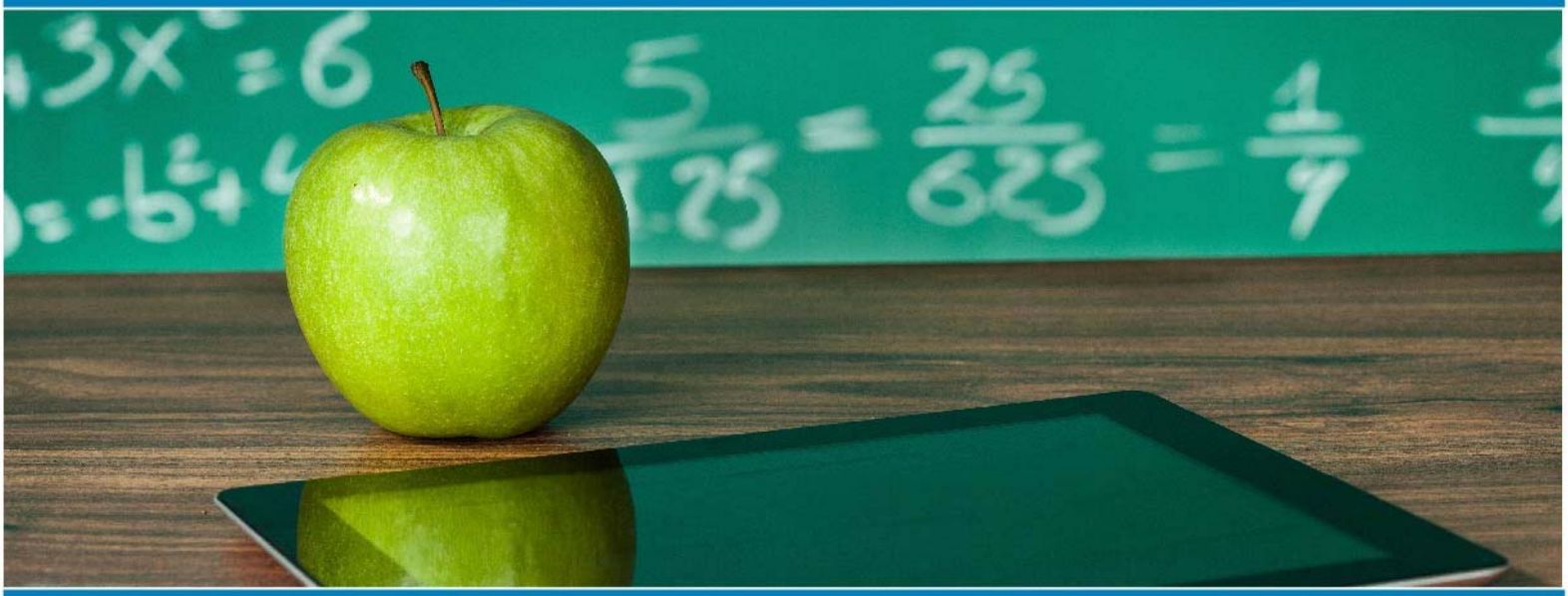
2014 No. 079



The Accuracy of School Classifications for Kentucky's Unbridled Learning Assessment and Accountability System

Prepared for:	1200003303, FFP	Prepared under:	Kentucky Department of Education Capital Plaza Tower, 18th Floor 500 Mero Street Frankfort, KY 40601
Authors:	Andrea L. Sinclair Emily R. Dickinson Arthur A. Thacker	Date:	February, 2015

Headquarters: 66 Canal Center Plaza, Suite 700, Alexandria, VA 22314 | Phone: 703.549.3611 | Fax: 703.549.9025 | **humrro.org**

The Accuracy of School Classifications for Kentucky's Unbridled Learning Assessment and Accountability System

Table of Contents

Chapter 1: Background 1

Chapter 2: KY Next Generation Learner Scores 2
- Introduction and Background 2
- Reliability Issues 2
  - Overall Calculation 2
  - Achievement 3
  - Gap 5
  - Growth 5
  - College and Career Readiness 6
  - Graduation Rates 6
  - Combining the Components 7
- Validity Issues 7
- Year-to-Year Comparison 7
- Discussion and Recommendations 8

Chapter 3: Program Reviews 9
- Introduction and Background 9
- Section 1: Quantitative Investigation of K-3 Program Review Scores 9
  - Item Level Investigations 9
  - Scale Level Analyses 11
  - Model Fit Comparisons 14
  - Summary of K-3 Analyses 16
- Section 2: Convergent Validity Investigation between Program Review Writing and K-PREP Writing 17
- Conclusions 20
- References 21

List of Tables

Table 2.1. Calculation of Next-Generation Learner Components 3
Table 2.2. Average Distribution Error for Each Proficiency Category across Tested Subjects 4
Table 2.3. Hypothetical Examples of the Same Achievement Score Derived from Different Combinations of NAPD Percentages 4
Table 2.4. Descriptive Statistics for Overall Next-Generation Learner Scores 8
Table 2.5. Overall Next-Generation Learner Score Ranges Associated with Each Accountability Classification 8

Table 2.6. Proportions of Variance (r^2) in Overall Next Generation Learner Score Explained by Each Component Separately 8
Table 2.7. Standardized Regression Coefficients for Model Predicting Overall Score from all Components 9
Table 2.8. Descriptive Statistics of Points Values of Next-Generation Learners Score Components by Classification Level 11
Table 2.9. Correlations between Component and Overall Scores: 2012-13 and 2013-14 8
Table 3.1. Internal Consistency Reliability Estimates for each K-3 Scale 12
Table 3.2. Fit Indices for CFA Single-Factor Models for each Scale 14
Table 3.3. Scale Intercorrelations for K-3 15
Table 3.4. Fit Indices for CFA Models 1 – 3 16
Table 3.5. Correlations between Schools' K-PREP Writing Scores and Program Review Writing Scores 19

List of Figures

Figure 2.1. Ranges of achievement points within elementary school classifications. 1
Figure 2.2. Ranges of gap points within elementary school classifications. 2
Figure 2.3. Ranges of growth points within elementary school classifications. 2
Figure 2.4. Ranges of achievement points within middle school classifications. 3
Figure 2.5. Ranges of gap points within middle school classifications. 3
Figure 2.6. Ranges of growth points within middle school classifications. 4
Figure 2.7. Ranges of college & career readiness points within middle school classifications. 4
Figure 2.8. Ranges of achievement points within high school classifications. 5
Figure 2.9. Ranges of gap points within high school classifications. 5
Figure 2.10. Ranges of growth points within high school classifications. 6
Figure 2.11. Ranges of college & career readiness points within high school classifications. 6
Figure 2.12. Ranges of graduation points within high school classifications. 7
Figure 3.1. Example of Single-Factor Measurement Model. 13
Figure 3.2. Simplified Measurement Model for Model 2. 16

The Accuracy of School Classifications for Kentucky's Unbridled Learning Assessment and Accountability System

Chapter 1: Background

The purpose of this report is to document evidence of the accuracy of school classifications under Unbridled Learning Accountability Model as implemented during the 2013-14 school. For several years, HumRRO has calculated the accuracy of school performance classifications as part of its quality control, validation, and research agenda on behalf of the Kentucky Department of Education. Prior calculations of school classification accuracy, though complex, were much more straightforward as they were based solely on assessment data. Under the current implementation of accountability model, a school's classification still relies heavily on student level assessment scores, but also on evidence of the quality of programs offered in arts & humanities, writing, and practical living and career studies.

The passage of Senate Bill 1 in 2009 required Kentucky to begin implementing a new assessment and accountability system during the 2011-12 school year. This marked the beginning of a three-phase approach to transitioning to the full Unbridled Learning Accountability Model in 2014-15. This model combines three major components: next-generation learners, next-generation instructional programs and support, and next-generation professionals, to calculate an overall indicator of the degree to which Kentucky schools and districts meet the expectation that all students are being prepared for success in college and in their careers. During the 2013-14 school year, only the next-generation learners and next-generation instructional programs and support were included in the calculation of school accountability scores.

These changes present several challenges for quantifying the accuracy of school performance classifications. Whereas school accountability scores in the past consisted only of a weighted average of student-level assessment scores within a school, they are now derived in part from multiple combinations of NAPD student classifications that include an allowance for the percentage of higher scoring students to offset the impact of the percentage of lower scoring students. Similarly, whereas accountability SEMs under the prior model were a weighted sum of assessment SEMs, such an approach is no longer applicable. The inclusion of student growth percentiles in the next-generation learners score further complicates the calculation because there is no clear consensus on how to quantify error variance in SGP scores. At the high school level, the possible combinations of multiple assessments for measuring college and career readiness creates yet another obstacle to identifying a common metric for error variance in accountability scores across schools. Overall accountability scores are also based on more than just the next-generation learner scores that have been the focus of this report. The next-generation instructional programs and supports component is based on self-reports which also lack a method for quantifying measurement error.

Given these changes, prior approaches to school classification are no longer appropriate, nor are direct comparisons to prior reports on school classification accuracy. This report summarizes a series of investigations related to the next-generation learners and next-generation programs and supports components of the Unbridled Learning Model. Chapter 2 explores issues for the reliability and validity of next-generation learners scores. Chapter 3 includes a quantitative analysis of K-3 program review data and a convergent validity investigation of program review writing and K-PREP writing.

Chapter 2: KY Next Generation Learner Scores

Introduction and Background

For the 2011-12 school year, overall accountability scores for schools and districts were based solely on the next-generation learners component, itself built upon several facets of student-level performance. Depending on the grade range, the next-generation learners score is based on three to five of the following components: achievement, gap, growth, readiness for college and career, and graduation rate.[1] In 2012-13 and 2013-14, a weighted combination of next-generation learners and next-generation instructional programs and support scores were used to determine the overall accountability score. In 2014-15, all three components will be combined, with 70% of the overall score attributed to the next-generation learners score, 20% to the next-generation instructional programs and support score, and 10% to the next-generation professionals score.

During each phase of implementation, interim cut scores have been set to determine schools' and districts' classification into Needs Improvement, Proficient, Distinguished, or School of Distinction categories. These cut scores were determined by locating the scores at the 70th, 90th, and 95th percentiles, respectively. Until all components of the accountability model are operational, cut scores will be reset each year. In 2014-15, cut scores will be locked in for a five-year period to give each school/district a clear improvement goal to work toward. These improvement goals include requiring schools/districts whose overall score is below the Proficiency cut to improve that score by one-third of a standard deviation over the five-year period, and requiring schools/districts identified as Proficient or higher to improve their overall score by approximately one-sixth of a standard deviation over five years.

From 2012-13 to the present, next-generation learner scores contributed 77% of the overall accountability index score, while next-generation instructional programs and support contributed 23%. The former, and the strongest contributor to overall scores, is largely based around students' scores on assessments (plus graduation rate at the high school level) while the latter is derived from ratings assigned by school-level committees. HumRRO routinely conducts studies of the accuracy of student-level classifications into Novice, Apprentice, Proficient, and Distinguished (NAPD categories), and therefore has some empirical evidence on which to base judgments about school classifications that are built in large part around these student-level classifications. Since no similar evidence is available regarding the reliability of program review scores, the remainder of this study will focus on the next-generation learners component only.

Reliability Issues

This section of the report will discuss issues related to the reliability of next-generation learners 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.

