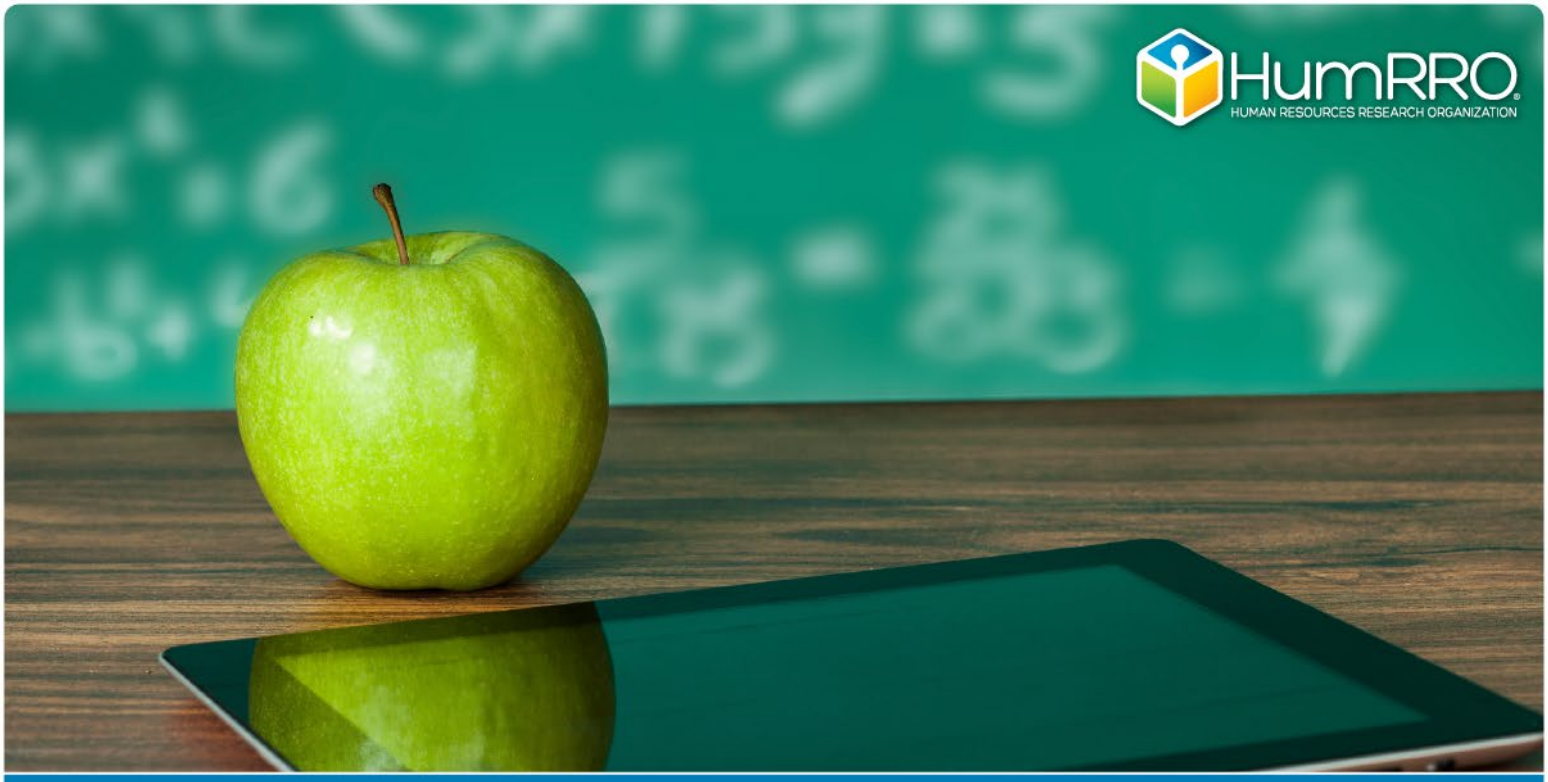


HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

2023 No. 059

School Classification Accuracy: Issues for Reliability and Validity

Prepared for: Kentucky Department of Education
Office of Assessment and Accountability
300 Sower Boulevard
Frankfort, KY 40601

Prepared under: Contract #1900004339

Authors: Hye-Jeong Choi
Emily R. Dickinson

Date: May 15, 2023

School Classification Accuracy: Issues for Reliability and Validity

Table of Contents

List of Tables

Table of Contents (Continued)

List of Figures

School Classification Accuracy: Issues for Reliability and Validity

Introduction

KRS 158.6455 requires the Kentucky Board of Education to create an accountability system to classify schools and districts that complies with the federal Every Student Succeeds Act of 2015 (ESSA). In Spring 2022, the Kentucky Department of Education implemented a new accountability model designed to meet ESSA requirements. Like previous systems, this new model uses students' state assessment scores to award points to schools for student academic performance. What has changed is the way in which these points are weighted and how they are combined with other indicators to derive school-level classifications.

Schools are assigned an overall accountability score, which is a weighted composite based on the following indicators:

- **State Assessment Results in Reading and Mathematics**. 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 to school, district and state levels. Schools are rated based on student performance levels, Novice (0 points); Apprentice (0.5 points); Proficient (1.0 point), Distinguished (1.25 points). Student performance is generated from the Kentucky Summative Assessment (KSA) and the Alternate KSA.
- **State Assessment Results in Science, Social Studies and Writing**. 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 to school, district, and state levels. Schools are rated based on student performance levels, Novice (0 points); Apprentice (0.5 points); Proficient (1.0 point), Distinguished (1.25 points). Student performance is generated from the Kentucky Summative Assessment (KSA) and the Alternate KSA.
- **English Learner Progress**. This component is based on improvement on the English Language Proficiency Exam by English Learners. English learners' progress is included in the calculation using an English learner progress table.[1]
- **Quality of School Climate and Safety**. This component is based on measures of the school environment. Perception data from surveys that measure insight into the school environment. Survey items are assigned scores of 0.00 for Strongly Disagree, and 33.33 for Disagree. The score of 66.66 is for Agree, and 100.00 for Strongly Agree. They are averaged for each question to get a question score. The question scores are then averaged to create an index.
- **Postsecondary Readiness (high school only)**. This component is based on attaining the necessary knowledge, skills, and dispositions to successfully transition to the next level of his or her education career. To demonstrate postsecondary readiness, high school students must earn a high school diploma or be classified as a Grade 12 nongraduate and meet the requirements for one type of readiness (Academic or Career).[2]

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

- **Graduation Rate (high school only)**. This component is based on the percentage of students earning a high school diploma compared to the cohort of students 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. At all grade spans, state assessment results in reading and mathematics is the indicator assigned the most weight. The English learner progress and 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 rate are only applied for high schools.

Table 1. Weighting of Accountability Indicators by Grade Span

Indicator	Elementary Weight	Middle Weight	High School Weight
State Assessment Results in Reading and Mathematics	51	46	45
State Assessment Results in Science, Social Studies, and Writing	40	45	20
English Learner Progress	5	5	5
Quality of School Climate and Safety	4	4	4
Postsecondary Readiness	NA	NA	20
Graduation Rate	NA	NA	6

Overall accountability scores are used to classify schools into performance levels. 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. Table 2 presents the score ranges for each accountability performance level rating at each of the three grade spans.

