HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

2024 No. 058

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:	Dea Mulolli Hye-Jeong Choi Emily Dickinson	**Date:**	May 15, 2024

School Classification Accuracy: Issues for Reliability and Validity

Table of Contents

Introduction iii

Reliability Issues vi

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

English Learner Progress viii

Quality of School Climate and Safety ix

Postsecondary Readiness ix

Graduation Rates x

Combining the Components x

Validity Issues xi

Discussion xix

References xxi

List of Tables

Table 1. Weighting of Accountability Indicators by Grade Span iv

Table 2. Score Ranges for Overall Accountability Ratings vi

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

Table 4. Descriptive Statistics for Overall Accountability Scores xi

Table 5. Overall Accountability Score Ranges Associated with Each Accountability Classification xi

Table 6. Descriptive Statistics of Points Values of Overall Accountability Score Components by Classification Level: Elementary Schools xiii

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

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

Table of Contents (Continued)

List of Figures

Figure 1. Elementary School Cut Scores for Each Indicator (KDE, 2023).....v
Figure 2. Accountability Calculation Incorporating Status and Change (KDE, 2023) v
Figure 3. Ranges of Reading and Math Indicator Scores Within Overall Classifications..... xvi
Figure 4. Ranges of Science, Social Studies, and Writing Indicator Scores Within Overall Classifications xvi
Figure 5. Ranges of English Learner Progress Indicator Scores Within Overall Classifications xvi
Figure 6. Ranges of Climate and Safety Indicator Scores Within Overall Classifications..... xvii
Figure 7. Ranges of Postsecondary Readiness Indicator Scores Within Overall Classifications xviii
Figure 8. Ranges of Graduation Rate Indicator Scores Within Overall Classifications.....xviii

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 initially changed was how these points are weighted and how they are 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 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 at 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 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 at 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. 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. 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. 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 whether a student has attained the necessary knowledge, skills, and dispositions to successfully transition to the next level of his or her education 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 nongraduate and meet the requirements for one type of readiness (Academic or Career).[2]

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 are the indicators assigned the most weight. 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.

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

In previous years, indicator scores simply reflected a school's current year score (Status), but indicator scores are now 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. 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.

Based on their Status and Change scores, individual schools are classified into one of five 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. As an example, Figure 1 presents Status and Change cut scores for elementary schools. Table 2 presents the score ranges for each accountability performance level rating at each of the three grade spans.

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

Figure 1. Elementary School Cut Scores for Each Indicator (KDE, 2023)

Elementary School Indicator Status Cut Scores

School Level	Indicators	Very Low	Low	Medium	High	Very High
Elementary School Status	State Assessment Results in Reading/Mathematics	0-31.9	32.0-53.9	54.0-69.9	70.0-80.9	81.0-125
	State Assessment Results in Science/ Social Studies/Writing	0-33.9	34.0-49.9	50.0-66.9	67.0-75.9	76.0-125
	English Learner Progress	0-33.9	34.0-47.9	48.0-57.9	58.0-64.9	65.0-140
	Quality of School Climate and Safety	0-66.9	67.0-73.9	74.0-76.9	77.0-81.9	82.0-100

Elementary School Indicator Change Cut Scores

School Level	Indicators	Declined Significantly	Declined	Maintained	Increased	Increased Significantly
Elementary School Change	State Assessment Results in Reading/Mathematics	-6.1 or less	-6.0 to -2.1	-2.0 to 0.0	0.1 to 6.9	7.0 or more
	State Assessment Results in Science/ Social Studies/Writing	-7.1 or less	-7.0 to -2.1	-2.0 to 0.0	0.1 to 8.9	9.0 or more
	English Learner Progress	-7.1 or less	-7.0 to -1.1	-1.0 to 0.0	0.1 to 22.9	23.0 or more
	Quality of School Climate and Safety	-5.1 or less	-5.0 to -2.1	-2.0 to 0.0	0.1 to 3.7	3.8 or more

2

Table 2. Score Ranges for Overall Accountability 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

Because overall school scores and ratings are based on a combination of indicators, there are multiple sources that can contain measurement error. First, each indicator has 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 important 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. However, a small percentage of schools had a lower classification, and a larger percentage of schools had a higher classification. 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 will discuss 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. The current report focuses on the characteristics of each indicator and related 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 skill 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 (NAPD: 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. School scores are computed by averaging the student-level points and multiplying by 100. The range of points for the state assessment results in reading and mathematics indicators across all grade levels for the 2022-2023 school year was 10.3 to 125.0. The range of points for the state assessment results in

science, social studies, and writing indicators across all grade levels for the 2022-2023 school year was 3.3 to 125.0. These results can be found later in this report in Tables 6 through 8.