Overall Calculation

A school's next-generation learner overall score is based on three to five components, depending on the grade range. Elementary schools' scores are derived from achievement, gap, and growth components. Middle schools include these components as well as a college and

[1] Details about next-generation learners score calculations are available at http://education.ky.gov/comm/UL/Documents/WHITE%20PAPER%20062612%20final.pdf.

career readiness component. High schools include these same four components as well as a graduation component. Table 2.1 summarizes the calculation of each of the five components.

Table 2.1. Calculation of Next-Generation Learner Components

Component	Points Calculation
Achievement	Percent of students scoring Distinguished + Percent of students scoring Proficient + 0.5 (Percent of students scoring Apprentice) + 0.5 (Percent of students scoring Distinguished) – 0.5(Percent of students scoring Novice)
Gap	Percent of non-duplicated count of Gap Group students scoring Proficient or Distinguished
Growth	Percent of students showing typical or higher growth based on Student Growth Percentiles.
CCR	**Middle school**: Percent of students meeting benchmarks on EXPLORE. **High school**: Percentage of students who have either met performance benchmarks on ACT, completed college placement tests, or attained career academic and technical benchmarks.
Graduation	The percent of students completing graduation requirements.

Notes: Gap Group is defined as students who are African American, Hispanic, Native American, Special education, Free/Reduced lunch-eligible, and/or Limited in English proficiency.

Each component is weighted and summed to calculate the overall score, and these weightings also vary by grade span. At the elementary level, growth is weighted slightly higher (40%) than achievement and gap, which are weighted equally (30%). At the middle school level, achievement, gap, and growth are equally weighted (28%), while college and career readiness contributes 16%. Among high schools, the five components are equally weighted at 20%.

The information on component calculations and different approaches to weighting illustrates the complexity of the new approach to measuring accountability, and this level of complexity has implications for the measurement of classification accuracy. The next sections will discuss the complications posed by each component of the overall next-generation learner score.

Achievement

The achievement component of the next-generation learner score is designed to recognize schools for students scoring above the Novice level of performance on K-PREP (grades 3 through 8) or end-of-course (high school) assessments. This is done by awarding a full point for each percent of students scoring Proficient or higher, as well as a half-point for each percent of students scoring in the Apprentice category. Bonus points may be awarded for the percentage of students scoring Distinguished (half-point per percent of students scoring at this level is also added), but this value is offset by the percentage of students scoring Novice (half-point per percent of students scoring at this level subtracted). Having more Novice than Distinguished students would result in no Distinguished bonus, but would not reduce the overall score any further (i.e., the percentage of Novice students won't reduce the overall score based on percent Proficient alone, but will reduce the amount of bonus points possible). Overall scores are also capped at 100 to reflect their basis on percentages. Schools with large Distinguished bonuses have their scores truncated at 100. The range of achievement points across all grade levels for the 2013-14 school year was 30.8 to 100.

It has been documented that student level classifications vary in terms of the amount of classification error, both across grade/subjects and across the performance categories.[2,3] 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, and this has implications for school level classification accuracy when some student-level classifications are weighted more heavily than others.

Table 2.2 illustrates the average distribution error across test content areas within each grade level for each student classification category. 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 K-PREP assessment demonstrates levels of classification accuracy that are comparable or higher than other state assessments.

Table 2.2. Average Distribution Error for Each Proficiency Category across Tested Subjects

	Novice	Apprentice	Proficient	Distinguished
Grade 3	0.81	2.23	0.94	1.17
Grade 4	1.44	1.38	2.58	2.23
Grade 5	1.35	1.12	1.00	1.54
Grade 6	4.57	1.75	3.60	0.68
Grade 7	1.94	1.66	1.69	1.85
Grade 8	1.23	1.01	2.35	2.13
High School	0.84	1.13	0.82	1.10

Notes: 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 .81%.

Table 2.2 demonstrates that although average levels of misclassification may be quite low overall, they do vary in magnitude across the NAPD categories, and across grade levels. The achievement component of the overall score is derived from some combination of the percentages of students scoring at each level, but the same score may reflect different combinations of these percentages. Table 2.3 depicts a hypothetical example.

Table 2.3. Hypothetical Examples of the Same Achievement Score Derived from Different Combinations of NAPD Percentages

	% Novice	% Apprentice	% Proficient	% Distinguished	Total Score

[2] Dickinson, E. R., Levinson, H., Thacker, A. A., & Hoffman, R. G. (2013).The accuracy of students' novice, apprentice, proficient, and distinguished classifications for the 2012 Kentucky performance rating for educational progress (K-PREP) tests (2013 No. 037). Alexandria, VA: Human Resources Research Organization.

[3] Dickinson, E. R., & Thacker, A. A. (2014).The accuracy of students' novice, apprentice, proficient, and distinguished classifications for the 2013 Kentucky end-of-course (EOC) tests. (2014 No. 024). Alexandria, VA: Human Resources Research Organization.

School 1	0	48	52	0	76
School 2	22	20	20	38	76

It is not surprising that there are multiple ways to achieve a particular achievement score, as demonstrated in Table 2.3. The Unbridled Learning 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. These differences in distributions of NAPD percentages do have implications for the measurement of classification accuracy, however. A school's achievement score that does not include Novice performing students, and therefore does not rely on the accuracy of student-level Novice classifications, will likely differ in its level of accuracy from a school whose achievement score does rely in part on the accuracy student-level Novice classifications.

Gap

The gap component of the next-generation learner score is designed to recognize schools for Proficient or higher performance among student members of subgroups that have historically experienced gaps in achievement. This is done by calculating the percentage of students performing at these levels based on a non-duplicated student count. In other words, though an individual student may belong to multiple subgroups identified as part of the larger gap group, each student would only be counted once. The range of gap points across all grade levels for the 2013-14 school year was 10.0 to 84.1.

Concerns have been raised regarding the possibility for such a model to mask the underperformance of particular student subgroups when all subgroups are essentially combined. In response to these concerns, the Kentucky Department of Education (KDE) has since required that the performance of individual subgroups be documented and that schools/districts can receive consequences when any individual subgroups are performing significantly below average.

Issues for classification accuracy within the gap component are similar to those described for the achievement component. Though the gap component is based on only the Proficient and Distinguished student-level classifications, schools may vary in the proportion of gap points that come from Proficient vs. Distinguished scoring students and these differences would affect the magnitude of classification error at the school level.

Growth

The growth component of the next-generation learner score is designed to recognize schools for students demonstrating growth on achievement measures at rates considered to be typical or higher compared to their academic peers.[4] For grades 4 through 8 (growth is not calculated for grade 3 because they do not have prior year scores), growth percentiles are calculated using scores on K-PREP, whereas high school growth is determined by comparing scores on ACT in 11th grade to those on PLAN in 10th grade. The range of growth points across all grade levels for the 2013-14 school year was 21.5 to 84.6.

[4] Betebenner, D. W. (2008). Norm- and criterion-referenced student growth. Retrieved from http://www.nciea.org.

The growth component differs from the achievement and gap components in part because it does not rely on student classifications. It does, however, rely on student assessment scores that contain some amount of measurement error. Criticisms of student growth percentile approaches include concerns over the effects of measurement error on identifying students' academic peers; depending on the reliability of a particular assessment, students may be misidentified as having the same prior grade-level test score (i.e., their underlying true scores are actually different) and this would have implications for the accuracy of growth percentiles.[5] Standard errors of measurement typically calculated during psychometric processing could be used to quantify the likelihood of a true score being some other possible score than what was observed. The implications of measurement error for student growth percentiles is an area where research is needed, however.[6]

College and Career Readiness

The college and career readiness component of the next-generation learner score is designed to recognize schools for students demonstrating readiness for postsecondary success. These readiness indicators at the high school level include meeting benchmarks (or passing) tests such as ACT, COMPASS, KYOTE, ASVAB, ACT WorkKeys, or KOSSA, or receiving an industry-recognized career certificate. Students may meet one or more of these indicators, so points for this component of the next-generation learners overall score are calculated based on a non-duplicated count across the indicators. Schools receive a point for each percentage of students identified as college and/or career ready. Bonus points are calculated by giving a half-point for each graduate meeting both the college and career ready benchmarks, dividing this value by the total number of graduates, and then adding this value to the initial percentage. Middle school college and career readiness is determined by the percentage of students meeting benchmarks on ACT Explore. The range of college and career readiness points across middle and high schools for the 2013-14 school year was 13.9 to 100.

College and career readiness, particularly at the high school level, is another component of the overall next-generation learners score that relies on several different assessment instruments which may be used in a wide array of 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.

Graduation Rates

The graduation component of the next-generation learner 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 2013-14 school year was 72.5 to 100.

---

[5] Haertel, E. (2009, December). Student growth data for productivity indicator systems. Paper presented at the Exploratory Seminar: Measurement Challenges within the Race to the Top Agenda, Princeton, NJ. Retrieved from http://www.k12center.org/rsc/pdf/HaertelPresenterSession2.pdf.

[6] Castellano, K. E. (2011). Unpacking student growth percentiles: statistical properties of regression-based approaches with implications for student and school classifications. Retrieved from http://ir.uiowa.edu/etd/931/.

Although this component of the next-generation learners score does not include multiple data sources or complex calculations, it does present its own limitation for the calculation of school-level classification accuracy. Because graduation rates are a single,-self reported value, there is no method for estimating their error variance. Prior research has assumed a combined reliability estimate of .7 to calculate error rates for graduation and other non-academic school-level performance indicators in the context of school classification accuracy[7]. This would not be appropriate here because graduation rate, barring simple human error, can be calculated without measurement error.

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. This process carries with it a number of assumptions about those item-level scores, most notably that they are independent of one another, or that the probability of one item response is not dependent on another item response. The preceding description of the calculation of next-generation learners scores illustrates how such an assumption would be violated. Because several component scores are built upon common information about student-level performance, then the probability of one component score is not entirely independent of other component scores. If this assumption is ignored, then classification accuracy would be overestimated by failing to appropriately combine the probabilities of individual component scores.

We have attempted to demonstrate the level of complexity of the next-generation learners overall accountability score and the issues this raises for quantifying the level of classification accuracy at the school and district level. Most significantly, the system lacks a common metric for variance across all components, which would be required if the error variance of the components were to be combined to determine an overall classification accuracy rate.

Validity Issues

This section of the report will discuss issues related to the validity of next-generation learners scores. Of particular interest are the relations between the component scores and the overall next-generation learner scores and the associated issues related to the interpretability of overall scores to stakeholders.

Schools are classified in part based on their overall next generation learners index score. Program reviews also impact classifications. This means that there is no simple next generation score that is associated with a shift in classification. Schools with identical next generation scores may be classified differently. This is why there is overlap in Table 2.4, which presents the range, mean, and standard deviation of overall scores for each grade span, while Table 2.5 highlights the lowest and highest overall score associated with each classification label for the three grade ranges. Table 2.5 indicates that the Needs Improvement category has the largest score ranges across the grade spans (score ranges between 23.4 and 42.5 points), while the Proficient ranges are much smaller (score ranges between 8.2and 10.8points). Also of note is

[7] 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). Alexandria, VA: Human Resources Research Organization.

the level of overlap between the accountability classifications in terms of next-generation learner scores. For example, one middle school classified as Needs Improvement has an unweighted next- generation learner score that is approximately 14 points higher than a middle school classified as Proficient.

Table 2.4. Descriptive Statistics for Overall Next-Generation Learner Scores

	Minimum	Maximum	Mean	Standard Deviation
Elementary (N=720)	26.1	89.2	60.69	9.21
Middle (N=329)	28.2	83.6	57.09	8.95
High (N=228)	41.6	89.3	63.91	7.52

Note: Based on unweighted overall next generation learner scores (i.e., not multiplied by .77).