Table 2. Score ranges for overall accountability ratings

School Level	Red	Orange	Yellow	Green	Blue
Elementary Schools	0-34.9	35.0-52.9	53.0-68.9	69.0-78.9	79.0 or more
Middle Schools	0-36.9	37.0-50.9	51.0-61.9	62.0-70.9	71.0 or more
High Schools	0-47.9	48.0-58.9	59.0-66.9	67.0-77.9	78.0 or more

Because overall school scores and ratings are based on a combination of indicators, each containing some measurement error component, it is important to determine the extent to which school classifications can be expected to be accurate. This 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 will discuss issues related to the reliability of the overall accountability scores. Of particular interest are the characteristics of the scoring process and how they pose limitations for the quantification of error variance in the overall score.

State Assessment Results in Reading, Mathematics, Science, Social Studies, and Writing

The state assessment results components of the overall accountability score are designed to recognize schools for students reaching the desired level of knowledge and skills as measured on state required academic assessments in reading, mathematics, science, social studies, and writing. Reading and mathematics performance are combined as one indicator, and science, social studies, and writing performance are combined as a separate indicator. Both are based on student performance on the KSA and the Alternate KSA, specifically the percentage of students classified at each performance level (Novice, Apprentice, Proficient, and Distinguished). For students classified as Novice, schools receive 0 points. For Apprentice classifications, schools receive 0.5 points. For Proficient classifications, schools receive 1 point. For Distinguished classifications, schools receive 1.25 points. The range of points for the state assessment results in reading and mathematics indicator across all grade levels for the 2021-2022 school year was 10.3 to 101.3. The range of points for the state assessment results in science, social studies, and writing indicator across all grade levels for the 2021-2022 school year was 16.0 to 97.6.

There are a couple of concerns regarding classification error. First, it has been documented that student level classifications vary in terms of the amount of classification error, both across grade/subjects and across performance categories (Crawford & Dickinson, 2022). Within a particular grade/subject, errors in classification that are averaged across the performance categories tend to cancel one another out, yielding average amounts of error that are relatively small. But misclassification levels tend to vary by performance categories, which has implications for school-level classification accuracy when some student-level classifications are weighted more heavily than others.

Table 3 illustrates the average distribution error across test content areas for each student classification category in each grade level. These values were calculated from the difference between expected and observed classifications for each NAPD category obtained separately for each grade/subject, which were then averaged across all the content areas tested at each grade level. Although rules of thumb have not been established for interpreting average distribution error among assessments of academic achievement, the KSA demonstrates classification accuracy levels comparable to or higher than other state assessments (Crawford et. al., 2021).

Table 3. Average Distribution Error for Each Proficiency Category across Tested Subjects

	Novice	Apprentice	Proficient	Distinguished
Grade 3	0.39	0.53	0.78	0.97
Grade 4	1.51	2.23	0.98	1.50
Grade 5	0.89	1.01	1.18	1.38
Grade 6	0.97	1.96	1.28	1.22
Grade 7	0.95	2.64	1.50	1.23
Grade 8	0.85	1.97	1.59	1.55
High School	1.55	3.24	2.00	2.05

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 reading and math is 0.39%.

Table 3 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. Overall, for instance, Apprentice shows relatively higher error than other classifications. The state assessment results components of the overall score are derived from some combination of the number of students scoring at each level, but the same indicator score may reflect different combinations of these student classifications. Table 4 depicts a hypothetical example of how the same state assessment results in reading & mathematics indicator score might reflect different combinations of student classifications.

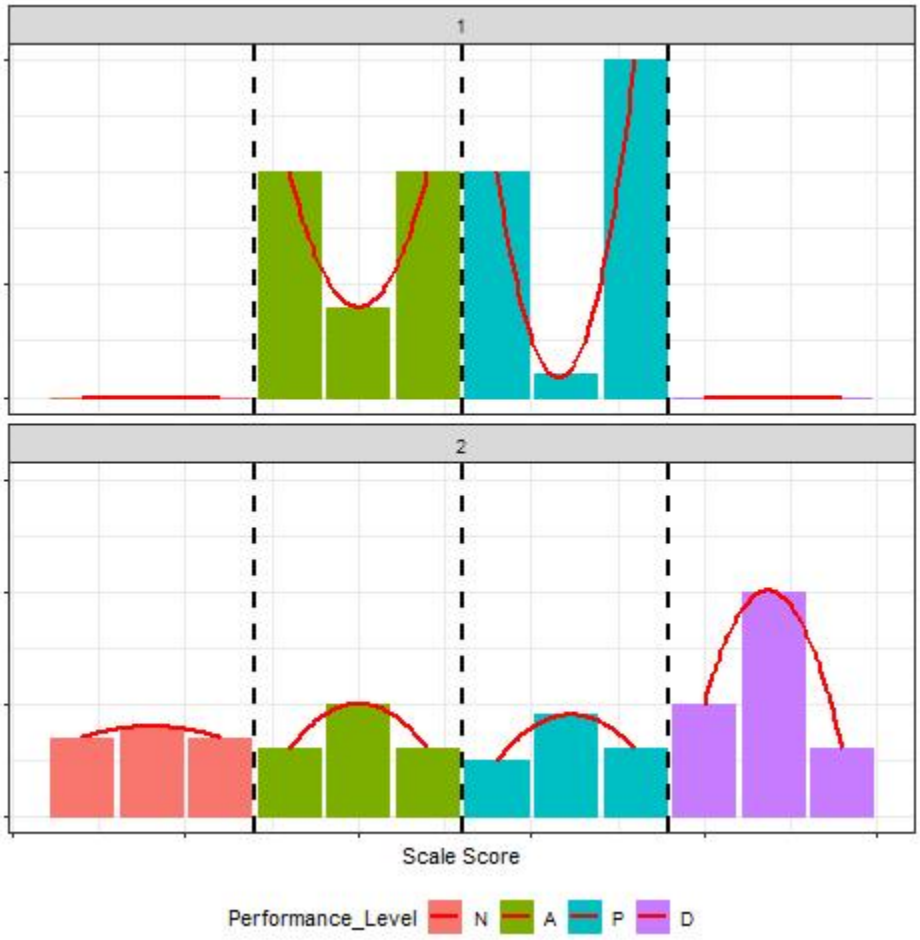
Table 4. Hypothetical Examples of the Same State Assessment Results in Reading & Mathematics Indicator Score Derived from Different Combinations of NAPD Percentages

	Novice	Apprentice	Proficient	Distinguished	Overall Score
School 1	0	48	52	0	76
School 2	22	22	20	36	76

It is not surprising that there are multiple ways to achieve a particular state assessment results indicator score (hereafter referred to as assessment indicator score), as demonstrated in Table 4. Kentucky's Accountability Model is intended to provide a balanced approach to measuring school performance and thus recognizes that accountability comprises a wide range of school and district work, and that high levels of performance may manifest in different ways. However, these differences in distributions of student NAPD classifications do have implications for the measurement of classification accuracy. Scores from a school that does not include students who scored in the Novice category do not rely on the accuracy of student-level Novice classifications. That school will likely differ in their level of accuracy from a school whose assessment indicator scores do rely in part on the accuracy of student-level Novice classifications.