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; Mulolli et al., 2023). 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. However, 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. Further, student-level classification errors from both the prior and current years have the potential to impact school-level classification via the inclusion of the Change score.

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, Kentucky's summative assessment has historically demonstrated classification accuracy levels comparable to or higher than other state assessments (Crawford et al., 2021). 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 3. Average Error Distribution for Each Proficiency Category across Tested Participants and Subjects

	Novice	Apprentice	Proficient	Distinguished
Grade 3	0.38	0.40	0.65	0.60
Grade 4	0.67	1.11	0.51	0.67
Grade 5	0.64	0.56	1.08	0.97
Grade 6	0.90	1.08	0.94	1.01
Grade 7	0.97	2.33	1.56	0.79
Grade 8	0.75	0.75	1.24	1.03
High School	0.68	1.88	1.19	0.76

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 0.38%.

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. For instance, Apprentice (Grades 7 and high school) shows relatively higher error rates 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.

English Learner Progress

The English learner progress (EL) 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 WIDA ACCESS or Alternate ACCESS performance (i.e., proficiency level) from last year to the current year using a table developed by KDE.[3] 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 that serve high percentages of at-risk students. The range of points for the English learner progress indicator across all grade levels for the 2022-2023 school year was 10.40 to 140.

Student proficiency levels range from 1.0 to 6.0 for ACCESS and A1-P2 for Alternative ACCESS, and these proficiency levels were 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 the data necessary to calculate the accuracy of Kentucky students' WIDA performance classifications. WIDA publishes an annual technical report that presents results indicating that ACCESS and Alternative ACCESS produce reliable and accurate classifications. WIDA (2023) reported 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). 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 under-estimation 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 2023, approximately 22% of Kentucky schools included the English learner progress indicator in their accountability calculation. If schools did not serve English learners, 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. The English language progress indicator score also contains measurement error associated with both a students' prior year and current year English language proficiency classification. The addition of Change to the accountability model did not impact the weighting of the English language progress indicator, which remains among the lowest of the accountability indicators.

However, there is a significant relationship between student achievement on standardized assessments and student demographic characteristics (ethnicity or socioeconomic status, etc.) that may impact any accountability indicators that are based on assessment performance. In other words, a school's student population may impact the school's achievement-related

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

indicator scores, and interventions may be designed to focus on the needs of particular student groups. Adding Change to the accountability model is one way to allow for school accountability scores to reflect academic performance improvements among all student groups.

Quality of School Climate and Safety

Scholars defined school climate as "the quality and character of school life," which is "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, p. 182). 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 the QSCS indicator across all grade levels for the 2022-2023 school year 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 et al., 2021; Dickinson & Thacker, 2022; Dickinson et al., 2023). 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, though this study was conducted before the inclusion of indicator Change scores in the accountability model.

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 Advanced 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 Career and Technical Education (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 2022-2023 school year was 52.0 to 125.0, with a mean of 96.16 and a standard deviation of 14.87.

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. Schools and districts report graduation rates. 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.

Although this component of the overall accountability score does not include multiple data sources or complex calculations, it does present limitations 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 standard error of measurement (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 & 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, the overall accountability score of schools not having EL indicator scores was significantly higher than that of schools having EL indicator scores. ($t=6.79$, $p<0.001$). The t-test scores are similar to the previous year ($t=6.18$), which may indicate that the effect of having more EL students on the overall score is consistent.

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 4 presents the range, mean, and standard deviation of overall accountability scores for each school level (elementary, middle, and high schools). Table 5 presents the same descriptive statistics for each performance level within each school level. As shown in Table 5, Level 5 has the largest score ranges at the elementary and middle school levels (40.4 and 29.5 points, respectively), whereas Level 4 has smaller ranges in the elementary school level (12.6 points), while Level 3 has the smallest range in middle school level (12.7 points). At the high school level, Level 5 has the largest score range (17.3 points), whereas Level 4 has the smallest score range (9.6 points).

