slight effect
- 0.2 - 0.49 = small effect
- 0.5 - 0.79 = moderate effect
- Greater than 0.8 = large effect

Cohen's *d* indicates the effect sizes for KSA academic performance indicators (i.e., SAS/RM and SAS/SC/SS/CS) tended to be large across all grades and all level comparisons. Cohen's *d* indicates a small to large effect size for elementary and middle school QSCS and a slight to large effect on high school QSCS. For ELP, Cohen's *d* indicates a slight to small effect size for elementary schools and a small to moderate effect for middle and high schools. For high school, the effect size for PSR or graduation rate varies from slight effect to large effects across accountability levels. For example, for both indicators, the mean difference between the lowest and the second lowest rating was large, and the effect size was large. In contrast, the effect size for the highest rating of both indicators was slight to small effect. The overall effect sizes are similar to those observed in the previous year (Mulolli et al., 2024).

Table 9. Descriptive Statistics of Points Values of Overall Accountability Score Components by Classification Level: Elementary Schools

Classification	Indicator	N	Min	Max	Mean	STD	Range	*d*
1	SAS/RM	58	12.3	45.5	28.7	8.1	33.2	
2	SAS/RM	138	31.0	63.0	47.8	6.5	32.0	2.7
3	SAS/RM	233	41.3	81.9	63.3	7.0	40.6	2.3
4	SAS/RM	172	63.0	91.2	76.9	6.0	28.2	2.1
5	SAS/RM	95	74.9	125.0	92.2	9.0	50.1	2.1
1	SAS/SC/SS/CW	58	9.9	46.0	28.1	7.7	36.1	
2	SAS/SC/SS/CW	138	21.1	66.2	45.4	7.7	45.1	2.2
3	SAS/SC/SS/CW	227	37.7	82.0	60.7	8.3	44.3	2.0
4	SAS/SC/SS/CW	168	52.0	97.7	73.6	7.7	45.7	1.7
5	SAS/SC/SS/CW	90	70.6	125.0	90.2	9.8	54.4	1.9
1	ELP	31	0.0	87.4	54.7	16.6	87.4	
2	ELP	51	18.5	82.8	56.1	14.7	64.3	0.1
3	ELP	56	16.1	115.0	62.2	21.1	98.9	0.4
4	ELP	36	26.8	112.8	66.5	19.8	86.0	0.2
5	ELP	13	51.1	112.6	75.8	14.7	61.5	0.6
1	QSCS	58	58.4	90.1	70.6	6.1	31.7	
2	QSCS	138	61.4	91.6	73.5	5.7	30.2	0.5
3	QSCS	233	16.1	115.0	62.2	21.1	98.9	0.6
4	QSCS	172	68.3	100.0	81.2	6.8	31.7	0.7
5	QSCS	95	65.4	100.0	84.8	8.0	34.6	0.5

Note. Min = Minimum; Ma x= Maximum; STD = Standard Deviation; *d* = Cohen's *d* for adjacent groups.

Table 10. Descriptive Statistics of Points Values of Overall Accountability Score Components by Classification Level: Middle Schools

Classification	Indicator	N	Min	Max	Mean	STD	Range	*d*
1	SAS/RM	23	8.0	40.0	29.6	7.2	32.0	
2	SAS/RM	63	32.1	64.9	47.3	6.9	32.8	2.5
3	SAS/RM	112	43.2	72.9	59.2	6.3	29.7	2.0
4	SAS/RM	90	59.6	85.2	72.8	5.6	25.6	2.2
5	SAS/RM	28	69.3	114.0	85.8	9.3	44.7	2.2
1	SAS/SC/SS/CW	23	0.0	39.0	26.4	8.5	39.0	
2	SAS/SC/SS/CW	63	22.7	54.1	41.9	6.6	31.4	2.3
3	SAS/SC/SS/CW	112	44.3	70.5	55.7	5.4	26.2	2.3
4	SAS/SC/SS/CW	87	59.6	85.2	72.8	5.6	25.6	2.2
5	SAS/SC/SS/CW	27	69.3	114.0	85.8	9.3	44.7	2.4
1	ELP	8	0.0	66.2	32.9	19.0	66.2	
2	ELP	20	22.7	54.1	41.9	6.6	31.4	0.1
3	ELP	18	0.0	96.7	31.5	25.8	96.7	0.1
4	ELP	6	0.0	63.9	27.9	22.1	63.9	0.1
5	ELP	3	19.7	53.1	34.6	17.0	33.4	0.4
1	QSCS	23	44.5	88.3	62.6	8.2	43.8	
2	QSCS	63	53.9	100.0	66.5	8.7	46.1	0.4
3	QSCS	112	54.3	87.6	88.0	6.0	33.3	0.3
4	QSCS	90	55.8	100.0	71.9	7.8	44.2	0.5
5	QSCS	28	61.3	91.8	72.2	7.1	30.5	0.1

Note. Mi n= Minimum; Max = Maximum; STD = Standard Deviation; *d* = Cohen's *d* for adjacent groups.