Second, Betebenner et al. (2008) argued that a school having students lying close to performance level cut points would be expected to have higher classification error rates (i.e., lower accuracy rates) than a school with students well away from those thresholds. This is a different kind of classification error. Figure 1 depicts hypothetical score distributions for the two schools in Table 4. The upper panel shows the scores have a U-shaped distribution for each performance level whereas the lower panel shows the scores have a normal or uniform distribution within each performance level. The school in the upper panel may have higher classification errors in student performance level classification.

Figure 1. Hypothetical Score Distribution of Students' Scores Within Each Performance Level



Note. Vertical black dotted lines indicated individual cut points and red curves are smoothed school distribution within each category.

Third, it is important to consider how a school's demographic compositions could impact the accuracy of accountability classifications. For example, average score reliability varies across grade and student subgroups (Pearson, 2022). Depending on the grade configuration of the school and the student populations served, the accuracy of school classifications could be impacted.

English Learner Progress

The English learner progress component of the overall accountability score is designed to recognize schools for non-native English-speaking students making progress in becoming proficient in English. This indicator is operationalized by comparing a student's WIDA ACCESS or Alternate ACCESS performance (i.e., proficiency level) from last year to the current year using a table developed by KDE (2017). Each tested student is assigned points based on this comparison, and the school indicator is calculated by averaging these points across students. Like the state assessment results indicators, eligible students who do not participate in testing receive the lowest possible proficiency level rating, which may differentially impact schools who serve high percentages of at-risk students. The range of points for the English learner progress indicator across all grade levels for the 2021-2022 school year was 7.3 to 77.4.

Student proficiency levels range from 1.0 to 6.0 for ACCESS and A1-P2 for Alternative ACCESS and these proficiency levels are used to create the English learner progress indicator table. KDE should interpret this table with caution, as WIDA (2023) indicates that proficiency levels are grade specific and so should not be compared across grades. WIDA (2023) rather suggests comparing scale scores across grades as a measure of student progress.

Also, as with any assessment, student level classification is impacted by score reliability. Unlike the KSA, we do not have access to data necessary to calculate the accuracy of Kentucky students' WIDA performance classifications. WIDA does publish an annual technical report that presents reliability and classification accuracy results. WIDA (2022) reported both Cronbach's alphas 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 (lower bound was about .6). However, these are not state specific. Using the available reliability coefficients estimated from all participating WIDA states would be possible, but this would not account for the possibility that these values might be an over- or -underestimation of reliability for Kentucky students specifically.

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 2022, approximately 20% of Kentucky schools included the English learner progress indicator in their accountability calculation. In these cases, the English learner progress indicator weight was distributed proportionally among the remaining indicators. Considering its relatively low weight, it stands to reason that the English learner progress indicator would have minimal impact on a school's classification and would likely only impact a school with an overall score near a cut point.

However, there is a significant relationship between student achievement on standardized assessments and student demographic characteristics (ethnicity or SES status, etc.) that may impact all three accountability indicators related to students' achievement.

Quality of School Climate and Safety

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. The survey consists of a series of statements (i.e., items) with which students are asked to indicate their level of agreement. All items are written such that a higher level of agreement indicates a more positive perception of

the school environment. For each student, survey items are assigned scores of 0.00 for each response of Strongly Disagree, and 33.33 for each response of Disagree. 66.66 for each response of Agree, and 100.00 for each Strongly Agree response. These item-level scores are then 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 QSCS indicator across all grade levels for the 2021-2022 school year was 47.5 to 98.3.

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 et al., 2021; 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. Dickinson & Thacker (2023) demonstrated that modifying the current accountability weighting scheme would have minimal impact on schools' overall accountability ratings.

Postsecondary Readiness

The postsecondary readiness component of the overall accountability score is designed to recognize schools for preparing students to demonstrate readiness for postsecondary success. 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 Advance Placement (AP), International Baccalaureate (IB), or Cambridge Advanced International (CAI), or other approved, nationally recognized examination, earning an approved industry certification, scoring at or above the benchmark on the CTE End-of-Program (EOP) assessment for articulated credit, or completing a KDE/Cabinet approved apprenticeship program. Schools receive a point for each student identified as postsecondary ready and a bonus (1.25 points) for each college-ready student who demonstrates career readiness in a high-demand career sector (e.g., advanced manufacturing, business and information technology, construction trades, healthcare, and transportation and logistics). The final indicator score is based on the total points assigned for students identified as postsecondary ready divided by the total number of graduates plus grade 12 non-graduates. The range of postsecondary readiness points across high schools for the 2021-2022 school year was 35.2 to 111.3, with a mean of 80.46 and a standard deviation of 12.22).

Postsecondary readiness is an accountability indicator that relies on several different assessment instruments that may be used in various combinations within a given school. 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 component of the overall accountability score is designed to recognize schools for students completing graduation requirements. Graduation rates are reported by schools and districts. The range of graduation points among high schools for the 2021-2022 school year was 68.8 to 100, with a mean of 93.29 and a standard deviation of 4.00).

