

# Program Reviews: A Quantitative Analysis of K-3 Program Review Data and a Convergent Validity Investigation of Program Review Writing and K-PREP Writing

## Final Report

**Prepared for:** Kentucky Department of Education  
Capital Plaza Tower, 17<sup>th</sup> Floor  
500 Mero Street  
Frankfort, KY 40601

**Authors:** Andrea L. Sinclair

**Date:** January 21, 2015

## Program Reviews: A Quantitative Analysis of K-3 Program Review Data and a Convergent Validity Investigation of Program Review Writing and K-PREP Writing

### Executive Summary

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 Kindergarten through Grade 3 (K-3) program review in its first operational year (i.e., 2013-2014). The findings 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 inflated. The scales (i.e., “standards”) that comprise the K-3 program review appear to be reasonably internally consistent, and findings from the single-factor confirmatory factor analyses (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 “Kentucky Systems of Intervention/Response to Intervention” (KSI/RtI) demonstrator might be distinct enough from the other items on the Curriculum & Instruction scale to warrant its own scale. Finally, an intercorrelated, four-factor model demonstrated the best fit with the K-3 program review data.

The second investigation in this report was a convergent validity investigation between schools’ program review Writing scores and their on-demand writing scores on the Kentucky Performance Rating for Educational Progress (K-PREP). Findings 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, Thacker, Koger, & Dickinson, 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 discriminant relations such that the magnitude of the correlations between program review Writing and K-PREP Writing tended to be stronger when correlating K-PREP writing with the Curriculum & Instruction and Formative & Summative Assessment scales than when correlating K-PREP writing with the Professional Learning and Administrative/Leadership Support & Monitoring scales. This provides some evidence of convergent and discriminant validity for the program review for Writing.

# Program Reviews: A Quantitative Analysis of K-3 Program Review Data and a Convergent Validity Investigation of Program Review Writing and K-PREP Writing

## Table of Contents

Executive Summary .....	i
Introduction and Background .....	1
Section 1: Quantitative Investigation of K-3 Program Review Scores .....	1
Item Level Investigations .....	1
Scale Level Analyses .....	3
Internal Consistency Reliability Estimates .....	3
Single-Factor Confirmatory Factor Analyses .....	4
Model Fit Comparisons.....	6
Summary of K-3 Analyses .....	8
Section 2: Convergent Validity Investigation between Program Review Writing and K-PREP Writing .....	9
Conclusions .....	12

## List of Tables

Table 1. Internal Consistency Reliability Estimates for each K-3 Scale .....	4
Table 2. Fit Indices for CFA Single-Factor Models for each Scale.....	6
Table 3. Scale Intercorrelations for K-3.....	7
Table 4. Fit Indices for CFA Models 1 – 3 .....	8
Table 5. Correlations between Schools' K-PREP Writing Scores and Program Review Writing Scores .....	11

## List of Figures

Figure 1. Example of Single-Factor Measurement Model.....	5
Figure 2. Simplified Measurement Model for Model 2.....	8

# Program Reviews: A Quantitative Analysis of K-3 Program Review Data and a Convergent Validity Investigation of Program Review Writing and K-PREP Writing

## Introduction and Background

This report represents the third in a series of HumRRO reports on program reviews in Kentucky public schools. The first 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). The first report 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. The 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 third report includes two parts. 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 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.

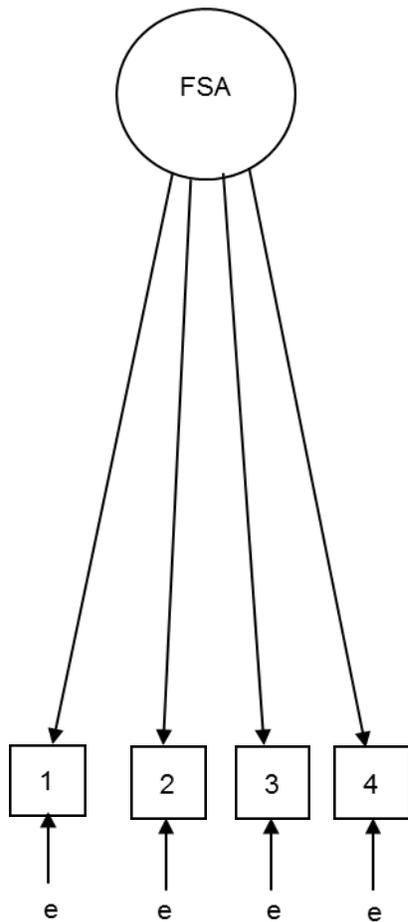
**Table 1. Internal Consistency Reliability Estimates for each K-3 Scale**