Table 11. Descriptive Statistics of Points Values of Overall Accountability Score Components by Classification Level: High Schools

Classification	Indicator	N	Min	Max	Mean	STD	Range	*d*
1	SAS/RM	17	4.0	37.0	25.9	8.7	33.0	
2	SAS/RM	44	19.8	80.4	58.1	7.9	60.6	1.8
3	SAS/RM	88	41.2	80.4	58.1	7.9	39.2	1.8
4	SAS/RM	53	50.9	88.5	70.4	8.2	37.6	1.6
5	SAS/RM	25	67.5	106.5	85.7	8.1	39.0	1.8
1	SAS/SC/SS/CW	17	0.00	57.3	28.5	14.1	57.3	
2	SAS/SC/SS/CW	44	21.7	87.5	41.1	14.3	65.8	0.8
3	SAS/SC/SS/CW	88	15.1	82.9	49.3	11.7	67.8	0.7
4	SAS/SC/SS/CW	53	41.7	85.3	58.1	10.2	43.6	0.8
5	SAS/SC/SS/CW	25	57.1	91.5	71.1	10.1	34.4	1.3
1	ELP	10	10.0	67.6	32.3	17.2	57.6	
2	ELP	13	0.0	46.5	28.6	14.9	46.5	0.3
3	ELP	21	6.7	67.9	31.9	12.8	61.2	0.3
4	ELP	8	18.2	51.5	34.1	10.7	33.3	0.1
5	ELP	1	-	-	-	-	-	-
1	QSCS	17	53.3	72.0	62.5	4.6	18.7	
2	QSCS	44	54.3	72.6	63.3	4.3	18.3	0.1
3	QSCS	88	40.8	84.6	65.2	5.9	43.8	0.4
4	QSCS	53	54.0	100.0	67.3	7.4	46.0	0.2
5	QSCS	25	56.8	98.2	70.6	9.0	41.4	0.5
1	PSR	16	20.3	95.5	74.7	18.4	75.2	
2	PSR	44	39.8	120.8	87.7	16.1	81.0	0.7
3	PSR	88	67.2	125.0	93.6	10.5	57.8	0.5
4	PSR	52	76.5	125.0	99.4	11.2	48.5	0.6
5	PSR	24	85.4	118.7	103.0	9.0	33.3	0.3
1	GR	17	81.6	100.0	91.6	5.7	18.4	
2	GR	44	86.1	100.0	95.3	3.6	13.9	0.7
3	GR	88	83.6	100.0	95.7	3.5	16.4	0.2
4	GR	53	90.8	100.0	96.6	2.5	9.2	0.3
5	GR	25	90.7	100.0	96.6	2.5	9.3	0.1

Note. Min = Minimum; Max = Maximum; STD = Standard Deviation; *d*= Cohen's *d* for adjacent groups.

Visual depictions of the distributions of component scores are another useful way to compare how the performance levels differ. Figures 5 through 10 depict the stair-step pattern between overall performance and each component of the overall accountability score, except for the ELP indicator. The boxes in the plot depict the interquartile range, or the middle 50% of scores, while the lines extending below and above the box depict the lower and upper quartiles, respectively. The circles that appear beyond the vertical lines depict outliers or extreme values. For the assessment results indicators, in particular, the interquartile ranges of the lower classification

levels tend to fall at or below the 25th percentile of the adjacent higher classification levels. There is more overlap among the remaining indicators across the accountability classifications.