Table 2.5. Overall Next-Generation Learner Score Ranges Associated with Each Accountability Classification

	Elementary		Middle		High	
	Low	High	Low	High	Low	High
Needs Improvement	26.1	66.6	28.2	70.7	41.6	65.0
Proficient	60.3	71.1	56.9	76.3	61.3	69.5
Distinguished	67.6	80.1	63.6	82.2	68.2	85.9
School of Distinction	71.5	89.2	65.6	83.6	71.5	89.3

Table reads: The lowest unweighted overall school next-generation learner score among elementary schools classified as Needs Improvement was 26.1. The highest unweighted overall next-generation learner score among Elementary schools classified as Needs Improvement was 66.6.

Because overall scores are comprised of three to five component points values, the next question of interest was the relative contribution of each component to the overall score. To address this question, we first correlated each component independently with the overall score. Squaring that value would then provide an indication of the proportion of variability in the overall score that could be explained by each component alone. As Table 2.6 indicates, achievement is the strongest predictor and gap is the second strongest predictor of overall next-generation learner score across the three grade levels.

Table 2.6. Proportions of Variance (r^2) in Overall Next Generation Learner Score Explained by Each Component Separately

	Elementary	Middle	High
Achievement	.861	.927	.819
Gap	.834	.874	.738
Growth	.621	.640	.517
CCR	NA	.743	.621
Graduation	NA	NA	.239

Notes: CCR= College and Career Readiness. Each component was correlated separately with the overall next-generation learners score.

Table reads: Achievement points explain approximately 86% of the variability in overall next-generation learners score among elementary schools when no other components are included in the prediction equation.

Next, regression models were run predicting overall score from the component points, entering the components into the model in order based on the magnitude of the explanatory value of the components presented in Table 2.6. While the results in Table 2.6 give a sense of the contribution of each component separately, Table 2.7 provides an indication of the contribution of each component to the overall score when all included components are taken into account.

Similar to Table 2.6, Table 2.7 indicates that the achievement component contributes most to the prediction of overall score at the elementary and middle school levels. At the high school level however, the college and career readiness component is the strongest predictor of overall score after accounting for the other components.

Table 2.7. Standardized Regression Coefficients for Model Predicting Overall Score from all Components

	Elementary	Middle	High
Achievement	.407	.350	.289
Gap	.356	.313	.234
Growth	.378	.233	.221
CCR	NA	.213	.401
Graduation	NA	NA	.127

Notes: CCR= College and Career Readiness. Components were added into the regression model in the order of the magnitude of values in Table 2.6.

Table reads: For every 1 unit increase in achievement points, overall next-generation learners score is expected to increase .407 standard deviations.

Another approach to understanding the different contributions of the component scores to school classifications is to explore 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. Table 2.8 presents descriptive statistics for the points values associated with each component within each classification level. One straightforward way to compare the score distributions within groups is to calculate a standardized mean difference score (Cohen's *d*) of adjacent categories. This statistic is based on the mean ($\overline{x}$), standard deviation (*s*), and sample size (*n*) of two groups and provides a standard metric for comparing group differences:

$$\text{Cohen's } d = \overline{x}_{\text{Group 1}} - \overline{x}_{\text{Group 2}})/s_{pooled} \qquad (1)$$

$$s_{pooled} = \sqrt{\left(\left(\left(n_{\text{Group 1}} - 1\right) * s_{\text{Group 1}}\right)\right) + \left(\left(n_{\text{Group 2}} - 1\right) * s_{\text{Group 2}}\right)\right) / \left(n_{\text{Group 1}} + n_{\text{Group 2}} - 2\right)\right)} \qquad (2)$$

Cohen's *d* values for most components across the grade levels tended to be high using established criteria (0.8 or greater), though a few would be considered medium (0.5-0.79)in magnitude. Cohen's *d* values were small (0.2-0.49) for five group comparisons. For example, mean differences between Schools of Distinctions and Distinguished schools were small on gap and growth components at the middle school level, and on growth component at the high school level. Small average differences between Schools of Distinction and Distinguished schools are not surprising given that these groups represent the 95th and 90th percentiles, respectively. Small mean differences were also observed between Distinguished and Proficient high schools and Proficient and Needs Improvement high schools on the graduation component. This is also not surprising considering that points on the graduation component tend to be high across all high schools.

Table 2.8. Descriptive Statistics of Points Values of Next-Generation Learners Score Components by Classification Level

	Elementary Schools														
	Needs Improvement			Proficient				Distinguished				School of Distinction			
	N	Mean	SD	N	Mean	SD	*d*	N	Mean	SD	*d*	N	Mean	SD	*d*
Achievement	371	64.88	8.93	187	78.07	5.26	1.80	93	86.96	5.18	1.70	69	92.10	4.25	1.08
Gap	371	40.05	7.11	187	51.51	4.80	1.89	93	57.27	5.93	1.07	69	65.36	7.55	1.19
Growth	365	55.47	7.63	185	62.95	5.85	1.10	92	66.89	4.44	0.76	66	71.23	4.92	0.93
	Middle Schools														
	Needs Improvement			Proficient				Distinguished				School of Distinction			
	N	Mean	SD	N	Mean	SD	*d*	N	Mean	SD	*d*	N	Mean	SD	*d*
Achievement	174	63.82	7.75	86	76.75	3.90	2.11	30	83.57	5.46	1.44	39	87.37	5.09	0.72
Gap	173	37.57	7.05	86	48.21	4.64	1.78	30	55.53	6.93	1.24	39	57.28	7.85	0.24
Growth	174	55.78	6.37	86	62.50	4.98	1.18	30	66.01	4.98	0.70	39	67.83	4.32	0.39
CCR	171	41.02	8.68	85	52.80	7.57	1.45	30	58.65	7.27	0.79	35	64.38	8.67	0.72
	High Schools														
	Needs Improvement			Proficient				Distinguished				School of Distinction			
	N	Mean	SD	N	Mean	SD	*d*	N	Mean	SD	*d*	N	Mean	SD	*d*
Achievement	88	52.94	6.57	81	63.02	5.09	1.72	40	71.70	7.82	1.32	19	80.24	8.20	1.07
Gap	87	29.82	6.03	81	36.20	4.26	1.22	40	43.73	6.33	1.40	19	51.29	7.93	1.05
Growth	88	49.24	6.10	81	56.07	5.80	1.15	40	61.05	6.93	0.78	19	64.56	8.98	0.44
CCR	88	61.70	13.98	81	76.34	9.88	1.21	40	85.42	10.49	0.89	19	90.04	6.75	0.52
Graduation	88	90.91	5.65	81	92.69	3.83	0.37	40	94.09	3.17	0.40	19	96.84	3.00	0.89

Notes: *d*= Cohen's *d* for adjacent groups. Proficient is compared to Needs Improvement, Distingished is compared to Proficient, and School of Distinction is compared to Distinguished.

Visual depictions of the distributions of component scores are another useful way to compare how the classification levels differ. Figures 2.1 through 2.12 depict the stair-step pattern between overall classification and each component of the next-generation learners score. 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 and stars that appear beyond the vertical lines depict outliers or extreme values. For the achievement and gap components 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 distributions of growth, college and career readiness, and graduation points across the accountability classifications.

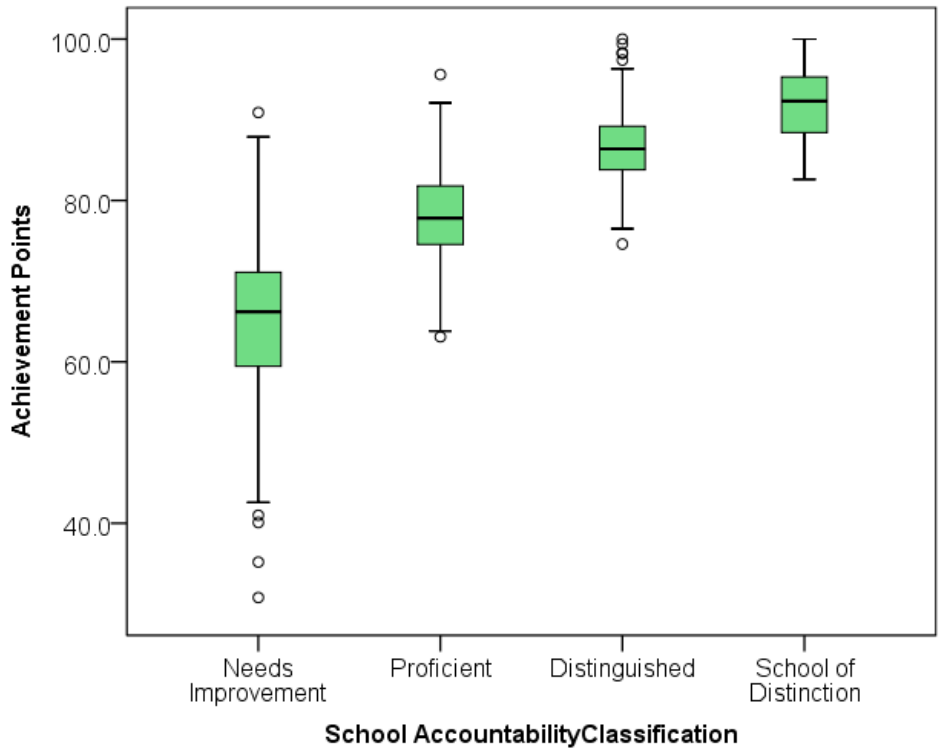


Figure 2.1. Ranges of achievement points within elementary school classifications.

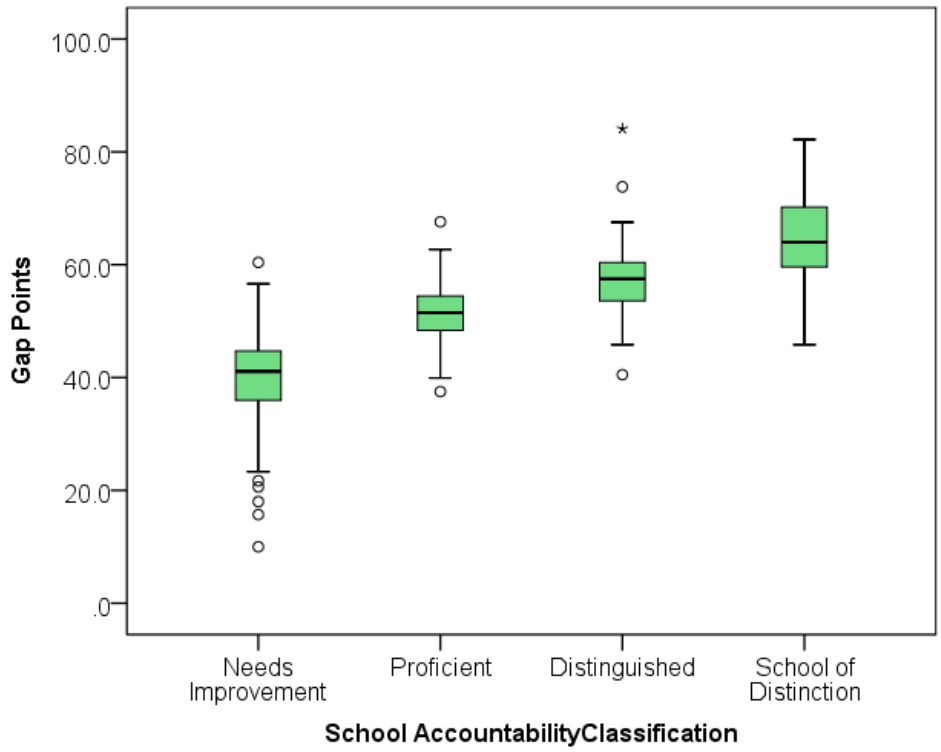


Figure 2.2. Ranges of gap points within elementary school classifications.