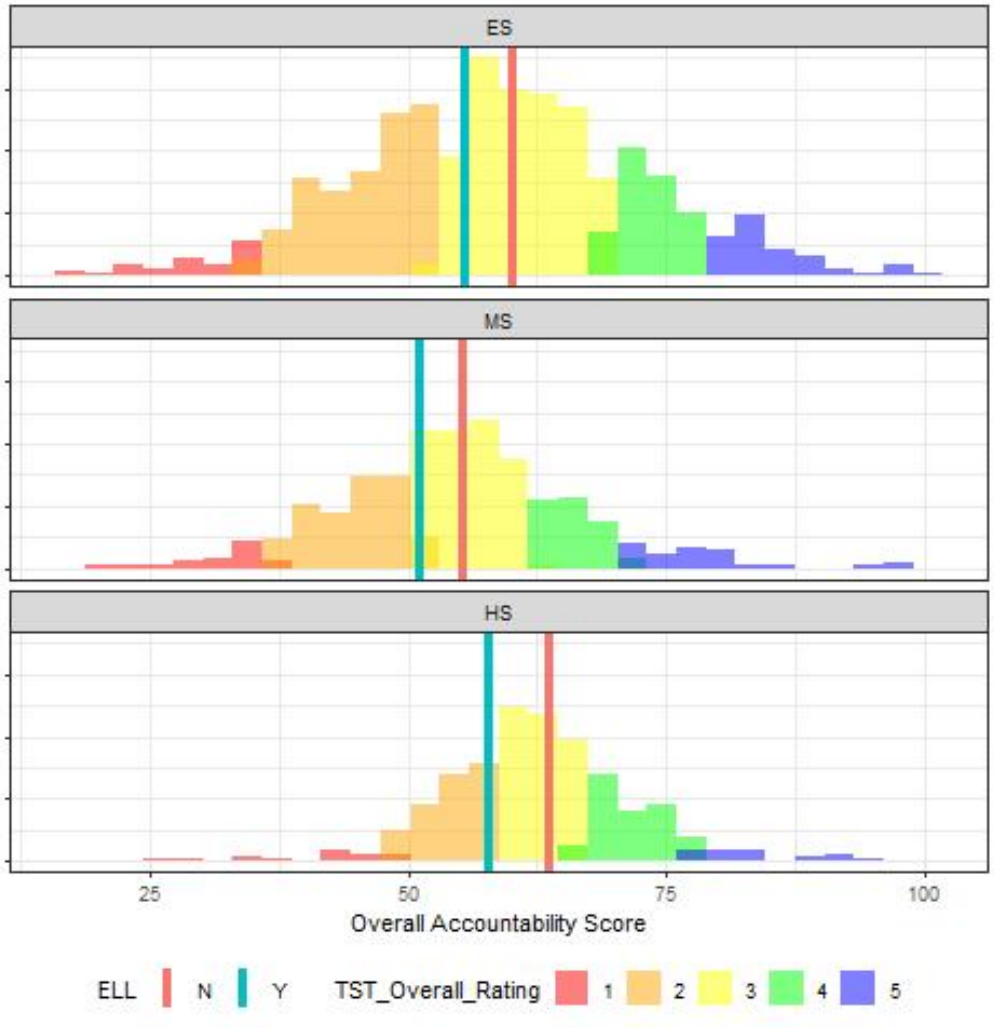
Although this component of the overall accountability score does not include multiple data sources or complex calculations, it does present its limitation for calculating school-level classification accuracy. Because graduation rates are a single, self-reported value, there is no method for estimating their error variance. Prior research explored the use of SEM values based on an assumed reliability of 1 (perfect reliability) and those based on an assumed reliability of 0 (total unreliability) and found only small differences in estimation error between these two assumptions (Hoffman and Wise, 2001, as cited in Hoffman & Dickinson, 2005). Research on school classification accuracy in Kentucky has since assumed a conservative reliability estimate of 0.7 to calculate error rates for graduation and other non-academic school-level performance indicators (Hoffman & Dickinson, 2005).

Combining the Components

Combining several component scores to create a single overall score is a scale-building process similar to developing an individual-level measurement. Rather than test items, component scores become the data points that could theoretically be used to calculate a reliability coefficient, derive a standard error of measurement, and calculate probabilities of true scores around observed scores. However, reliability coefficients are not available, or would be impractical to calculate, for some of the accountability indicators, thereby limiting the extent to which school-level classification accuracy can be quantified.

Another consideration is the treatment of missing indicators. If a particular indicator is not included for a school (e.g., EL indicator excluded due to the school not serving any students classified as EL), then the weight of that indicator is distributed proportionally among the remaining indicators. In some cases, schools may be missing more than one indicator. One potential concern is if the pattern of missing indicators is systematic rather than random. For example, Figure 2 depicts overall school score distributions for schools having EL indicator scores and schools that do not have EL indicator scores. The overall accountability score of schools not having EL indicator scores was significantly higher than schools having EL indicator scores. ($t=6.18$, $p<0.001$).

Figure 2. Comparison Of Overall School Accountability Score Distribution Within Each Classification Level Between Schools With and Without EL Indicator Scores



Note. The vertical lines are the means of overall accountability scores for the schools having EL indicator score (blue) and not having EL indicator score (red).

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 5 presents the range, mean, and standard deviation of overall scores for each school level (elementary, middle, and high schools). Table 6 presents the same descriptive statistics for each classification. As shown in Table 6, the Level 5 category has the largest score ranges at the elementary and middle school levels (20.6 and 25.5 points, respectively), whereas the Level 4 category has smaller ranges (9.6 and 8.6, respectively). At the high school level, the Level 1 category has the largest score range (21.5 points), whereas the Level 3 category has the smallest score range (7.9 points).

Table 5. Descriptive Statistics for Overall Accountability Scores

	Minimum	Maximum	Mean	Standard Deviation
Elementary (N=903)	16.6	99.8	59.00	13.68
Middle (N=500)	18.9	96.7	54.73	11.20
High (N=406)	26.2	93.9	62.75	9.08

Table 6. Overall Accountability Score Ranges Associated with Each Accountability Classification

	Classification	N	Min	Max	Range	Mean	STD
Elementary	1	35	16.6	34.9	18.3	29.33	5.16
	2	273	35.0	52.9	17.9	46.20	4.92
	3	395	53.0	68.9	15.9	61.21	4.48
	4	133	69.0	78.6	9.6	73.31	2.72
	5	67	79.2	99.8	20.6	85.23	4.91
Middle	1	24	18.9	36.7	17.8	32.03	4.37
	2	147	37.0	50.8	13.8	45.03	4.05
	3	215	51.0	61.9	10.9	55.94	3.08
	4	78	62.1	70.8	8.7	65.99	2.47
	5	36	71.2	96.7	25.5	77.87	6.70
High	1	15	26.2	47.7	21.5	41.08	6.92
	2	109	48.1	58.8	10.7	54.69	3.00
	3	170	59.0	66.9	7.9	62.75	2.38
	4	94	67.2	77.9	10.7	71.53	3.12
	5	18	78.2	93.9	15.7	83.89	5.23

Note. Min=minimum; Max=maximum; STD=standard deviation; *d*= Cohen's *d* for adjacent groups.

Because the overall accountability score combines multiple indicator scores, a key piece of validity evidence is to document how well each component differentiates between the classification categories. One way this can be done is through an analysis of the distribution of the component scores within each category and the extent to which there is overlap in component scores across the categories. This analysis consists of calculating descriptive statistics for each classification level within each grade span (e.g., elementary schools classified as Level 1, elementary schools classified as level 2, etc.) Tables 7-9 present the number of schools at each classification 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 7 shows that there were 35 elementary schools classified as Level 1 on the reading and mathematics assessment results component. Among those schools, the lowest score on this accountability component was 10.3, the highest score was 34.6, and the mean score was 25.34.

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 performance indicators (i.e., RD & MA and SC, SS, & WR) tended large across all grades and all classification comparison. Cohen's *d* indicates moderate effect size for QSCS and small effect size for EL. For high school, the effect size for PSR or graduate rate varies small to large 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. However, for PSR, the effect size for the highest category and the second highest category was small (0.28).