Figure 5. Ranges of Reading and Math Indicator Scores Within Overall Classifications

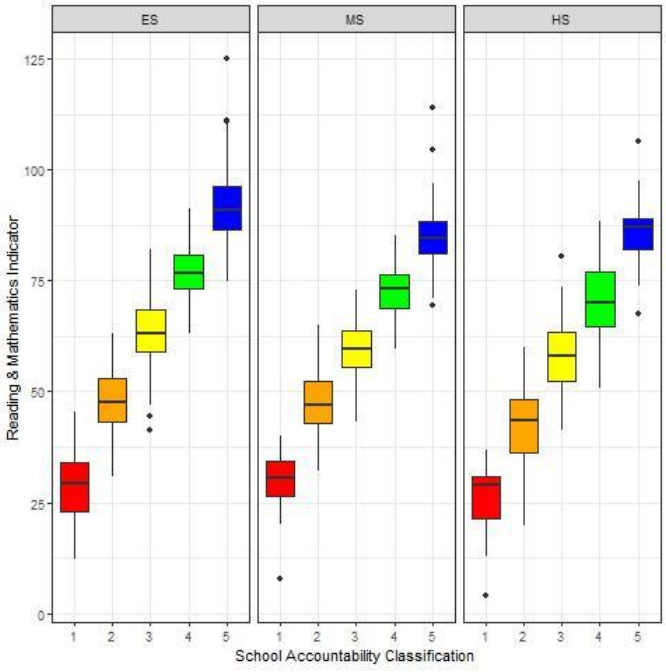


Figure 6. Ranges of Science, Social Studies, and Writing Indicator Scores Within Overall Classifications

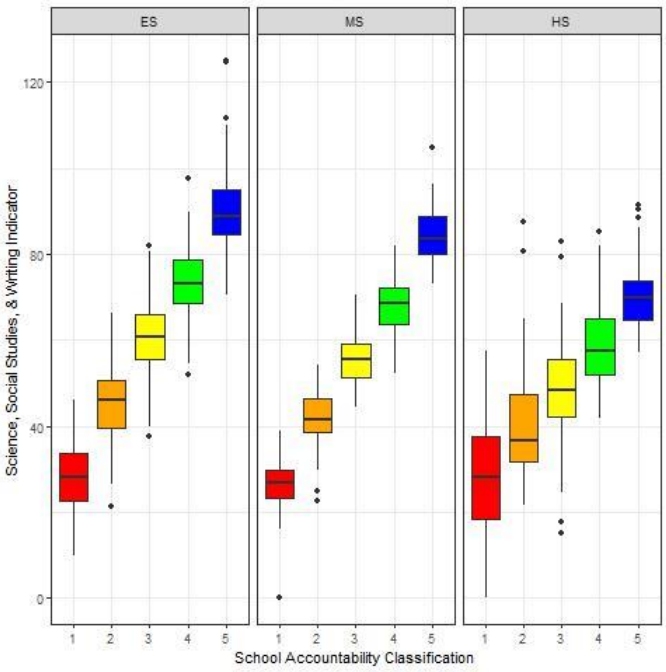


Figure 7. Ranges of English Learner Progress Indicator Scores Within Overall Classifications

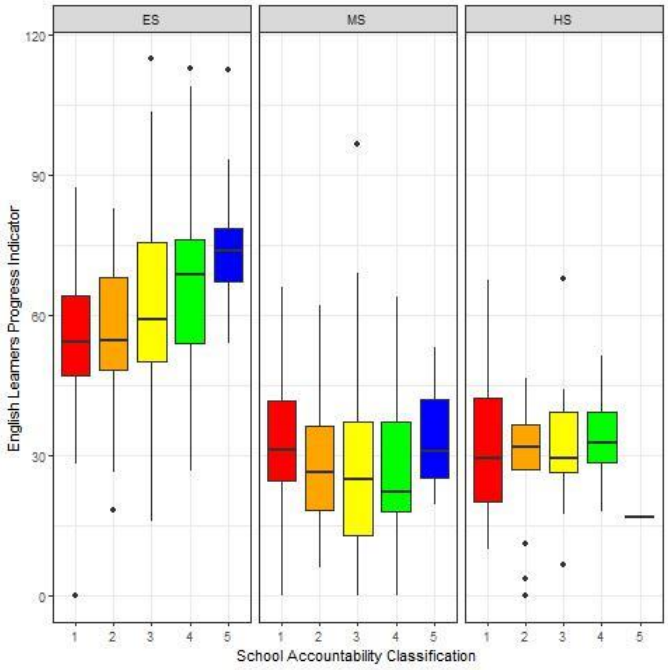


Figure 8. Ranges of Climate and Safety Indicator Scores Within Overall Classifications