Table 4. Descriptive Statistics for Overall Accountability Scores

School Level	Min	Max	Mean	STD
Elementary (N=716)	13.1	123.4	66.4	17.0
Middle (N=316)	16.0	106.6	58.5	15.4
High (N=227)	31.1	98.3	66.9	11.7

Table 5. Overall Accountability Score Ranges Associated with Each Accountability Classification

School Level	Performance Rating	N	Min	Max	Range	Mean	STD
Elementary	1	32	13.1	37.6	24.5	29.4	6.7
	2	149	38.1	54.9	16.8	47.9	4.6
	3	221	55.0	69.9	14.9	62.5	4.4
	4	187	70.1	82.7	12.6	75.6	3.6
	5	127	83.0	123.4	40.4	90.6	7.5
Middle	1	24	16.0	35.8	19.8	29.2	5.9
	2	78	36.1	50.9	14.8	45.1	4.4
	3	101	51.2	63.9	12.7	57.9	3.7
	4	80	64.0	76.8	12.8	70.1	3.7
	5	33	77.1	106.6	29.5	85.0	7.4
High	1	14	31.1	47.2	16.1	40.4	5.3
	2	38	49.2	59.9	10.7	54.7	3.0
	3	87	60.1	70.9	10.8	65.3	3.2
	4	65	71.0	80.6	9.6	75.2	2.8
	5	23	81.0	98.3	17.3	85.8	4.4

Note. Min=minimum; Max=maximum; STD=standard deviation

Because the overall accountability score combines multiple indicator scores, a key piece of validity evidence is to document how well each component differentiates between performance levels. One way that can be done is through an analysis of 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 6-8 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 6 shows that there were 32 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 12.7, the highest score was 39.0, and the mean score was 28.10.

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., RD & MA and SC, SS, & WR) 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 EL, 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, whereas 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 (Choi & Dickinson, 2023).

Table 6. 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	32	12.7	39.0	28.1	26.3	7.3	
2	RD & MA	149	27.6	65.0	46.8	37.4	7.4	2.59
3	RD & MA	221	40.5	79.8	61.6	39.3	6.3	2.19
4	RD & MA	187	54.4	94.9	75.2	40.5	6.3	2.16
5	RD & MA	127	70.3	125.0	91.4	54.7	9.5	2.15
1	SC, SS, & WR	31	3.3	45.9	25.5	42.6	9.5	
2	SC, SS, & WR	144	27.9	64.2	45.3	36.3	7.6	2.46
3	SC, SS, & WR	219	36.9	80.5	61.9	43.6	7.9	2.13
4	SC, SS, & WR	182	57.1	105.9	75.7	48.8	7.9	1.80
5	SC, SS, & WR	120	62.6	125.0	90.0	62.4	9.8	1.70
1	EL	11	40.7	99.9	73.7	59.2	16.7	
2	EL	58	31.4	111.2	76.4	79.8	17.1	0.14
3	EL	56	10.4	108.7	77.0	98.3	19.2	0.03
4	EL	41	45.6	140.0	82.4	94.4	21.2	0.18
5	EL	16	54.4	131.9	90.7	77.5	26.9	0.28
1	QSCS	32	59.5	80.6	69.8	21.1	5.2	
2	QSCS	149	62.7	92.3	73.8	29.6	5.2	0.82
3	QSCS	221	60.2	92.2	76.5	32.0	5.2	0.49
4	QSCS	187	68.4	100.0	78.3	31.6	5.7	0.35
5	QSCS	127	69.0	100.0	82.8	31.0	7.3	0.69

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

Table 7. 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	13.1	42.3	29.2	30.3	7.2	
2	RD & MA	78	31.5	60.7	29.2	46.8	6.8	2.32
3	RD & MA	101	36.9	71.6	34.7	58.7	6.3	1.98
4	RD & MA	80	61.7	84.4	22.7	72.7	5.6	2.35
5	RD & MA	33	74.7	125	50.3	89.7	10.8	2.48
1	SC, SS, & WR	23	7.6	36.5	28.9	25.7	6.7	
2	SC, SS, & WR	78	29.0	64.7	35.7	42.4	6.3	2.60
3	SC, SS, & WR	100	43.9	77.4	33.5	56.5	5.6	2.53
4	SC, SS, & WR	80	51.5	86.4	34.9	67.8	6.4	2.02
5	SC, SS, & WR	29	67.4	100.0	32.6	80.5	7.8	2.01
1	EL	12	1.2	26.2	25.0	13.8	8.8	
2	EL	19	0.0	43.5	43.5	21.7	11.6	0.58
3	EL	10	4.6	61.4	56.8	28.8	16.8	0.51
4	EL	7	11.2	59.0	47.8	38.3	16.9	0.25
5	EL	2	31.1	49.0	17.9	40.1	12.7	0.49
1	QSCS	24	51.0	83.4	32.4	61.1	6.6	
2	QSCS	78	54.1	79.6	25.5	62.9	5.1	0.24
3	QSCS	101	55.2	86.0	30.8	66.7	6.5	0.62
4	QSCS	80	57.2	86.1	28.9	68.1	6.0	0.24
5	QSCS	33	61.3	99.6	38.3	74.6	9.5	0.89