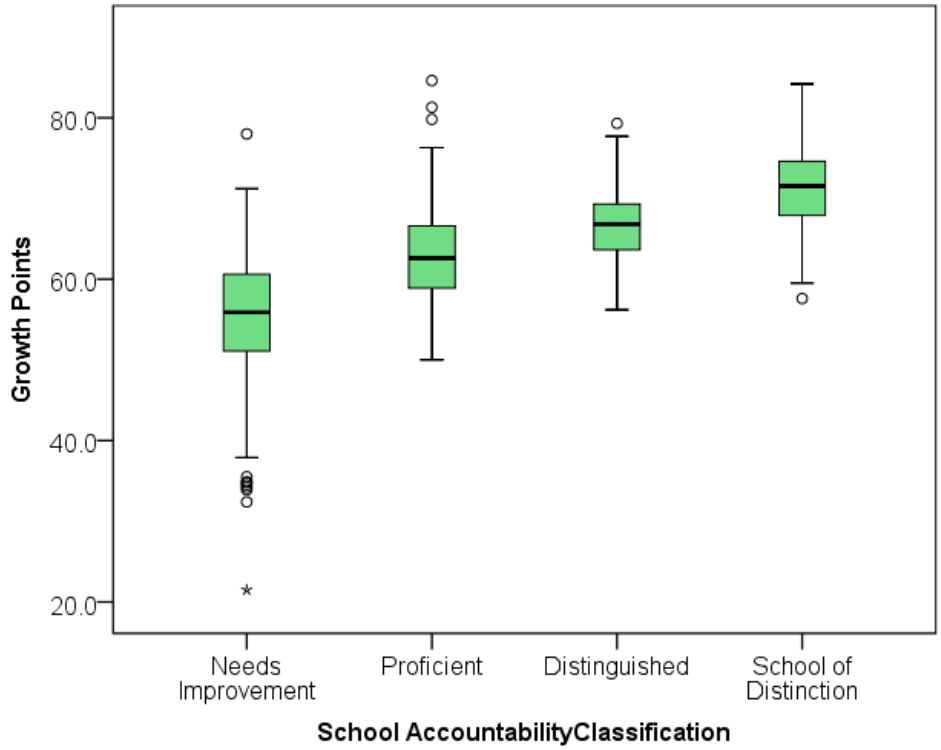


Figure 2.3. Ranges of growth points within elementary school classifications.

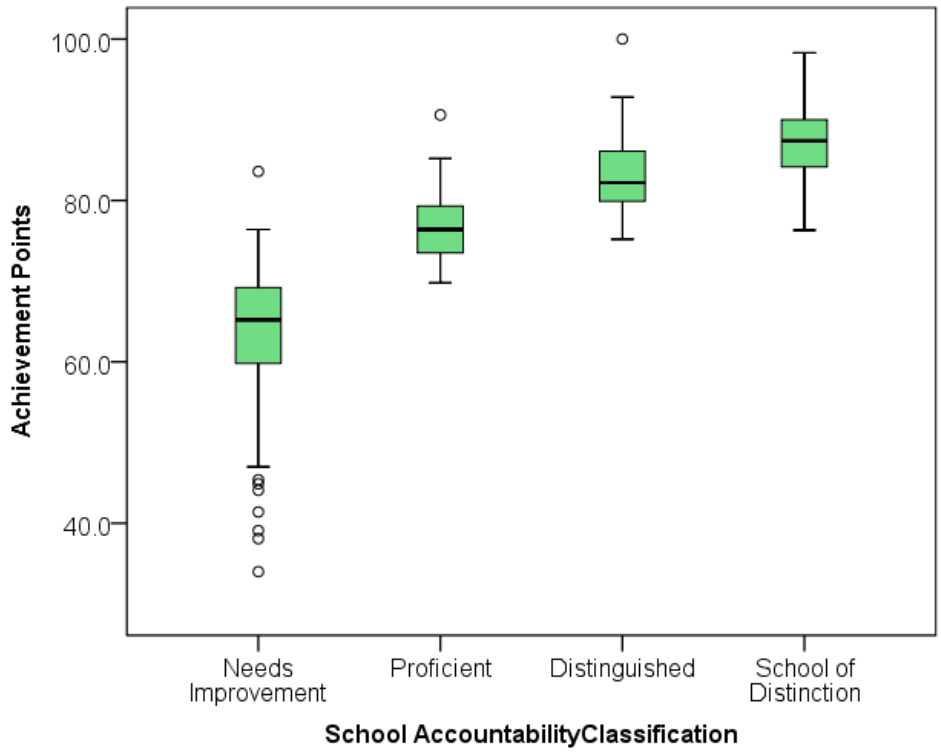


Figure 2.4. Ranges of achievement points within middle school classifications.

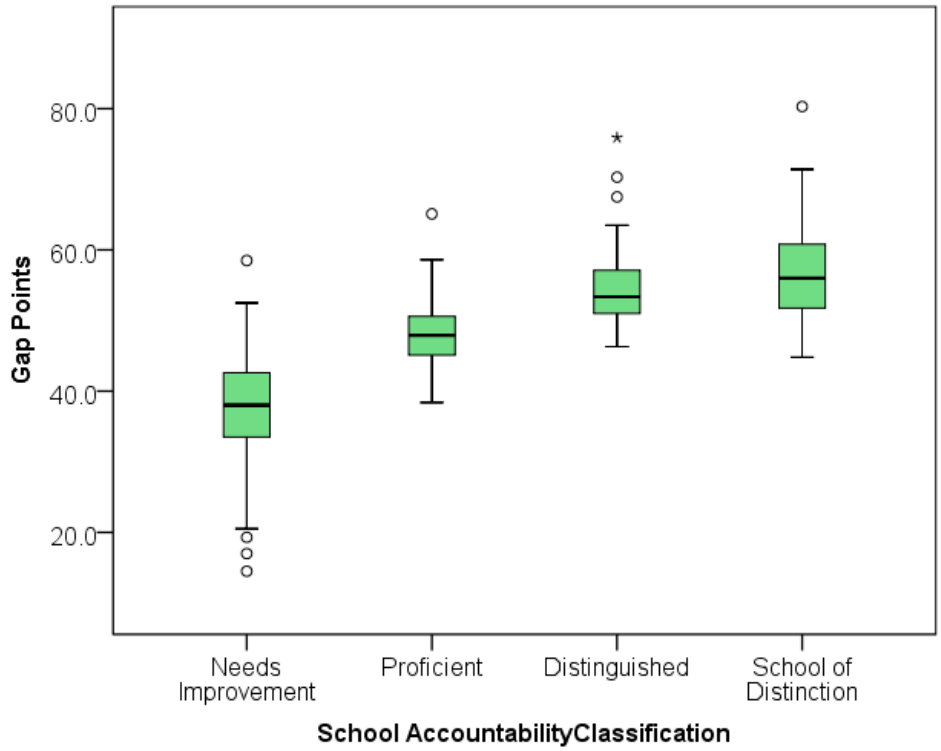


Figure 2.5. Ranges of gap points within middle school classifications.

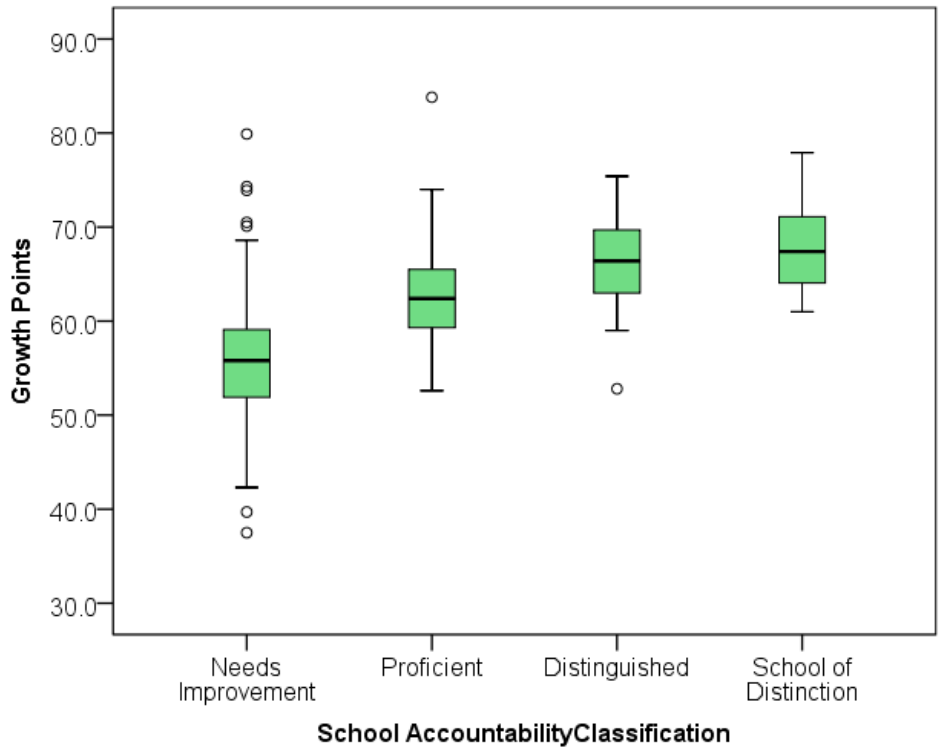


Figure 2.6. Ranges of growth points within middle school classifications.

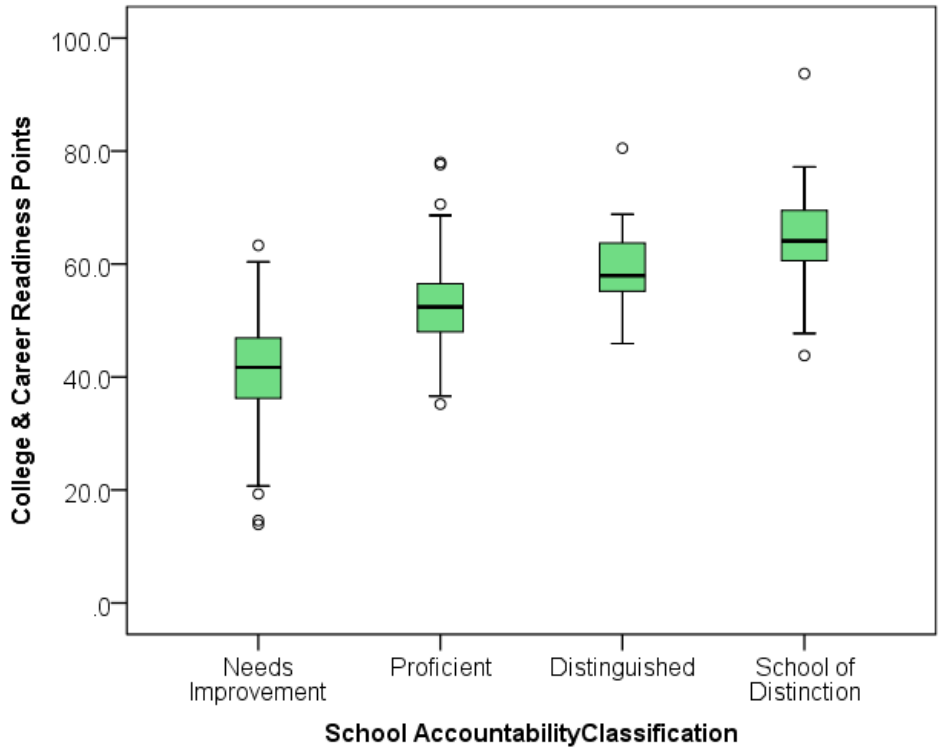


Figure 2.7. Ranges of college & career readiness points within middle school classifications.