Table 7. Descriptive Statistics of Points Values of Overall Accountability Score Components by Classification Level: Elementary Schools

Classification	Component	N	Min	Max	Mean	Range	STD	*d*
1	RD & MA	35	10.3	34.6	25.34	24.3	6.07	
2	RD & MA	273	28.9	57.4	44.74	28.5	6.20	3.14
3	RD & MA	395	44.6	75.3	61.39	30.7	5.71	2.82
4	RD & MA	133	62.1	85.2	74.90	23.1	3.94	2.54
5	RD & MA	67	73.3	101.3	87.38	28.0	5.13	2.86
1	SC, SS, & WR	34	18.0	40.3	29.30	22.3	5.48	
2	SC, SS, & WR	267	28.0	60.4	44.94	32.4	5.74	2.74
3	SC, SS, & WR	390	43.9	73.3	59.59	29.4	5.68	2.57
4	SC, SS, & WR	126	60.8	84.2	71.21	23.4	4.65	2.13
5	SC, SS, & WR	64	72.1	97.6	83.00	25.5	6.54	2.20
1	EL	12	27.6	61.9	49.63	34.3	10.38	
2	EL	90	23.9	77.4	50.59	53.5	9.53	0.10
3	EL	80	28.2	71.8	52.26	43.6	9.52	0.17
4	EL	33	30.8	75.6	54.97	44.8	10.73	0.27
5	EL	6	41.9	73.2	58.50	31.3	13.34	0.32
1	QSCS	35	62.5	76.6	70.69	14.1	3.17	
2	QSCS	273	66.3	85.7	75.15	19.4	2.99	1.48
3	QSCS	395	69.5	98.3	77.27	28.8	3.62	0.63
4	QSCS	133	71.7	88.9	78.45	17.2	3.46	0.33
5	QSCS	67	73.1	94.3	80.96	21.2	4.76	0.64

Note. Min=minimum; Max=maximum; STD=standard deviation; *d*= Cohen's *d* for adjacent groups.

Table 8. Descriptive Statistics of Points Values of Overall Accountability Score Components by Classification Level: Middle Schools

Classification	Component	N	Min	Max	Range	Mean	STD	*d*
1	RD & MA	24	19.2	37.8	18.6	32.77	5.12	
2	RD & MA	147	35.1	56.2	21.1	46.46	4.62	2.92
3	RD & MA	215	48.8	67.8	19.0	58.68	4.01	2.86
4	RD & MA	78	61.5	76.8	15.3	68.31	3.13	2.54
5	RD & MA	36	71.2	98.5	27.3	80.89	7.19	2.63
1	SC, SS, & WR	23	16.0	37.7	21.7	28.93	5.22	
2	SC, SS, & WR	144	32.1	53.8	21.7	41.98	4.87	2.65
3	SC, SS, & WR	215	42.0	63.2	21.2	52.52	4.13	2.37
4	SC, SS, & WR	77	55.5	72.5	17.0	64.01	3.81	2.84
5	SC, SS, & WR	32	66.3	96.8	30.5	75.63	7.44	2.26
1	EL	9	16.6	34.2	17.6	25.37	6.47	
2	EL	23	10.6	50.8	40.2	24.45	8.86	-0.11
3	EL	23	15.1	34.3	19.2	24.87	4.77	0.06
4	EL	11	10.0	35.1	25.1	23.98	7.07	0.16
5	EL	3	29.6	44.1	14.5	36.03	7.39	1.69
1	QSCS	24	51.3	73.7	22.4	61.18	4.24	
2	QSCS	147	54.9	78.1	23.2	65.03	4.18	0.92
3	QSCS	215	56.2	81.8	25.6	67.25	3.77	0.56
4	QSCS	78	61.0	78.5	17.5	68.63	3.98	0.36
5	QSCS	36	64.7	81.9	17.2	70.96	3.73	0.60

Note. Min=minimum; Max=maximum; STD=standard deviation; *d*= Cohen's *d* for adjacent groups.

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

Classification	Component	N	Min	Max	Mean	Range	STD	*d*
1	RD & MA	15	14.8	38.8	32.18	24.0	7.04	
2	RD & MA	109	37.8	59.6	47.77	21.8	5.20	2.87
3	RD & MA	170	40.3	68.0	57.19	27.7	5.39	1.77
4	RD & MA	94	59.4	85.9	68.32	26.5	5.77	2.02
5	RD & MA	18	72.4	98.8	85.96	26.4	8.61	2.80
1	SC, SS, & WR	15	16.3	36.3	27.18	20.0	6.39	
2	SC, SS, & WR	109	29.4	51.5	41.05	22.1	5.76	2.38
3	SC, SS, & WR	170	35.8	62.0	49.18	26.2	5.55	1.45
4	SC, SS, & WR	94	44.5	77.9	57.05	33.4	6.66	1.32
5	SC, SS, & WR	18	63.2	87.0	71.67	23.8	7.07	2.17
1	EL	8	13.3	36.5	24.05	23.2	7.88	
2	EL	18	7.3	40.0	24.41	32.7	9.77	0.04
3	EL	32	8.4	55.5	23.84	47.1	11.35	-0.05
4	EL	7	16.7	45.8	31.26	29.1	10.22	0.66
5	EL	0	--	--	--	--	--	NA
1	QSCS	15	53.2	62.7	59.15	9.5	2.84	
2	QSCS	109	47.5	69.1	60.32	21.6	3.63	0.33
3	QSCS	170	54.3	72.4	61.55	18.1	3.42	0.35
4	QSCS	92	57.1	71.6	63.72	14.5	3.86	0.61
5	QSCS	18	60.3	72.9	67.09	12.6	4.04	0.87
1	PSR	15	35.2	81.9	59.77	46.7	16.37	
2	PSR	109	42.6	100.0	72.62	57.4	10.77	1.11
3	PSR	170	48.6	104.9	81.73	56.3	9.47	0.91
4	PSR	94	66.0	111.3	88.57	45.3	8.84	0.74
5	PSR	18	84.9	96.6	90.91	11.7	3.80	0.28
1	Grad	15	81.1	94.7	86.07	13.6	4.34	
2	Grad	109	83.4	100.0	92.40	16.6	3.57	1.72
3	Grad	170	68.8	98.9	93.34	30.1	3.76	0.26
4	Grad	94	85.7	100.0	94.67	14.3	3.29	0.37
5	Grad	18	93.5	100.0	96.90	6.5	1.92	0.71

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 classification levels differ. Figures 3 through 8 generally depict the stair-step pattern between overall classification and each component of the overall accountability score, except for the EL 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 a more overlap among the remaining indicators across the accountability classifications.