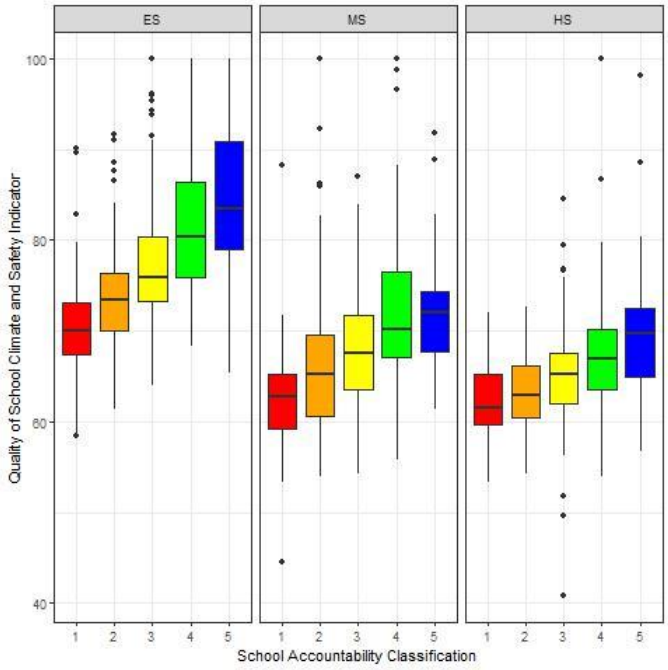


Figure 9. Ranges of Postsecondary Readiness Indicator Scores Within Overall Classifications

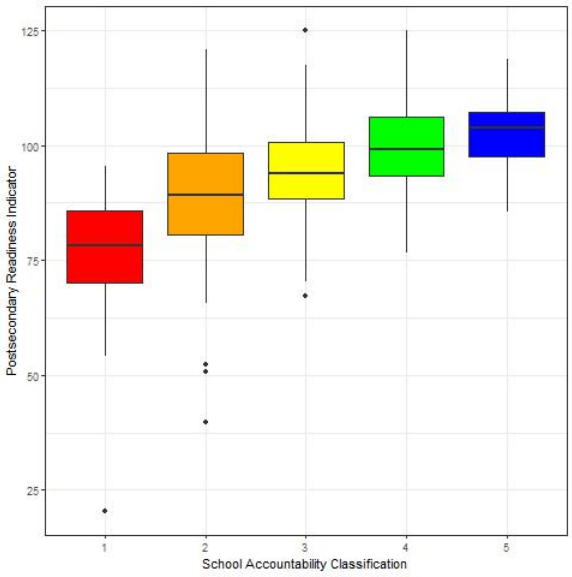
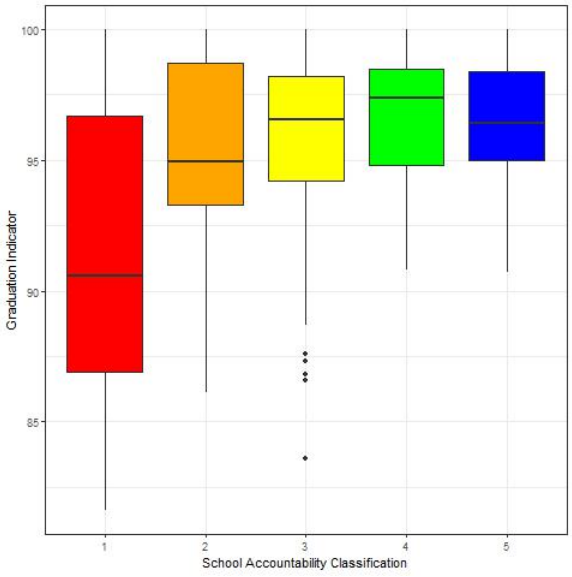


Figure 10. Ranges of Graduation Rate Indicator Scores Within Overall Classifications



Lastly, Table 12 presents descriptive statistics of Change score including mean, standard deviation, minimum and maximum. Change score can be negative and positive. Except for Elementary ELP, Change scores for all other indicators were positive indicating school status scores increased in 2023-2024. The table also includes t-test results; the significant positive t-test result indicates Change score was statistically significantly different from zero. KDE may wish to further explore the actions taken by the schools which showed the highest change score over the year to better understand what interventions and/or supports were associated with such change, or if particular schools could benefit from specialized supports.

