2013 No. 038

Exploration of the Scoring of the K-PREP On-Demand Writing Assessment

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

Authors: Arthur A. Thacker
Emily R. Dickinson
Heather Levinson

Date: June, 2013

HumRRO®
Human Resources Research Organization

Human Resources Research Organization (HumRRO)
66 Canal Center Plaza, Suite 700, Alexandria, Virginia 22314-1578
Phone: 703.549.3611 | Fax: 703.549.9661 | www.humrro.org

HumRRO
Human Resources Research Organization

2013 No. 038

Exploration of the Scoring of the K-PREP On-Demand Writing Assessment

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

Authors: Arthur A. Thacker
Emily R. Dickinson
Heather Levinson

Date: June, 2013

Exploration of the Scoring of the K-PREP On-Demand Writing Assessment

Table of Contents

Overview of K-PREP Writing Assessment 1
Purpose of Study 2
Analyses and Results 2
Discussion and Conclusions 4
References 6

List of Tables

Table 1 K-PREP On-Demand Writing Test Blueprint 1
Table 2. Correlations between Unadjusted and Adjusted Writing Scores and Reading and Math Scale Scores, by Grade Level 3
Table 3. Percentage of Writing Proficiency Level Classifications Based on Original Writing Scale Score and on Single Rater Scores, by Rater and Grade Level 1
Table 4. Rater Agreement for the Grade 5 ODW Assessment 1
Table 5. Rater Agreement for the Grade 6 ODW Assessment 1
Table 6. Rater Agreement for the Grade 8 ODW Assessment 1
Table 7. Rater Agreement for the Grade 10 ODW Assessment 2
Table 8. Rater Agreement for the Grade 11 ODW Assessment 2
Table 9. Percentage of Calculated Differences between Original NAPD Categorizations and Categorizations Based on Single Rater, by Rater and Grade Level 3
Table 10. Correlations between Original Adjusted Writing Scores and Single-Rater Writing Scores and Reading and Math Scale Scores 3

EXPLORATION OF THE SCORING OF THE K-PREP ON-DEMAND WRITING ASSESSMENT

Overview of K-PREP Writing Assessment

The Kentucky next-generation learners accountability model is anchored in college and career readiness for all students. Thereading, writing, and mathematics standards were adopted in 2009 through Senate Bill 1 from the Common Core Standards (referred to as the Kentucky Common Academic Standards or KCAS). These standards were developed through the National Governors Association Center for Best Practices and the Council of Chief State School Officers. Prior to the implementation of KCAS, the Commonwealth Accountability Testing System (CATS) included the Kentucky Core Content Test (KCCT) with a section of multiple-choice items and an on-demand writing component in grades 5 and 8, and Writing Portfolios produced annually and scored in 4th, 7th, and 12th grades for accountability purposes.

Beginning in the 2011-12 school year, the Kentucky Performance Rating for Educational Progress (K-PREP) replaced KCCT. Writing is now assessed in grades 5, 6, 8, 10 and 11, as shown in Table 1. For writing, K-PREP includes two types of on-demand writing tasks, stand-alone prompts that outline a situation, and passage-based prompts. These prompts assess writing specific to genre (e.g. narrative, informative/exploratory, and argumentative) and goals for writing conventions (e.g., organization and style). Prompts and scoring guides were developed through collaboration between the testing contractor and KDE.

Table 1 K-PREP On-Demand Writing Test Blueprint

Grade	Stand-Alone A	Stand-Alone B	Passage-Based
5	Narrative	Opinion	Informative/Explanatory
6	Narrative	Argumentative	Informative/Explanatory
8	Narrative	Informative/Explanatory	Argumentative
10	Informative/Explanatory	Informative/Explanatory	Argumentative
11	Argumentative	Argumentative	Informative/Explanatory

The changes modeled in the writing portion of K-PREP are a response to significant shifts in writing instruction and assessment required following the adoption of KCAS. Writing instruction includes writing to prompts and authentic texts, including content from English language arts, history, and science. An important change in KCAS is the integration of writing with reading. Students apply one or more writing standards to respond to a literary or informational text, using evidence from the text to support analyses and reflection. Students may present analysis as argument or explanation.

Across grades, writing mirrors the shift in reading from narrative to informational texts, reflecting vertical growth. All grades give evidence from sources of sufficient text complexity, focus on coherent writing, and integrated reading standards. Only grades 3-5, however, engage in narrative writing. The elementary grades also write argument and explanatory writing, but for middle and high school, these types of writing dominate instruction and assessment. Each type of writing is based on purpose. Narratives convey experiences, real or imaginary, to inform,

instruct, persuade, or entertain. Explanatory writing clarifies or increases the knowledge and understanding of readers. Argument persuades the reader or asks the reader to adopt a certain point of view using the structure of claim, reasoning, and evidence.