Figure 3. Ranges of Reading and Math Indicator Scores Within Overall Classifications

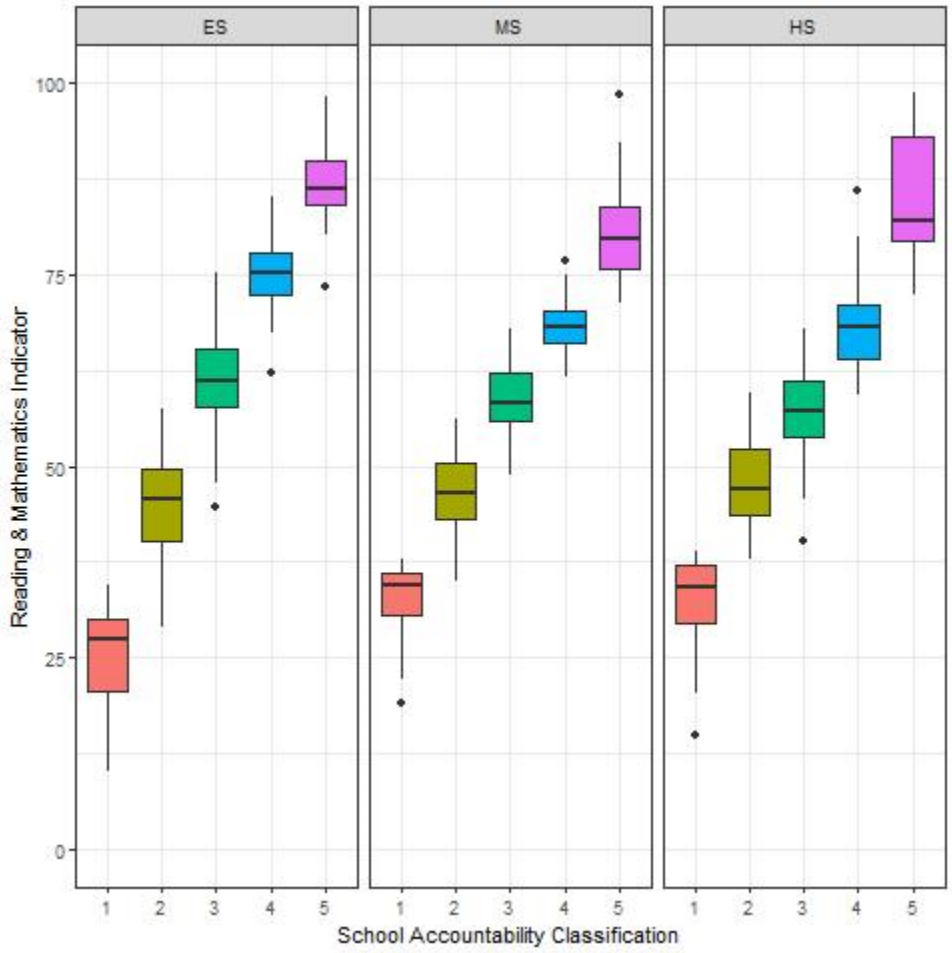


Figure 4. Ranges of Science, Social Studies, and Writing Indicator Scores Within Overall Classifications

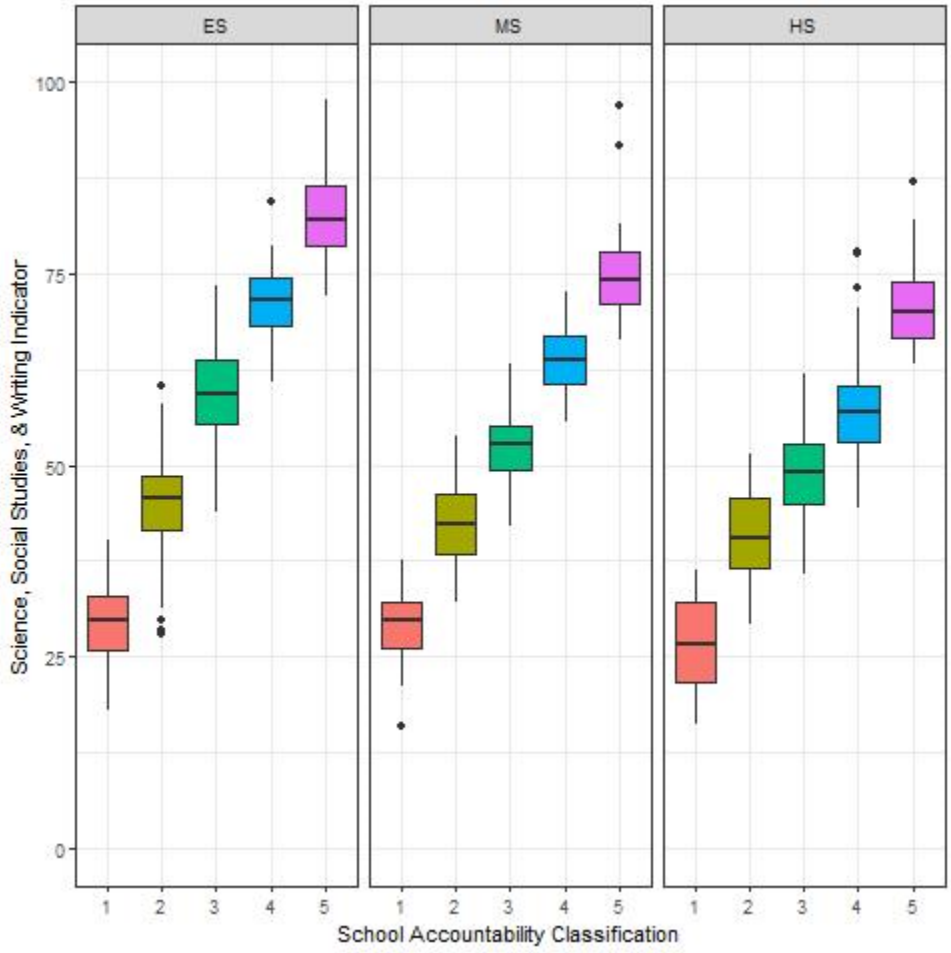


Figure 5. Ranges of English Language Progress Indicator Scores Within Overall Classifications