Table 12. Descriptive Statistics of Change Scores

Indicator	N	Mean	STD	Min	Max	t-test	p-val*
Elementary							
SAS/RM	696	1.04	6.30	-25.9	32.1	4.37	**<.0001**
SAS/SC/SS/CS	674	0.20	7.74	-28.7	33.6	0.67	0.506
ELP	167	-3.00	12.30	-37.5	40.6	-3.15	**0.002**
QSCS	696	0.38	3.17	-8.6	12.4	3.19	**0.002**
Middle							
SAS/RM	316	1.08	4.95	-15.2	21.1	3.88	**0.000**
SAS/SC/SS/CS	309	1.36	5.94	-16.5	21.5	4.03	**<.0001**
ELP	49	3.09	12.43	-14.0	45.8	1.74	0.088
QSCS	316	1.02	3.38	-9.0	16.9	5.37	**<.0001**
High							
SAS/RM	227	0.55	7.86	-28.7	29.7	1.05	0.297
SAS/SC/SS/CS	224	0.17	7.38	-26.2	24.9	0.34	0.736
ELP	44	2.21	8.60	-13.0	26.7	1.70	0.096
QSCS	227	1.72	3.29	-9.7	16.8	7.85	**<.0001**
PSR	222	2.50	7.62	-28.3	33.3	4.89	**<.0001**
GR	227	0.69	1.77	-5.9	7.1	5.84	**<.0001**

* Bolded entries indicate statistical significance at $p < 0.05$.

Discussion

The current accountability system includes various factors to evaluate schools' efforts to improve student achievement. As with the previous accountability model, the overall accountability score still relies most heavily on student-level performance classifications based on academic assessment performance. The accountability indicators for which data are available have been demonstrated to show high levels of reliability, thereby supporting that the system is designed to classify schools accurately.

Utilizing historical data, we posited several statistics to estimate the composite score reliability and school classification accuracy for the high schools that had complete data across all six indicator scores. While the results indicated that the classification accuracy was robust, the results also showed an inherent challenge in precisely classifying schools whose performance scores are near rating category cut scores.

The complexity of the model, however, does not allow for a straightforward quantification of reliability or for a calculation of error variance for the composite scores. Given that there are limitations to the quality of reliability evidence at the aggregate level, it is even more important to identify evidence to support the validity of school classifications.

For schools classified at the highest levels, it is important to verify that they are performing at relatively high levels among all the indicators included. Otherwise, this might call into question the interpretability and utility of the overall performance ratings for key stakeholders. At the

middle levels of the overall rating scale, schools would be expected to have more of a mix of performance on the various indicators, and schools at the lowest rating level would be expected to be performing relatively low on most indicators. KDE applies a series of cut scores to classify schools on each accountability indicator, providing schools with a more robust depiction of their relative strengths and areas for improvement. The present study generally found the expected pattern among the indicator scores. This, taken into consideration with recent standards validation work conducted by HumRRO (Mulolli et al., 2025) supports the validity of Kentucky's school-level accountability classifications.

As of the 2022-2023 school year, KSA's accountability is based not only on a school's current status but also on the amount of change schools have experienced in each component since the previous school year. Including a change score in the accountability calculation presumably allows all schools a better chance to demonstrate their improvement. This also introduces more complexity to the model that further exacerbates estimating the accuracy of school classifications. Change under the Kentucky model is evaluated based on a school's rating in the current year relative to its prior year's rating. Thus, the overall rating reflects a compounding of the classification error from each of the included years for each accountability component as well as a confounding of cohorts. On the other hand, accounting for change may enhance the validity of school classifications by recognizing the adjustments that schools make from year to year in response to feedback from the system. School ratings improve when students' scores increase for any indicator across years, and their ratings may decline if students' scores decline. This combination of status and change may help schools better understand how their efforts toward improving student learning play out in the accountability system. It is also possible that monitoring improvement may have a motivating effect on educators.

Quantifying error variance for both status and change is most complicated for the postsecondary readiness indicator. Schools may choose from a menu of measurement options when reporting students' postsecondary readiness. This yields the possibility that a school's prior year and current year postsecondary readiness indicator scores are each based on a different combination of assessments. While offering multiple options for measuring postsecondary readiness supports the validity of these indicator scores (by allowing students to choose a readiness indicator that matches their postsecondary plans), it introduces further complexity to quantifying the accuracy of school classifications. It should also be concerning if lower-performing schools and higher-performing schools show different patterns of the indicators of post-secondary readiness (e.g., higher performers using mostly college entrance exams while lower performers use mostly Career and Technical Education [CTE] exams). The ways schools meet this requirement should be monitored, in addition to the overall results, to ensure that students have an equitable opportunity to demonstrate readiness.