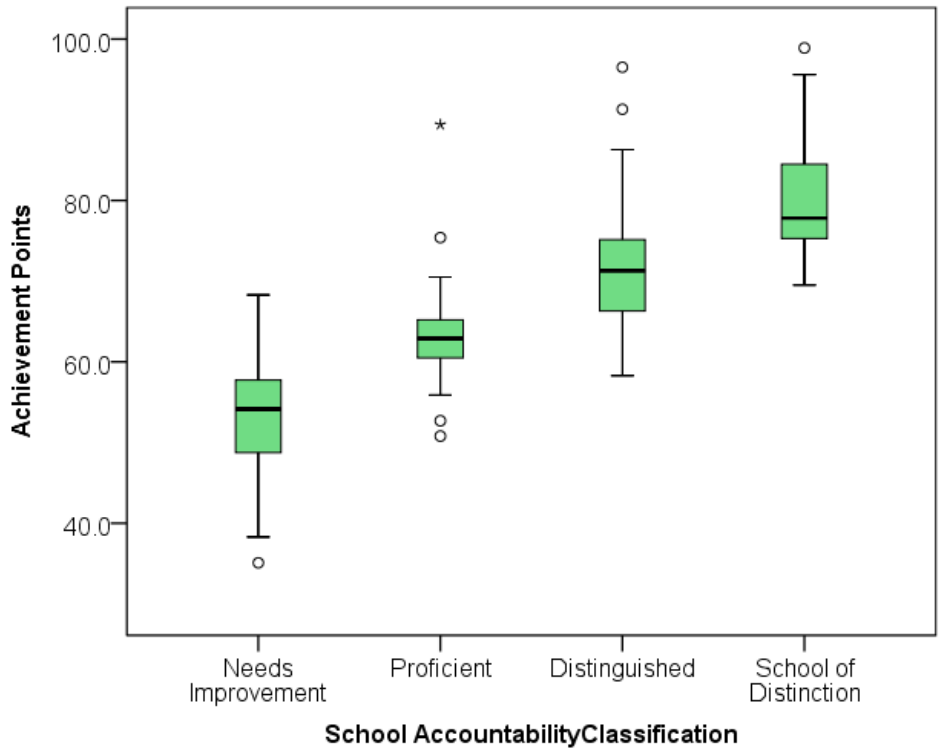


Figure 2.8. Ranges of achievement points within high school classifications.

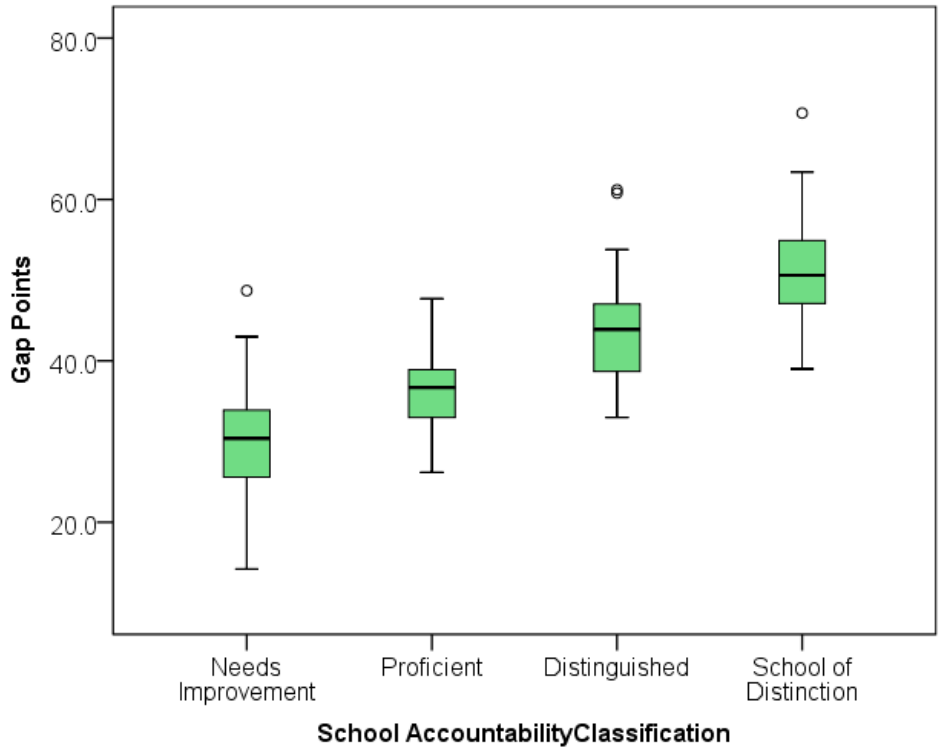


Figure 2.9. Ranges of gap points within high school classifications.

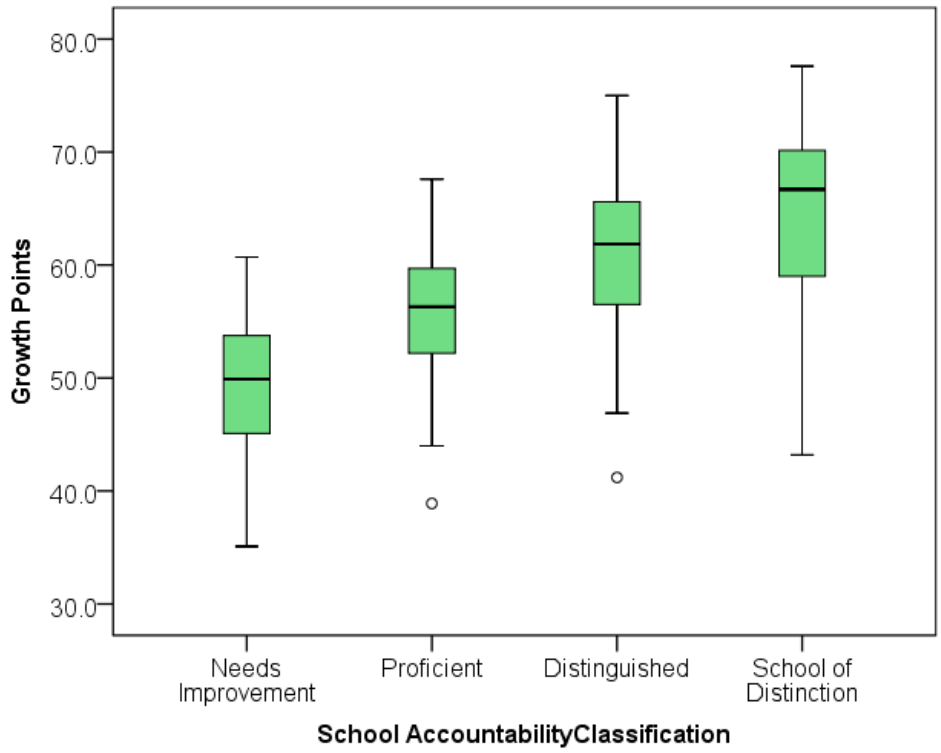


Figure 2.10. Ranges of growth points within high school classifications.

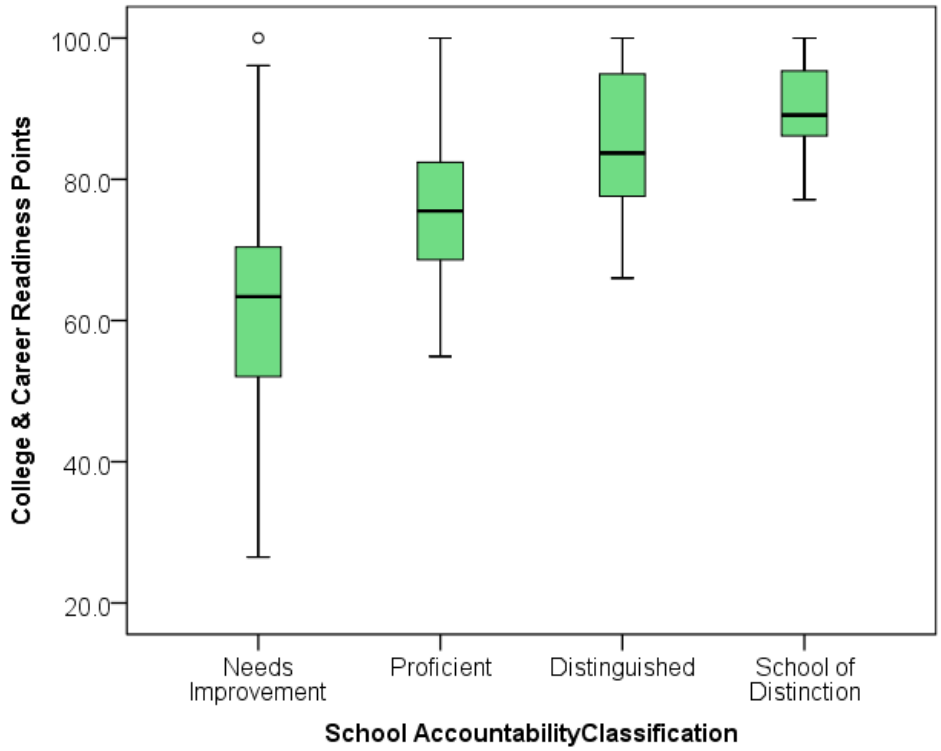


Figure 2.11. Ranges of college & career readiness points within high school classifications.

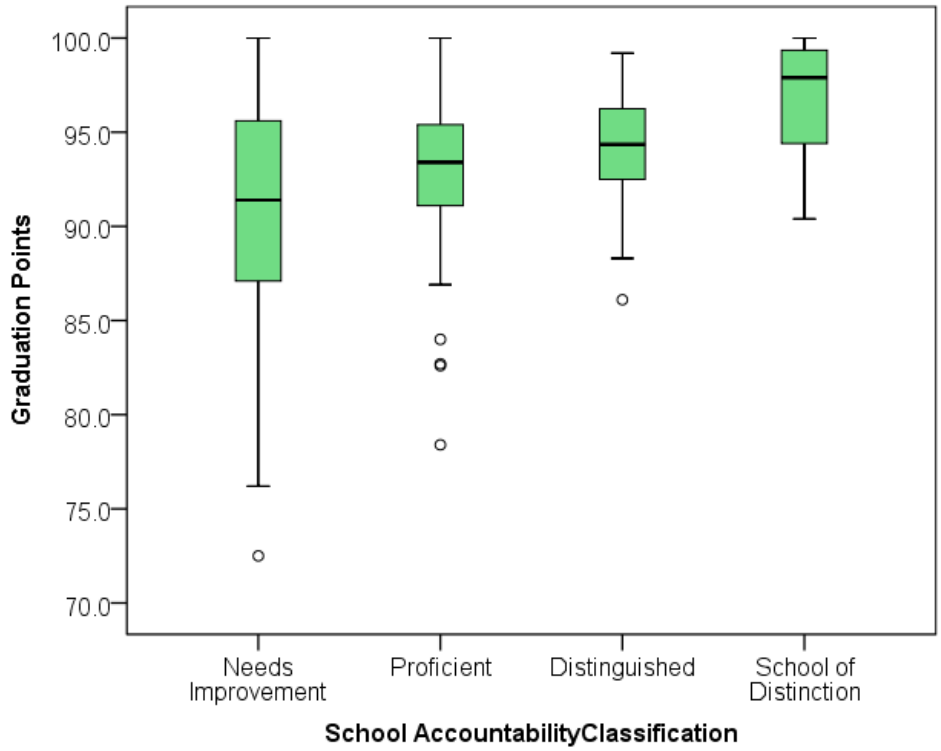


Figure 2.12. Ranges of graduation points within high school classifications.