Scale	# Items	<i>M</i>	<i>SD</i>	$\alpha$
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 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 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 1. Example of Single-Factor Measurement Model.**

Table 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. <sup>1</sup>

**Table 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	<b>0.041</b>
PL	2.096	2	<b>0.008</b>	<b>1.000</b>	<b>1.000</b>	<b>0.009</b>
ALSM	81.701*	14	0.081	<b>0.961</b>	0.942	<b>0.032</b>

*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 shows that the scales are highly correlated.

<sup>1</sup> 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. 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, 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 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 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 4. Fit Indices for CFA Models 1 – 3**

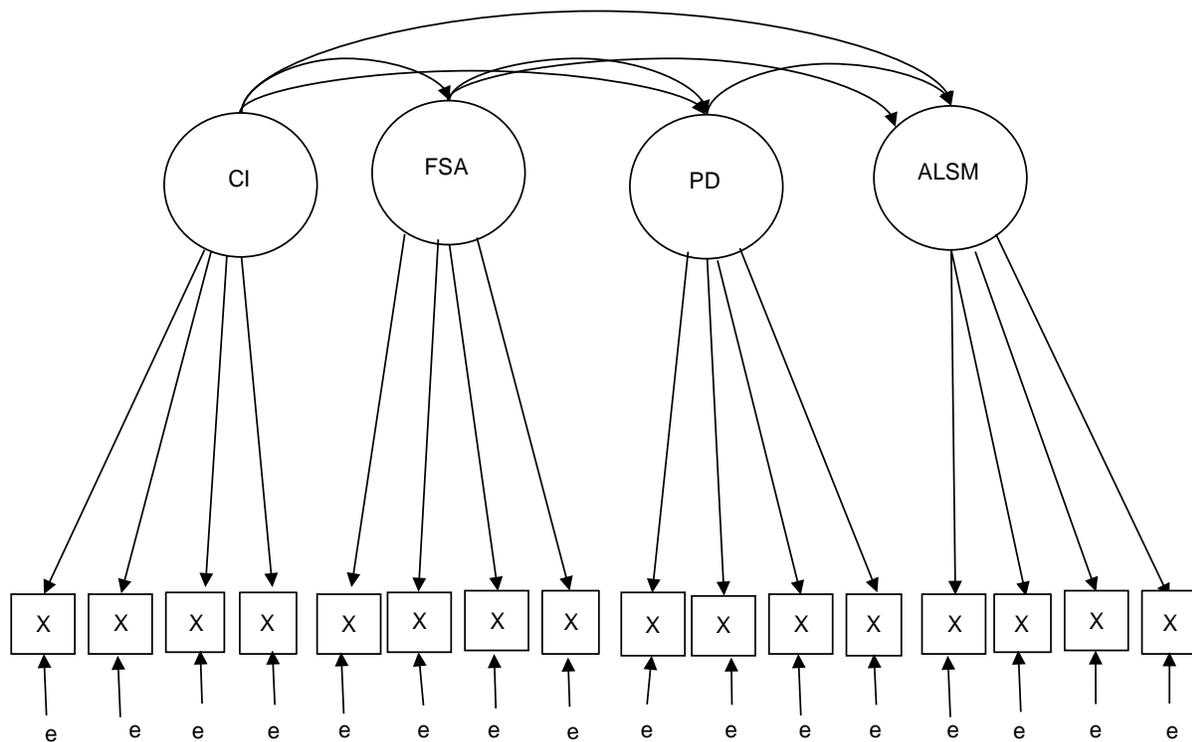
Scale	Chi-Square	df	RMSEA	CFI	TLI	SRMR	AIC
Model 1	1857.322*	350	0.077	0.824	0.810	0.056	27783.216
Model 2	1241.450*	344	0.060	0.895	0.885	<b>0.046</b>	27179.343
Model 3	1272.522*	346	0.060	0.892	0.882	<b>0.047</b>	27206.415

*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 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 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 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 5. Correlations between Schools' K-PREP Writing Scores and Program Review Writing Scores**

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

Note. <sup>a</sup>First set of n counts are for 2012-2013 and second set are for 2013-2014.

<sup>b</sup>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>

<sup>c</sup>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 discriminant 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.