Future school-level classification accuracy research conducted by HumRRO will examine school classification accuracy more closely. Such a study would simulate various options for scoring in each category and examine the classification accuracy of test cases similar to those experienced by schools in Kentucky. Estimations of accuracy could then be generated based on a continuum of pathways to school performance categories. For example, it would be possible to compare the accuracy of a school with relatively static, but high, indicator scores with a school in the same category with lower indicator scores, but that improved substantially on multiple indicators. This would not yield an accuracy estimate for each school but could provide context for interpreting school accountability fluctuations from year to year (e.g., how much variability Kentucky assumes is due to measurement error versus true changes in school performance).

Currently, eligible students who do not participate in testing are assigned the lowest possible proficiency level rating for accountability purposes. KDE should consider monitoring the characteristics of these students to determine any patterns. It is possible, for example, that schools serving higher percentages of at-risk students are differentially impacted by this scoring policy.

We also continue to recommend investigating the impact of accountability designations, and changes in designation, on schools. This could eventually require school visits but could begin using extant data. For example, if a school's designation drops (e.g., going from blue to yellow), does that decline impact the results of the climate and safety survey? It is important to document school's reactions to accountability designations, positive and negative, to determine the effectiveness of school-level improvement efforts, and to guard against unintended negative consequences for students. Monitoring how accountability results are interpreted and how they impact schools and students is vital to ensuring the validity and fairness of the accountability system.

References

ACT Inc. (2024). *ACT technical report.* Retrieved March 28, 2025, from https://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.pdf

Choi, H-J., Dickinson, E. R. & Thacker, A.A. (2024) *Exploring the relations among change, status, and overall performance in Kentucky's school accountability system.* Human Resources Research Organization.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge Academic.

Cohen, J., Mccabe, E. M., Michelli, N. M., & Pickeral, T. (2009). School climate: Research, policy, practice, and teacher education. *Teachers College Record (1970)*, *111*(1), 180–213. https://doi.org/10.1177/016146810911100108.

Crawford, B. F., & Dickinson, E. R. (2022). *The accuracy of students' novice, apprentice, proficient, and distinguished classifications for the 2022 Kentucky Summative Assessment (KSA) tests* (2022 No. 112). Human Resources Research Organization.

Dickinson, E. R., & Thacker, A. A. (2022). *Analysis of the 2022 Quality of School Climate and Safety (QSCS) Survey*. Human Resources Research Organization.

Hoffman, R. G., & Dickinson, E. R. (2005). *The accuracy of school classification for the 2004 accountability cycle of the Kentucky Commonwealth Accountability Testing System*. (FR-05-26). Human Resources Research Organization.

Kentucky Department of Education. (2024). Kentucky's Accountability System. Kentucky Department of Education. Retrieved March 28, 2025, from https://www.education.ky.gov/AA/distsupp/Documents/Comprehensive_Overview_Kentucky_Accountability_System.pdf

Lee, J. J., Dickinson, E. R., & Thacker, A. A. (2020). *The quality of school climate and safety survey: Confirmatory factor analysis study*. Human Resources Research Organization.

Mosier, C. I. (1943). On the reliability of a weighted composite. *Psychometrika, 8(*3), 161-168. https://doi.org/10.1007/BF02288700.

Mulolli, D., Dickinson, E. R., & Thacker, A. A. (2025). *Standards validation for Kentucky's accountability system* (2025 No. 023). Human Resources Research Organization.

Mulolli, D., Dickinson, E. R., & Thacker, A. A. (2025). *The accuracy of students' novice, apprentice, proficient, and distinguished classifications for the 2024 Kentucky. Summative Assessment (KSA) tests* (2025 No. 004). Human Resources Research Organization.

Mulolli, D., Choi, H-J., & Dickinson, E. R. (2024). *School classification accuracy: Issues for reliability and validity*. Human Resources Research Organization.

Pearson (2024). *Kentucky summative assessments 2023-2024 technical manual.* Retrieved March 28, 2025, from

https://www.education.ky.gov/AA/Reports/Documents/KY1161509_KY_SP25_TechReport.pdf

WIDA. (2025). *ACCESS for ELLs Interpretive Guide for Score Reports Grades K-12*. Board of Regents of the University of Wisconsin System. Retrieved March 28, 2025, from https://wida.wisc.edu/sites/default/files/resource/Interpretive-Guide.pdf

Zimmerman, D. W. (2009). The reliability of difference scores in populations and samples. *Journal of Educational Measurement*, 46(1), 19-42.