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: High Schools

Classification	Component	N	Min	Max	Mean	Range	STD	*d*
1	RD & MA	14	10.3	44.3	34.0	25.4	9.2	
2	RD & MA	38	20.8	56.4	35.6	40.7	8.4	1.77
3	RD & MA	87	31.8	72.0	40.2	55.6	8.0	1.88
4	RD & MA	65	51.7	90.6	38.9	68.3	7.2	1.64
5	RD & MA	23	70.7	103.8	33.1	85.1	8.0	2.29
1	SC, SS, & WR	14	9.7	41.9	32.2	25.2	9.7	
2	SC, SS, & WR	38	21.1	73.4	52.3	43.0	11.8	1.45
3	SC, SS, & WR	87	27.1	95.1	68.0	50.1	10.9	0.71
4	SC, SS, & WR	63	29.5	90.6	61.1	60.1	12.2	0.86
5	SC, SS, & WR	22	70.7	103.8	33.1	85.1	8.0	0.67
1	EL	9	13.7	37.7	24.0	.62222	8.9	
2	EL	10	4.7	47.9	43.2	27.9	12.8	0.12
3	EL	18	3.1	67.1	64.0	30.8	15.7	0.33
4	EL	8	11.2	44.8	33.6	33.0	13.1	0.52
5	EL	0	-	-	-	-	-	-
1	QSCS	14	53.0	68.9	15.9	59.2	4.4	
2	QSCS	38	46.5	70.9	24.4	60.0	5.5	0.19
3	QSCS	87	49.9	80.0	30.1	61.6	5.4	0.41
4	QSCS	65	53.1	81.8	28.7	64.9	6.2	0.58
5	QSCS	23	61.4	81.8	20.4	67.2	5.1	0.32
1	PSR	14	57.6	95.1	37.5	74.3	11.5	
2	PSR	37	52.0	125.0	73.0	87.8	16.2	1.03
3	PSR	87	69.6	125.0	55.4	95.8	12.5	0.49
4	PSR	64	78.2	125.0	46.8	103.0	11.4	0.62
5	PSR	23	82.8	125.0	42.2	105.3	10.6	0.30
1	Grad	14	83.6	100.0	16.4	88.1	4.5	
2	Grad	38	81.6	100.0	18.4	93.3	4.6	1.10
3	Grad	87	85.3	100.0	14.7	95.0	3.2	0.48
4	Grad	65	79.3	100.0	20.7	95.3	3.6	0.14
5	Grad	23	88.9	100.0	11.1	96.2	3.7	0.10

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 2 through 7 depict the stair-step pattern between overall performance 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 more overlap among the remaining indicators across the accountability classifications.

Figure 2. Ranges of Reading and Math Indicator Scores Within Overall Classifications

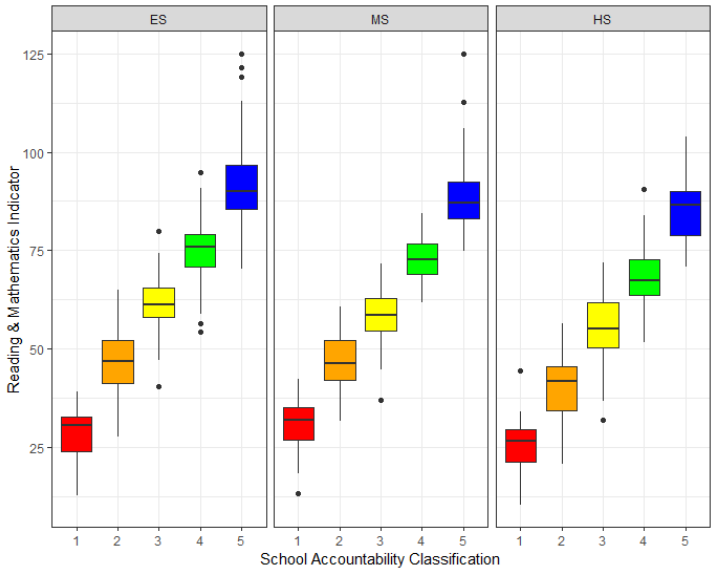


Figure 3. Ranges of Science, Social Studies, and Writing Indicator Scores Within Overall Classifications