Year-to-Year Comparison

One way to document both reliability and validity within the next-generation learners accountability system is to determine the correlation between scores across multiple years of system implementation. Correlations between scores would provide some indication of the consistency of measurement over time, as well as the appropriateness of inferences about school performance based on accountability scores.

Correlations between school accountability scores from 2012-13 and 2013-14 are expected to be positive and moderate-to-large for two major reasons. First, given Kentucky schools' commitment to providing high-quality educational experiences to its students, we would not anticipate that a large number of schools would experience a decline in their accountability score over time. In fact, very low or negative correlations would raise questions about the appropriateness of the next-generation learners accountability model for making inferences about school performance. Second, even assuming schools' best efforts to improve performance, it is reasonable to expect that some schools will maintain their score status, while other schools will either improve considerably or lose some ground. As such, we would not expect a perfect correlation from year to year either.

Table 2.9 presents correlations between both the component scores and the overall scores for schools with available data for 2012-13 and 2013-14. As expected, all correlations are positive and moderate to large. The weakest correlations were between the growth components, particularly among elementary and middle schools. This is likely

due to the growth component being the most sensitive to regression toward the mean. The growth component has comparatively larger error, meaning that measurement error may contribute substantially to very high or very low scores. In subsequent years, there will be a tendency for the scores to be closer to the mean because the error is distributed randomly among all of the scores. It is less likely the error will cause scores to be impacted in the same direction two years in a row. Table 2.9. Correlations between Component and Overall Scores: 2012-13 and 2013-14

	Achievement	Gap	Growth	C & C Readiness	Graduation	Overall
Elementary	.901	.817	.374	NA	NA	.765
Middle	.873	.815	.339	.788	NA	.838
High	.885	.777	.609	.760	.920	.883

Discussion and Recommendations

The present study has sought to highlight issues related to the reliability and validity of school accountability scores in their current form, with a focus on the next-generation learner component. There is evidence to support the validity of inferences about school performance based on the next-generation learner scores, but areas of overlap in component scores may make distinctions between the classifications that are difficult for practitioners to understand. It may be beneficial to revisit the relative weighting of components within the next-generation learners score calculation and/or to identify objective criteria within each component and use an approach akin to the standards setting procedures typically used for identifying assessment benchmarks.

As the new accountability model continues to be rolled out, it is important to take into consideration limitations of the model. While the Unbridled Learning model may reflect improvements over prior approaches to measuring school accountability from a validity standpoint, these changes have led to increased difficulty in quantifying the accuracy and reliability of school classifications.

As Kentucky considers the implications of the accountability model, it would be beneficial to determine how the component level results are used by schools. If schools tend to focus on any particular component, or set of components, it may be beneficial to consider the validity of those components and their potential for exhibiting change. Schools can be expected to actively address those components where they think they can have the largest impact. Kentucky's accountability system will benefit from knowing where schools place their focus.

Chapter 3: Program Reviews

Introduction and Background

This chapter represents the third in a series of HumRRO investigations of program reviews in Kentucky public schools. A prior, separate report included analyses of the data collected in the pilot year (2011-2012) of the program reviews for Arts & Humanities, Practical Living & Career Studies (PL/CS), and Writing (Sinclair & Thacker, 2013a). This included a quantitative investigation of the program review scores in which items (i.e., "characteristics") comprising the program review measures were flagged based on several criteria such as, little or no variance in ratings, very high item-total correlations, or very low item-total correlations. Flagged items were recommended for additional review by content experts. The first report also included guidance on setting performance cut scores on each of the program review areas. Finally, recommendations were also provided for modifying the ASSIST software system used for collecting performance review ratings to help improve the quality of the data collected from schools. A second report (Sinclair & Thacker, 2013b) further investigated the quantitative properties of the data collected for Arts & Humanities, PL/CS, and Writing in their first operational year (2012-2013), with a focus on scale reliability and unidimensionality. The second report also compared performance level classifications based on compensatory and non-compensatory scoring models. The recommendation was made to continue with the compensatory scoring model that was applied to the 2012-2013 data because it produced more favorable results and has the benefit of greater parsimony.

This chapter is presented in two sections. First, it includes a quantitative analysis of the data collected on the Kindergarten through Grade 3 (K-3) program review area from its first operational year (2013-2014). Similar quantitative analyses that were applied to the data from the Arts & Humanities, PL/CS, and Writing program review areas (Sinclair & Thacker, 2013a; 2013b) are applied to the data from the K-3 program review. Second, this report also includes a convergent validity investigation between schools' K-PREP on-demand writing scores and their program review Writing scores.

Section 1: Quantitative Investigation of K-3 Program Review Scores

To gain insight into the quality and usefulness of the data obtained from the K-3 program review several analyses were conducted. All analyses were conducted on the 734 schools for which K-3 data was available from the 2013-2014 academic year.

Item Level Investigations

First, we investigated the frequency distribution of ratings on each K-3 item. Schools rated the K-3 items on a 4-point scale where 0 = No Implementation, 1 = Needs Improvement, 2 = Proficient, and 3 = Distinguished. Items were flagged if they demonstrated little variance in ratings. "Little variance" was operationalized as 80% or more of the schools providing the same rating on an item. If an item has little variance (i.e., nearly all schools provide the same rating), then the item may not be very informative. That is, if the overwhelming majority of schools provide the same rating, then that item contributes very little information with regard to making determinations about the relative level of a school's performance. Furthermore, because such items have little variance, they do not co-vary with other items in meaningful ways. We investigated the frequency distribution of ratings for each of the K-3 items, and found that there were no items for which 80% or more of schools selected the same rating. Consequently, none of the 28 items that comprise the K-3 program review measure appear to suffer from a severe

lack of variance in ratings based on an investigation of the frequency distributions. As an additional check, we investigated the standard deviation of each item. There were two items that had standard deviations of less than 0.50 on the four-point scale, indicating that schools responded very uniformly to these two items—meaning that these items did not discriminate very effectively between schools with stronger K-3 programs and schools with weaker K-3 programs. Those two items were: “To what extent do teachers consistently collaborate with others on their team or grade level to plan instructional units, including common assessments and supplemental activities to ensure that each student has access to the curriculum and supports necessary to attain the curriculum?” and “To what extent does the School leadership/SBDM committee continually monitor the availability of resources in an effort to thoughtfully allocate sufficient blocks of instructional time and developmentally appropriate resources needed to support an effective K-3 program?” The low standard deviations for these items suggest that these items might benefit from additional review by content experts. Perhaps the item stems and/or performance level descriptors (i.e., “anchors”) could be revisited and further articulated to further differentiate the performance levels.

Next, we investigated the frequency distribution of ratings to identify whether there were items for which a majority of schools selected the “No Implementation” rating, as this might be an indication of a disconnect between expectations and the reality of what schools are able to accomplish in their K-3 programs. There were no instances of items for which a majority of schools selected “No Implementation.” In fact, the “No Implementation” rating was used very seldom. Its most frequent occurrence was for the item, “To what extent does the SBDM committee establish and enact a process to at least annually analyze data related to the implementation and impact of policies and practices specifically for the K-3 program?” Even with this being the item with the most frequent occurrence of the “No Implementation” rating, only 2% of the 734 schools selected this rating. Given that 2013-2014 was the first operational year for K-3 program reviews, it is somewhat surprising that not more schools selected “no implementation” for this item.

Nearly 75% of elementary schools in Kentucky had a K-3 program review score corresponding to a performance level classification of proficient or distinguished; however, the percentages of elementary schools scoring proficient/distinguished on the program review areas for Arts & Humanities, Practical Living/Career Studies, and Writing were lower (63%, 61%, and 66%, respectively). Moreover, the percentages of elementary schools scoring proficient/distinguished on K-PREP Reading, Mathematics, Science, Social Studies, and Writing were 55%, 49%, 71%, 58%, and 39%, respectively. Given the lower percentages of proficient/distinguished for other program review areas and for K-PREP, this suggests that schools might have inflated their K-3 program review ratings.

We computed the mean rating on each item across all schools. Nearly 90% of the items (25) had mean ratings of 2.0 or higher, meaning that the average school rated itself as at least "proficient" on nearly every K-3 item. Three items had mean ratings of less than 2.0. Those items were:

- "To what extent do teachers consistently involve students in defining and/or writing learning targets (using clear and precise language) that are essential to standard attainment? To what extent can students describe what it takes to achieve the target (the success criteria)? To what extent is instruction planned to directly ensure that students meet the targets and ultimately have opportunities to demonstrate understanding of the standard as a whole?" (*M* = 1.87, *SD* = 0.50)
- "To what extent does the school regularly communicate intervention services and progress with the families of those students identified for intervention? To what extent is family communication focused on improving student learning?" (*M* = 1.96, *SD* = 0.58)
- "To what extent do K-3 teachers collaborate with community, business and postsecondary partners through advisory committees, work exchange programs and/or community groups?" (*M* = 1.95, *SD* = 0.52).

Scale Level Analyses

Internal Consistency Reliability Estimates

As with the other program review areas, the K-3 program review consists of four "scales" or standards, which are: Curriculum & Instruction, Formative & Summative Assessment, Professional Learning, and Administrative/Leadership Support & Monitoring. Each scale is comprised of multiple subtopics, or "demonstrators," as they are referred to in the program review materials. As was done with the initial program review areas (see Sinclair & Thacker, 2013a; Sinclair & Thacker, 2013b), we investigated the internal consistency of each of the K-3 scales.

First, we calculated Cronbach's coefficient alphas as estimates of the internal consistency reliability of each of the scales. When individual items relate to the same concept (i.e., typically used as an indication of scale unidimensionality), then the scale will be more reliable. Generally, reliability estimates of .90 and higher are considered excellent, reliability estimates between .80 - .89 are considered good, and reliability estimates between .70 and .79 are considered adequate. The results in Table 3.1 show that the coefficient alphas for all scales are between .73 - .89, indicating adequate to good internal consistency reliability. It should be noted that it is possible to get a reliable scale using items with reasonably poor internal consistency if the scale contains enough items. For example, 10 items that have an average inter-item correlation of only 0.2 will still produce a scale with a reliability of 0.71. Similarly, if the average correlation among five items is 0.5, the alpha coefficient will be approximately 0.83, but if the number of items is 10—with the same average correlation—the alpha coefficient will be 0.91. Consequently, because the Curriculum & Instruction scale has the most items, it is not all that surprising that the alpha coefficient is highest for this scale. Another caveat to note is that the alpha coefficients are likely inflated to an unknown degree due to rater effects. Because a different rater (e.g., the school principal) rates each school, the variance attributed to raters cannot be estimated. As such, the coefficient alphas are inflated to an unknown degree.

Table 3.1. Internal Consistency Reliability Estimates for each K-3 Scale

Scale	# Items	*M*	*SD*	α
CI	13	2.26	0.36	.89
FSA	4	2.17	0.44	.78
PL	4	2.18	0.42	.73
ALSM	7	2.21	0.43	.85

Note. CI = Curriculum & Instruction; FSA = Formative & Summative Assessment; PL = Professional Learning; and ALSM = Administrative/Leadership Support & Monitoring.
M = Mean; *SD* = Standard deviation.