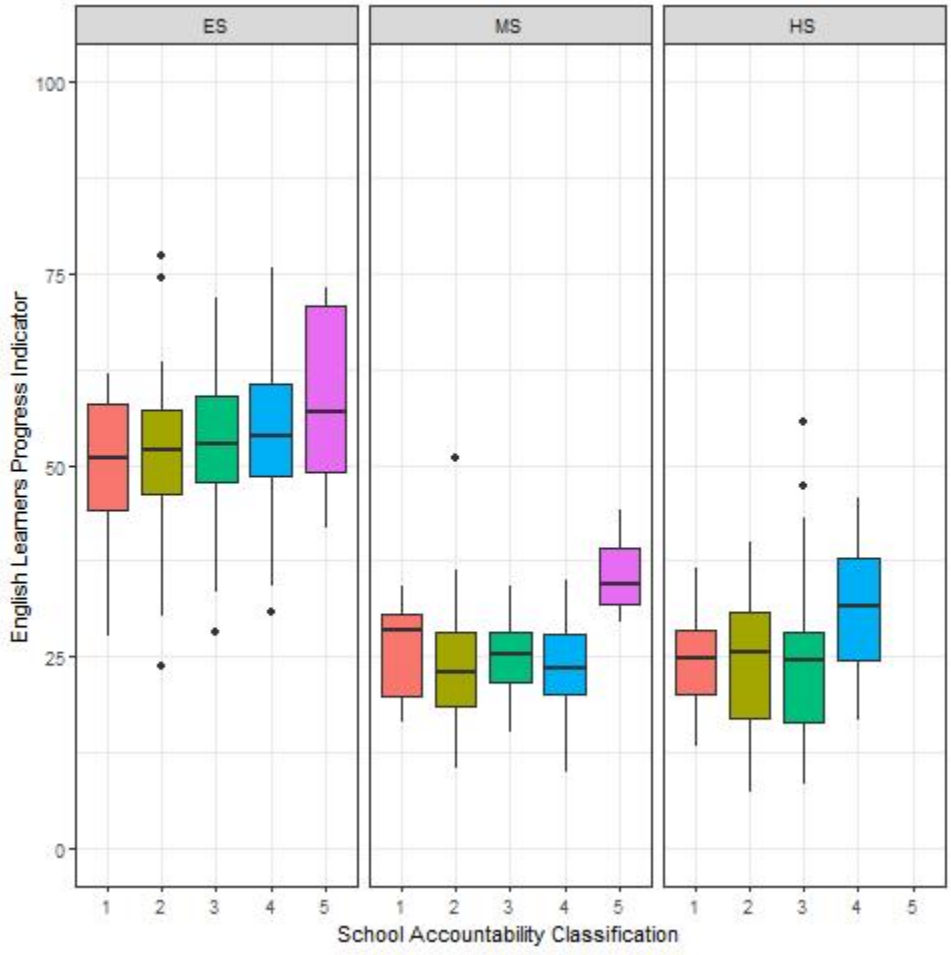


Figure 6. Ranges of Climate and Safety Indicator Scores Within Overall Classifications

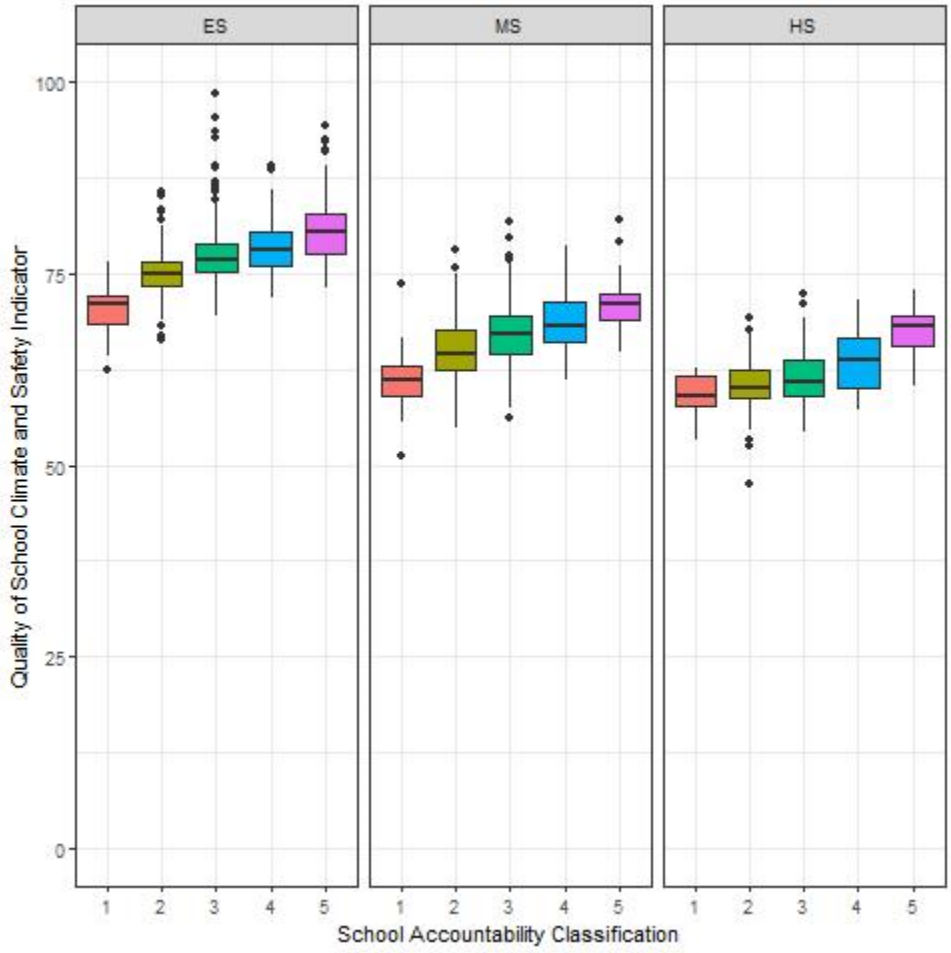


Figure 7. Ranges of Postsecondary Readiness Indicator Scores Within Overall Classifications