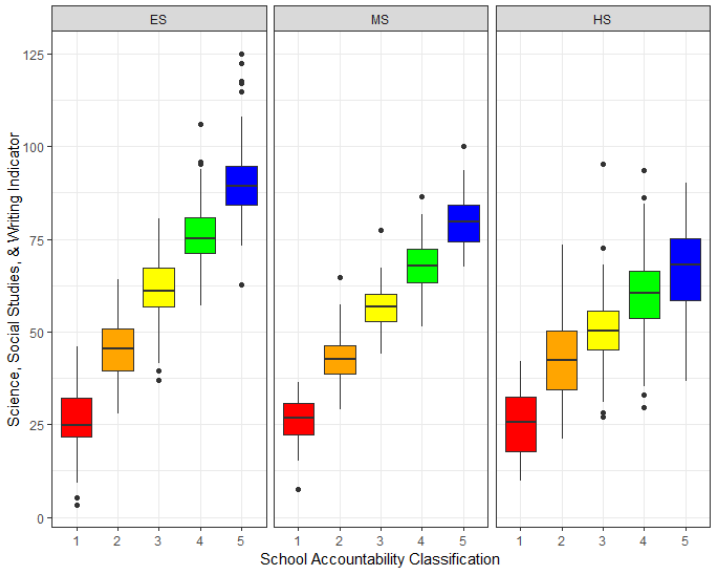


Figure 4. Ranges of English Learner Progress Indicator Scores Within Overall Classifications

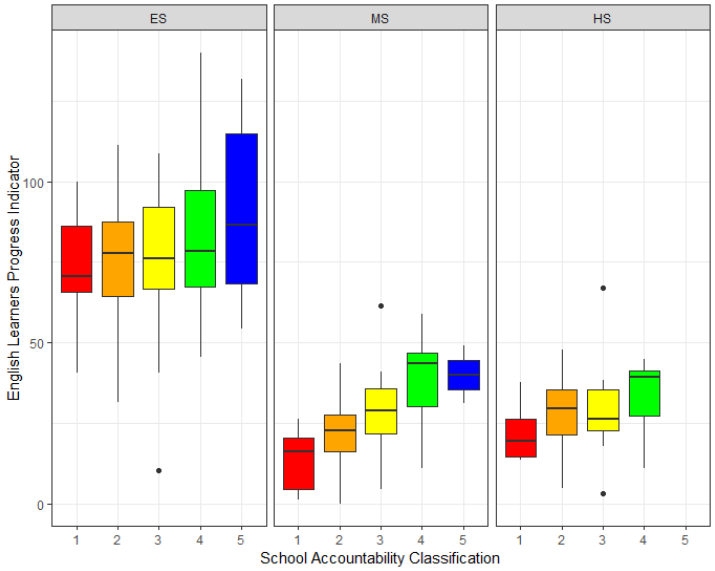


Figure 5. Ranges of Climate and Safety Indicator Scores Within Overall Classifications

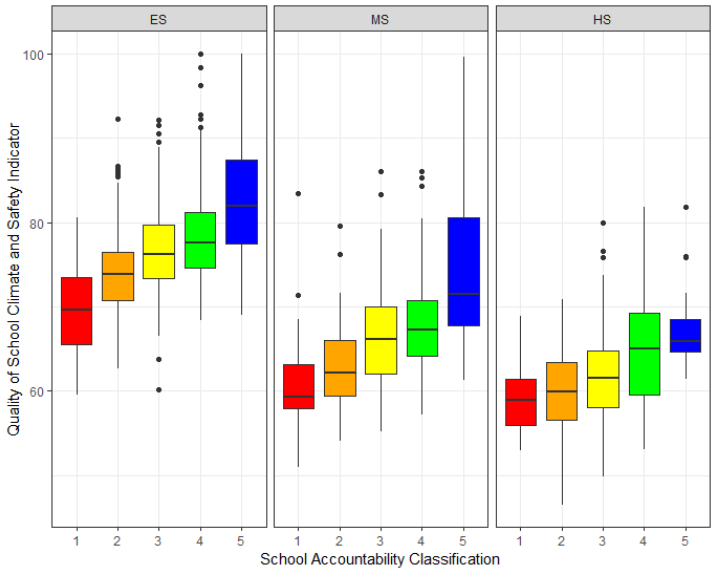


Figure 6. Ranges of Postsecondary Readiness Indicator Scores Within Overall Classifications