Single-Factor Confirmatory Factor Analyses

To further investigate the unidimensionality of the scales, we a conducted single-factor confirmatory factor analysis (CFA) on each scale using Mplus version 7 (Muthén & Muthén, 1998-2012). CFA allows the user to specify an *a priori* measurement model (by constraining parameters of the model), in which the relation between observed variables (i.e., characteristic ratings) and latent variables (i.e., constructs) is hypothesized. The covariance matrix implied by the hypothesized model is evaluated against the observed data matrix, thereby allowing quantification of model fit. Figure 3.1 represents the measurement model for the Formative & Summative Assessment scale. In this example, the four FSA items (denoted by the boxes) are explained by the latent variable FSA (denoted by the circle), as well as unique variance specific to each item.

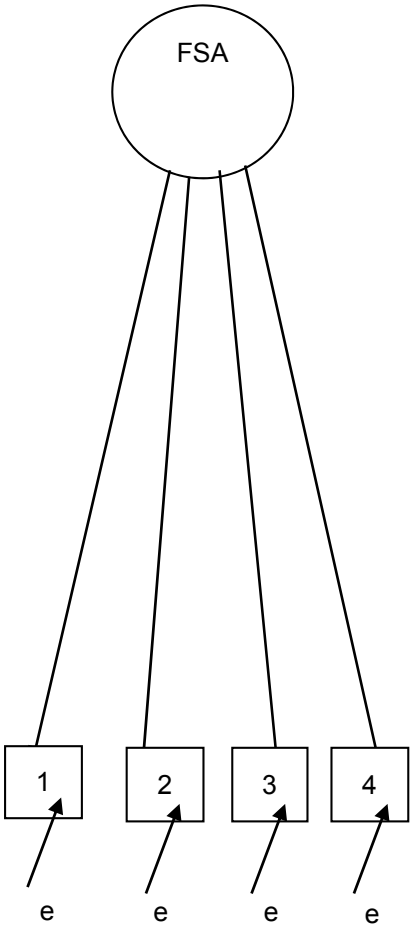


Figure 3.1. Example of Single-Factor Measurement Model.

Table 3.2 reports the degrees of freedom (df) and selected fit indices for the single-factor models for each of the four K-3 scales. In each of these single-factor models the scale (i.e., factor) is identified as the latent variable onto which all the items comprising that scale load. The chi-square values for all models are statistically significant at $p < 0.001$, indicating poor model fit, for all models *except* for Professional Learning. It should be noted that the chi-square test is *not* generally relied on as an index of overall model fit in models tested on samples larger than 200, as chi-square values are sample-size dependent. Root Mean Square Error of Approximation (RMSEA) values below .05 are generally indicative of good model fit, and values below 0.08 are generally indicative of reasonable model fit (lower is better) (Browne & Cudeck, 1989). The Professional Learning model demonstrated an RMSEA below 0.08, and the Administrative/Leadership Support & Monitoring model was very close (RMSEA = 0.081) to meeting that threshold. Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) values above 0.95 are generally indicative of good model fit, and values above .90 are generally indicative of reasonable model fit (higher is better) (Hu & Bentler, 1999). Of the four scales, only the Curriculum & Instruction scale did not meet either of these thresholds. Finally, Standardized Root Mean Square Residual (SRMR) values below 0.05 are generally indicative of good model

fit, and values below 0.08 are generally indicative of reasonable model fit (lower is better) (Browne & Cudeck, 1989); the models for Formative & Summative Assessment, Professional Learning, and Administrative/Leadership Support & Monitoring all exhibited SRMR values indicating good model fit, and the Curriculum & Instruction model exhibited an SRMR value indicating reasonable model fit. In summary, the results from the single-factor CFAs provide some evidence to support the unidimensionality of the scales, although evidence of unidimensionality was strongest for Professional Learning, which demonstrated acceptable model fit across all fit indices. Evidence of unidimensionality was weakest for Curriculum & Instruction, which only demonstrated acceptable model fit on the SRMR index. [8]

Table 3.2. Fit Indices for CFA Single-Factor Models for each Scale

Scale	Chi-Square	*df*	RMSEA	CFI	TLI	SRMR
CI	563.730*	65	0.102	0.857	0.829	*0.063*
FSA	45.876*	2	0.173	*0.949*	0.847	***0.041***
PL	2.096	2	***0.008***	***1.000***	***1.000***	***0.009***
ALSM	81.701*	14	0.081	***0.961***	*0.942*	***0.032***

Note. df = degrees of freedom; RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; TLI = Tucker Lewis Index; SRMR = Standardized Square Root Mean Residual.
*Indicates chi-square value is statistically significant ($p < .001$).
Bold italicized values indicate good model fit.
Italicized values indicate reasonable model fit.
CI = Curriculum & Instruction; FSA = Formative & Summative Assessment; PL = Professional Learning; and ALSM = Administrative/Leadership Support & Monitoring.

Given that the Curriculum & Instruction scale only demonstrated acceptable model fit on one of the fit indices, as an exploratory analysis, we conducted a principal components analysis using SPSS 22.0 (IBM, 2013) to help determine if the demonstrators that comprise the measure are distinct enough to emerge as separate components. Based on eigenvalues greater than 1.0, two components were extracted. All six items (i.e., "characteristics") comprising the "Kentucky Systems of Intervention/Response to Intervention" (KSI/RtI) demonstrator loaded onto one component, and the remaining seven items (which comprise the "Student Access," "Aligned & Rigorous Curriculum," and "Instructional Strategies" demonstrators) loaded onto a second component. This suggests that the "KSI/RtI" demonstrator might be distinct enough from Curriculum & Instruction to warrant its own scale.

Model Fit Comparisons

Next, three different CFA models were tested and compared to attempt to explain the underlying structure of the K-3 rubric. First, we investigated the intercorrelations among the four K-3 scales. Table 3.3 shows that the scales are highly correlated.

[8] It should be noted that the same caveat noted under the Scale Level Analyses section—that is, estimates being inflated to an unknown degree due to the fact that a different rater rated each school—also applies to the CFA results.

Table 3.3. Scale Intercorrelations for K-3

Scale	CI	FSA	PL	ALSM
CI	1.00			
FSA	.73	1.00		
PL	.67	.63	1.00	
ALSM	.65	.62	.67	1.00

Note. Correlations below the diagonal are on the observed variables, and correlations above the diagonal are on the latent variables.
CI = Curriculum & Instruction; FSA = Formative & Summative Assessment; PL = Professional Learning; and ALSM = Administrative/Leadership Support & Monitoring.

Given the intercorrelations among the scales, an intercorrelated, four-factor model was one of the models tested. We also tested a general factor model and second-order factor model. The three models and a description of each follows:

- Model 1: A single, general factor model with all 28 "characteristics" (i.e., items) loading onto a single K-3 factor.
- Model 2: An intercorrelated, four-factor model in which the four factors correspond to the four scales.
- Model 3: A model with a second-order factor responsible for the four, correlated factors. Given the reasonably high intercorrelations between the factors as seen in Table 3.3, this suggests that the correlations among the four factors might be accounted for by a higher-order factor. Consequently, we tested this third model in which the four-factor model has a second-order, general K-3 factor.

The fit indices displayed in Table 3.4 indicate that Model 1 had RMSEA and SRMR values indicating *reasonable* model fit, and that Model 2 and Model 3 had RMSEA values indicating *reasonable* model fit and SRMR values indicating *good* model fit. The Akaike Information Criterion (AIC) is a comparative measure of model fit (Akaike, 1974). Lower values indicate better fit. Model 2 has the lowest AIC value of all three models. Moreover, Model 2 also demonstrated the most favorable values for the CFI, TLI, and SRMR fit indices. Therefore, Model 2 represents the best fit of all three models. Figure 3.2 depicts the measurement model for Model 2. Given that Model 2 fit the data better than Model 3 this suggests that the second-order factor of "K-3" is likely too broad to be considered a construct. This makes intuitive sense given that the K-3 program entails a wide range of content ranging, for example, from mathematics to writing.

Table 3.4. Fit Indices for CFA Models 1 – 3

Scale	Chi-Square	*df*	RMSEA	CFI	TLI	SRMR	AIC
Model 1	1857.322*	350	*0.077*	0.824	0.810	*0.056*	27783.216
Model 2	1241.450*	344	*0.060*	0.895	0.885	***0.046***	27179.343
Model 3	1272.522*	346	*0.060*	0.892	0.882	***0.047***	27206.415

Note. df = degrees of freedom; RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; TLI = Tucker Lewis Index; SRMR = Standardized Square Root Mean Residual; AIC = Akaike Information Criterion.
*Indicates chi-square value is statistically significant ($p < .001$).
Bold italicized values indicate good model fit.
Italicized values indicate reasonable model fit.

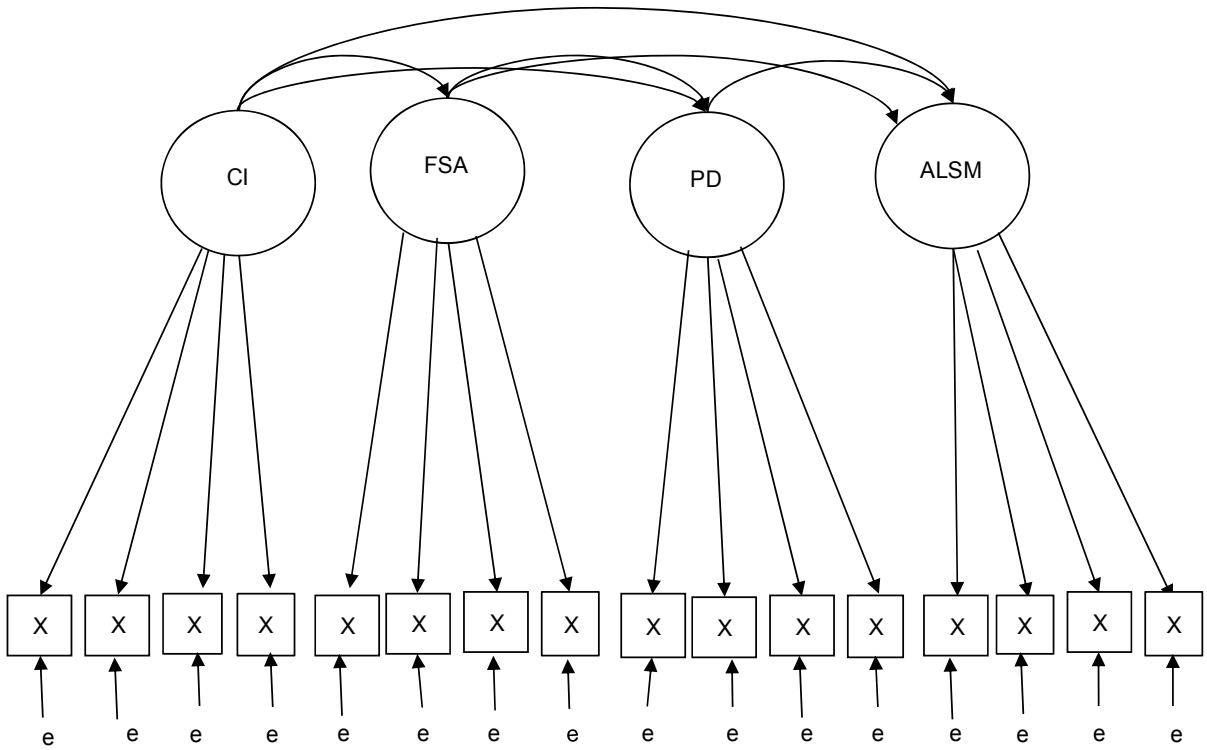


Figure 3.2. Simplified Measurement Model for Model 2.