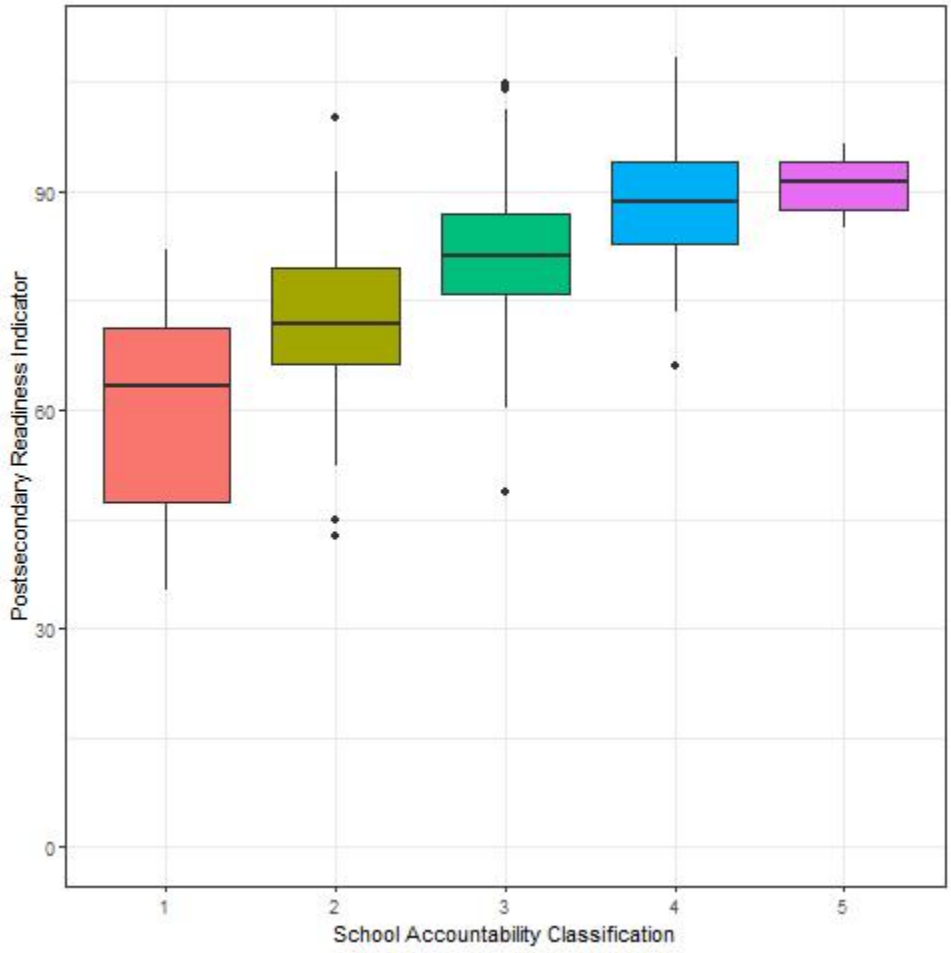
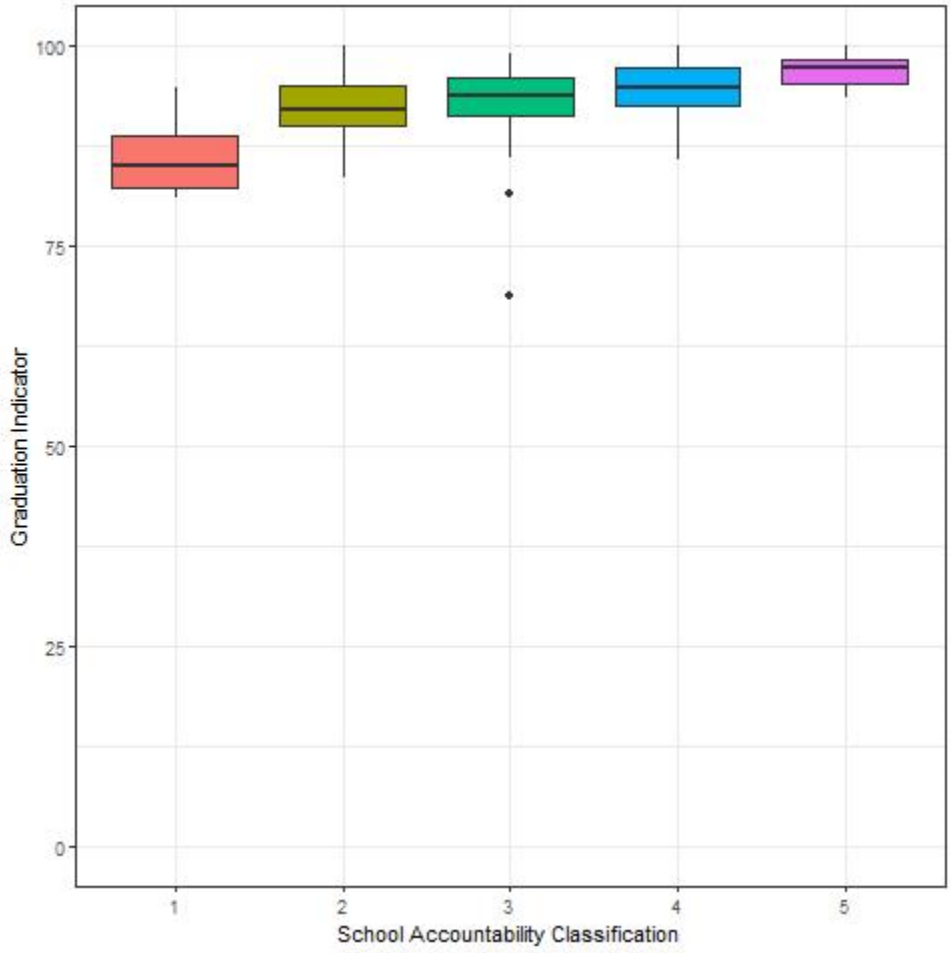


Figure 8. Ranges of Graduation Rate Indicator Scores Within Overall Classifications



Discussion and Recommendations

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 by design. 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 accurately classify schools. The complexity of the model, however, does not allow for a straightforward quantification of reliability or for error variance of 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 more 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, supporting the validity of Kentucky's school-level accountability classifications.

Reliance on current status measures raises a number of issues of fairness to schools (Linn, 2002; Meyers, 2000), particularly for schools serving poor and/or initially low-achieving students. Beginning in the 2022-2023 school year, accountability will be based on not only a school's current status, but also the amount of change schools have experienced on each component since the previous school year. Table 10 is an example KDE presented in "Kentucky's Accountability System at a Glance." KDE also indicated that "tables will be developed for each indicator and may be different." This could be a good opportunity for some of those schools to demonstrate their improvement. However, this will also introduce more complexity into the model that will likely further complicate estimating the accuracy of school classifications. Change under the Kentucky model will be evaluated based on a school's rating in the current year relative to its prior year's rating. Thus, the overall rating will reflect a compounding of the classification error from each of the included years for each accountability component. On the other hand, accounting for change will enhance the validity of school classifications by recognizing the adjustments that schools make from year to year in response to feedback from the system.

Table 10. 5-by-by Colored Table for Status and Change Table

LEVEL	Declined Significantly form Prior Year	Declined form Prior Year	Maintained from Prior Year	Increased from Prior Year	Increased Significantly from Prior Year
Very High in Current Year	Yellow	Green	Blue	Blue	Blue
High in Current year	Yellow	Yellow	Green	Green	Blue
Medium in Current Year	Orange	Orange	Yellow	Green	Green
Low in Current Year	Red	Orange	Orange	Yellow	Yellow
Very Low in Current Year	Red	Red	Red	Orange	Yellow

Note. This table was retrieved from https://education.ky.gov/AA/Acct/Documents/Accountability_at_a_Glance_2021-2022.pdf.

References

Betebenner, D. W., Shang, Y., Xiang, Y., Zhao, Y., & Yue, X. (2008). The impact of performance level misclassification on the accuracy and precision of percent at performance level measures. *Journal of Educational Measurement, 45*(2), 119-137.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge Academic.

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.

Crawford, B. F., & Dickinson, E. R., & Thacker A. A. (2021). *The accuracy of students' novice, apprentice, proficient, and distinguished classifications for the 2022 Kentucky Summative Assessment (KSA) tests* (2021 No. 156). Human Resources Research Organization.

Dickinson, E. R., & Thacker, A. A. (2023). *Exploring the School Climate and Safety Indicator*. 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.

Dickinson, E. R., Thacker, A. A., & Paulsen, J. (2021). *Analysis of the 2021 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. (2017,September). Commonwealth of Kentucky Revised Consolidated State Plan Under the Every Student Succeeds Act. Kentucky Department of Education, Frankfort, KY.

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.

Linn, R. (2000). Assessments and accountability. Educational Researcher, 29(2), 4–16.

Meyers, R. H. (2000). Value-added indicators: A powerful tool for evaluating science and mathematics programs and policies. *NISE Brief 3*(3). Madison: University of Wisconsin-Madison, National Center for Improving Science Education.

Pearson (2022). Kentucky Summative Assessment, 2021-22 Yearbook.

WIDA (2022). *Annual Technical Report for ACCESS for ELLs Online English Language Proficiency Test Series 502, 2020–2021 Administration*. https://wida.wisc.edu/sites/default/files/resource/ACCESS-Online%20ATR-Redacted.pdf

WIDA (2023). *ACCESS for ELLs Interpretive Guide for Score Reports Grades K-12*. Board of Regents of the University of Wisconsin System. https://wida.wisc.edu/sites/default/files/resource/Interpretive-Guide.pdf