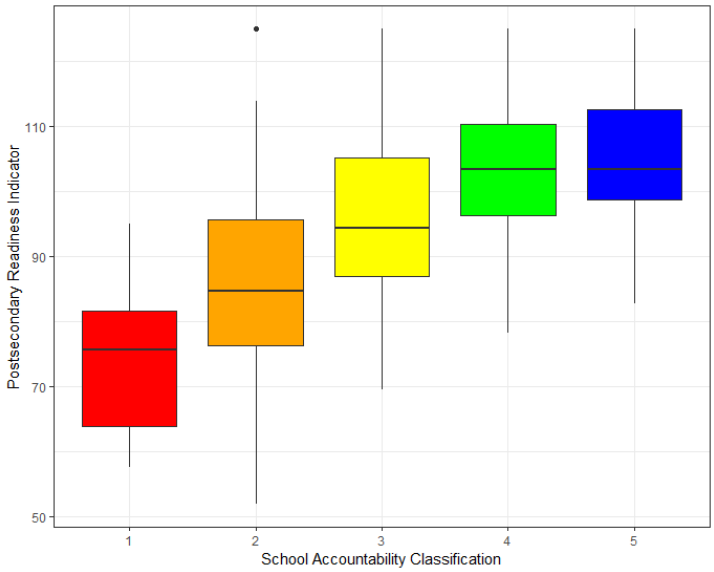
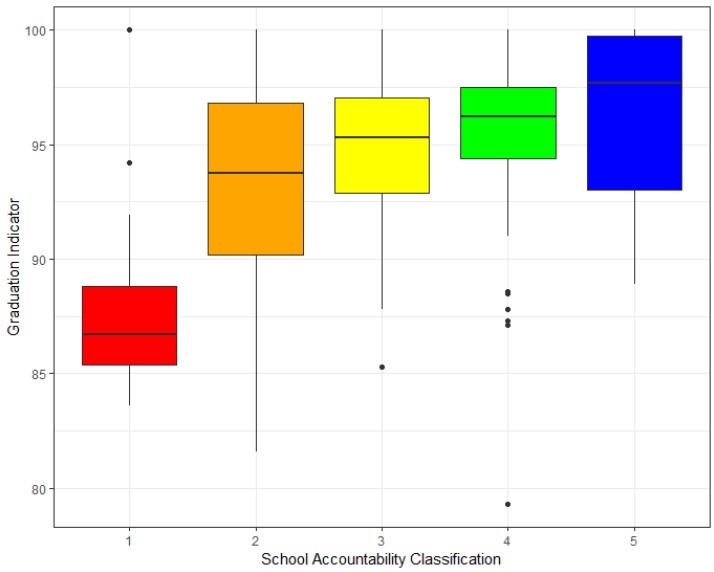


Figure 7. Ranges of Graduation Rate Indicator Scores Within Overall Classifications



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. The complexity of the model, however, does not allow for a straightforward quantification of reliability or for a calculation and comparison 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, supporting the validity of Kentucky's school-level accountability classifications.

Reliance on previously used status measures raises several issues of fairness for schools (Linn, 2002; Meyers, 2000), particularly for schools serving poor and/or initially low-achieving students. Beginning in 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. However, 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 improve 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 post-secondary 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 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.

To support future school-level classification accuracy research, we recommend that KDE authorize a simulation study that more closely examines the contribution of error introduced by adding Change to the accountability system and how it impacts school-level classification. This 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 should Kentucky assume is due to measurement error versus true changes in school performance).

We also recommend an investigation of the impact of accountability designations, and changes in designation, on schools. This could require school visits, but some quantitative work could begin using existing 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

Choi, H-J., & Dickinson, E. R. (2023). *School classification accuracy: Issues for reliability and validity*. Human Resources Research Organization.

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.

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 et al (2023) missing or it should be Dickson & Thacker (2023)?

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.

Hoffman & Wise (2001)

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.

Linn (2002)???

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

Mulolli, D., Crawford, B. F., & Dickinson, E. R. (2023). *The accuracy of students' novice, apprentice, proficient, and distinguished classifications for the 2023 Kentucky. Summative Assessment (KSA) tests* (2024 No. 143). Human Resources Research Organization.

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.

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