This is a simplified measurement model in the sense that the number of observed variables (i.e., items) is more than four for CI and ALSM.

Summary of K-3 Analyses

In summary, the results from the K-3 analyses indicate that none of the items appear to suffer from a severe lack of variance in ratings, although the pattern of scores on the K-3 program

review, when compared with scores from other program review areas and with K-PREP scores, suggest that K-3 ratings might be somewhat inflated.

The scales that comprise the K-3 program review appear to be reasonably internally consistent. Findings from the single-factor CFAs provide additional evidence for the unidimensionality of the scales, with the possible exception of the Curriculum & Instruction scale, which only demonstrated acceptable model fit on one of the selected fit indices. An exploratory principal components analysis of the items on the Curriculum & Instruction scale provide evidence of two components such that all of the items comprising the "Kentucky Systems of Intervention/Response to Intervention" (KSI/RtI) demonstrator load on one component and all of the remaining items load on a second component. This suggests that the KSI/RtI demonstrator might be distinct enough from Curriculum & Instruction to warrant its own scale (i.e., standard), and that the high internal consistency reliability (i.e., coefficient alpha) reported for the scale might be an artifact of the relatively large number of items comprising the scale (13), and not the homogeneity of the items on the scale.

Finally, three CFA models were tested and compared to attempt to explain the underlying structure of the K-3 rubric. An intercorrelated, four-factor model fit the data better than a single, general factor model and better than a model with a second-order factor. This suggests that the second-order factor of "K-3" is likely too broad to be considered a construct, which makes intuitive sense given the wide range of content areas encompassed under the K-3 program.

Section 2: Convergent Validity Investigation between Program Review Writing and K-PREP Writing

For this section of the report, we investigated the relation between schools' program review Writing scores and their K-PREP Writing scores. Given that both measures purport to assess writing, we expect that schools' scores on the two measures should be positively and appropriately correlated with one another. This has been referred to as the "Goldilocks" range (Hoffman, 1998); that is, not so highly correlated as to indicate that the measures do not have important differences, but not so low as to indicate that they measure entirely different content. While we expect schools' program review Writing scores to be positively correlated with their K-PREP Writing scores we do not expect the correlations to be particularly strong. First, the two tests cannot be perfectly correlated because the relation is affected by error variance. Second, even though program review Writing and K-PREP Writing are both assessing writing, the two are not assessing identical content. Third, the formats of the measures are very different. Program review Writing scores are ratings made by school leaders based on collected evidence/artifacts of the writing program within the school, whereas K-PREP Writing scores are based on student responses to on-demand writing prompts, which are administered in grades 5, 6, 8, 10 and 11.

Because prior research has demonstrated reasonable support for the unidimensionality of the scales comprising the program review measure for Writing (Sinclair & Thacker, 2013b), correlations were computed between schools' K-PREP Writing scores and their scores on *each* of the program review Writing scales (i.e., "standards"). The four scales for the program review measure for writing and their supporting demonstrators are:

1. Curriculum & Instruction:
    - Student Access
    - Aligned and Rigorous Curriculum
    - Instructional Strategies
    - Student Performance
2. Formative & Summative Assessment
    - Variety of Assessment
    - Expectations for Student Learning
    - Response to Assessment
3. Professional Learning
    - Planning
    - Participation
    - Teacher Leadership
4. Administrative/Leadership Support & Monitoring:
    - Shared Vision
    - Time and Resources
    - Policies and Monitoring
    - Principal Leadership

Given that the "Curriculum & Instruction" and "Formative & Summative Assessment" scales are most directly relevant to student performance, we expect correlations to be higher between these scales and K-PREP Writing, and lower between the other scales and K-PREP Writing. If this pattern emerges, then this would provide some convergent and discriminant validity evidence for the Writing program review.

At the time of this report, two years' worth of operational data were available for program review Writing. Consequently, within-year correlations were computed between schools' scores on program review Writing and K-PREP Writing for 2012-2013 and again for 2013-2014. All correlations were computed by grade span.

The results displayed in Table 3.5 (see bold italicized values) indicate that the schools' program review Writing scores and their K-RPEP writing scores are indeed positively correlated. The magnitude of the correlations range from weak to moderate (r = .15 to r =.35). These correlations are weaker in magnitude than what has been found in prior convergent validity investigations involving other assessments in Kentucky; for example, correlations between ACT mathematics scores and mathematics scores on the Kentucky state assessment (which, at the time of the study, was the Kentucky Core Content Test) were correlated at r = .59 (Dickinson & Thacker, 2009). However, the finding that the correlations between K-RPEP Writing and program review Writing are less than correlations reported in prior convergent validity investigations in Kentucky is not surprising given that in the present study the two measures being correlated have very different formats—one is a student-level assessment in which students respond to writing prompts and the other is a program-level review by school leadership of the evidence and artifacts of the writing program in the school. The convergent validity correlation cited in the aforementioned research was based on two measures with much more similar formats (i.e., ACT Mathematics and KCCT Mathematics).

In general, the pattern of results presented in Table 3.5 provides some support for the posited convergent and discriminant relations. The magnitude of the correlations between program review Writing scores and K-PREP Writing scores tended to be slightly stronger when correlating K-RPEP writing with the Curriculum & Instruction and Formative & Summative Assessment scales (mean r = .29) than when correlating K-RPEP writing with the Professional Learning and Administrative/Leadership Support & Monitoring scales (mean r = .23).

Table 3.5. Correlations between Schools' K-PREP Writing Scores and Program Review Writing Scores

		2012 – 2013						2013 - 2014					
Grade Span [a](n)	Measure	KPREP WR	PR WR CI	PR WR FSA	PR WR PL	PR WR ALSM	PR WR Total	KPREP WR	PR WR CI	PR WR FSA	PR WR PL	PR WR ALSM	PR WR Total
Elem School (n=699) (n=689)	[b]KPREP WR	1.00						1.00					
	PR WR CI	***.33***	1.00					***.32***	1.00				
	PR WR FSA	***.35***	.83	1.00				***.32***	.79	1.00			
	PR WR PL	***.27***	.77	.75	1.00			***.26***	.66	.68	1.00		
	PR WR ALSM	***.28***	.72	.68	.79	1.00		***.25***	.61	.63	.72	1.00	
	[c]PR WR Total	***.34***	.91	.89	.92	.89	1.00	***.33***	.86	.88	.89	.87	1.00
Middle School (n=327) (n=329)	KPREP WR	1.00						1.00					
	PR WR CI	***.31***	1.00					***.28***	1.00				
	PR WR FSA	***.29***	.87	1.00				***.25***	.79	1.00			
	PR WR PL	***.20***	.76	.76	1.00			***.18***	.69	.70	1.00		
	PR WR ALSM	***.27***	.76	.74	.78	1.00		***.25***	.62	.63	.69	1.00	
	PR WR Total	***.29***	.92	.91	.91	.90	1.00	***.27***	.87	.89	.89	.85	1.00
High School (n=229) (n=228)	KPREP WR	1.00						1.00					
	PR WR CI	***.28***	1.00					***.28***	1.00				
	PR WR FSA	***.23***	.83	1.00				***.18***	.75	1.00			
	PR WR PL	***.15***	.71	.72	1.00			***.14***	.67	.65	1.00		
	PR WR ALSM	***.24***	.68	.67	.75	1.00		***.25***	.58	.52	.70	1.00	
	PR WR Total	***.25***	.89	.90	.90	.88	1.00	***.24***	.86	.84	.89	.83	1.00

Note. [a]First set of n counts are for 2012-2013 and second set are for 2013-2014.
[b]KPREP WR is the K-PREP Writing score representing the NAPD Calculation variable in the Accountability Achievement Level data file downloaded from http://applications.education.ky.gov/SRC/DataSets.aspx
[c]PR WR Total is the total (summative) points across the four program review scales.
WR = Writing; PR = Program Review; CI = Curriculum & Instruction; FSA = Formative & Summative Assessment; PL = Professional Learning; and ALSM = Administrative/Leadership Support & Monitoring.
Bold italicized values represent the correlation coefficients of greatest interest in the current investigation

Conclusions

This report includes two separate investigations of the Kentucky program reviews. The first was an investigation of the quantitative properties of data collected from the K-3 program review in its first operational year (i.e., 2013-2014), and the second was a convergent validity investigation between schools' program review Writing scores and their K-PREP Writing scores.

The findings from the first investigation indicate that the K-3 program produced data with reasonably sound quantitative properties. None of the items comprising the K-3 measure appear to suffer from a severe lack of variance in ratings. However, there is some evidence to suggest that K-3 ratings might be somewhat inflated. The scales that comprise the K-3 program review appear to be reasonably internally consistent. Findings from the single-factor CFAs provide additional evidence for the unidimensionality of the scales, with the possible exception of the Curriculum & Instruction scale. An exploratory principal components analysis revealed that the items (i.e., characteristics) comprising the KSI/RtI demonstrator might be distinct enough from the other items on the Curriculum & Instruction scale to warrant its own scale (i.e., standard). Finally, an intercorrelated, four-factor model demonstrated the best fit with the K-3 program review data.

The findings from the convergent validity investigation between schools' program review Writing scores and their K-PREP Writing scores indicate that the program review Writing scores are positively correlated with K-PREP Writing scores. While the magnitude of the correlations were weak to moderate, which is less than prior convergent validity studies involving assessments in Kentucky (e.g., Dickinson & Thacker, 2009; Sinclair, et al., 2008), this is likely due to the very different formats of the two measures. The pattern of the correlations was consistent with the posited convergent and discrimant relations such that the magnitude of the correlations between program review Writing and K-PREP Writing tended to be stronger when correlating K-RPEP writing with the Curriculum & Instruction and Formative & Summative Assessment scales than when correlating K-RPEP writing with the Professional Learning and Administrative/Leadership Support & Monitoring scales. This provides some evidence of convergent and discriminant validity for the program review scores for Writing.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716 – 723.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24,* 445 – 455.

Dickinson, E.R., & Thacker, A.A. (2009). Relations among Kentucky's Core Content Test, ACT scores, and students' self-reported high school grades. (FR-09-32). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G., (1998). Relationships among Kentucky's open-response assessments, ACT scores, and students' self-reported high school grades. (FR-98-27). Alexandria, VA: Human Resources Research Organization.

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1-55.

IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

Muthén, L. K., & Muthén, B. O. (1998-2012). Mplus Users Guide (7th ed.). Los Angeles, CA: Muthén & Muthén.

Sinclair, A.L. & Thacker, A.A. (2013a). Description of the development of scores for the 2011-2012 program reviews and an investigation of their quantitative properties. (FR-13-04). Alexandria, VA: Human Resources Research Organization.

Sinclair, A.L. & Thacker, A.A. (2013b). Analysis of 2012-2013 program review data: Follow-up investigation of the 2011-2012 pilot data. (FR-13-069). Alexandria, VA: Human Resources Research Organization.

Sinclair, A.L., Thacker, A.A., Koger, L.E., & Dickinson, E.R. (2008). Relations between students' scores on the revised 2007 KCCT and the prior KCCT. (FR-08-11). Alexandria, VA: Human Resources Research Organization.