Principals of assessment in writing also have shifted in response to KCAS. Consistent with the Next-Generation Learners accountability model, assessment is intended to measure achievement across a wide spectrum to determine if students are on track to be college and career ready. Assessments should also model work worth doing in classroom in order to promote high quality student work and student learning, with better interaction with formative assessments. Last, the assessment model should be sustainable and scalable, demonstrating careful plans to validate assessment scores and claims made based on scores. The present study addresses the validity of K-PREP writing scores by exploring potential threats to validity during scoring.

Purpose of Study

The purpose of the present study is to explore potential issues related to the scoring of the K-PREP writing assessments. K-PREP writing test items are independently scored by two trained raters. Before a rater is considered qualified for the "live" scoring of items, s/he must meet certain criteria. For example, each rater must demonstrate an adequate level of agreement with other raters on what the appropriate score for a particular item should be.

During scoring of the 2012 K-PREP writing assessment, it was noted that very few students received scores of '4'. The writing test is comprised of two prompts, each scored by two reviewers, for a total of eight points per prompt (16 total points possible after summing across prompts). So, when we discuss the writing scale, it is important to remember which scale we are discussing. The full test scale is '16' points. The scale for a given prompt is '8' points, but that scale is created by adding the ratings given by two raters, each of whom can only assign a rating of '0-4.' At the rater-level, there are only four available points for the scale. If the raters do not use the '4' rating, the scale is reduced to 3 points (0 is reserved for blank or entirely off topic responses). This creates a potential issue when we observe that ratings of '2' are extremely common. Given that inter-rater agreement is judged acceptable if raters assign adjacent scores, then always applying a rating of '2' would be adequate for qualification. This is not to suggest that raters are engaging in such activity, but does point to potential threats to the validity of writing scores when the mechanism for gauging rating accuracy does not function as intended.

Analyses and Results

This study addresses one aspect of the validity of writing scores as indicators of student-level writing ability. Writing scores should correlate highly with reading scores given that the two measurement constructs (writing and reading) are closely related. Similarly, because the constructs of writing and mathematics would be expected to be less related than writing and reading, we would expect the writing-to-reading correlation to be higher than the writing-to-math correlation. These correlations are referred to as validity coefficients. Convergent validity coefficients refer to the correlation between like or similar constructs, in this case writing-to-reading. Discriminant validity coefficients refer to the correlation between dissimilar constructs, in this case writing-to-math. The strength of these correlations can vary considerably due to the

reliability of the assessments or other factors, but the pattern of higher convergent than discriminant validity coefficients should hold if the tests are genuinely measuring the constructs in the way they were designed.

Table 2 presents correlations between writing scale scores and reading and math scale scores. Correlations are presented for both unadjusted scores and scores that were adjusted for rater severity/leniency (adjustment computed by Dr. William Auty, data provided to HumRRO and matched to KDE-provided files). N-counts represent the number of students with scores in all three subject areas at each grade level. As high schools now use end-of-course (EOC) assessments, some students may test in reading and math in the 9th or 12th grade, at which time writing is not assessed. Thus, correlations reflect those students who were tested in reading and/or math at the same grade level as writing.

Table 2. Correlations between Unadjusted and Adjusted Writing Scores and Reading and Math Scale Scores, by Grade Level

		Grade				
		5 (n=50,650)	6 (n=50,291)	8 (n=49,080)	10 (n=44,413)	11 (n=22,265)
Unadjusted	Reading	.581	.642	.642	.698	.508
	Math	.541	.604	.591	.488	.404
Adjusted	Reading	.542	.616	.604	.669	.489**
	Math	.504	.577	.554	.462*	.384

Note: All correlations statistically significant at p <.001.
*Correlation based on 11,249students who had math scale score data in 10th grade.
**Correlation based on 626 students who had reading scale score data in 11th grade.

Table 2, as expected, shows that reading and writing were highly correlated. The somewhat lower correlation in high school is not overly worrisome due to the much smaller sample size. Math and writing are also highly correlated, but not so highly as reading and writing for any grade level. This pattern represents validity evidence that the construct being measured by the writing test is more similar to the construct measured by the reading test than the math test. It should be noted that all assessment scores can be expected to be positively correlated at the student level. This is due to the very strong tendency of higher-ability students to be strong across all subjects and weaker students to be weaker across all subjects. Unadjusted scores were consistently more strongly correlated than adjusted writing scores, though the magnitude of the difference was quite small.

Next, students were rescored using only scores from a single rater, which were doubled to produce a score on the 16-point scale. This process was done for both raters, and then performance levels were assigned based on each new score. Raters were randomly assigned based on whether they were listed as ‘Rater 1” or ‘2’ in the data file. Table 3 presents the percentage of students in each NAPD category based on the original writing score, along with the percentages based on the single-rater scores. Across the grade levels, using scores based on a single rater tended to decrease the proportion of students scoring Proficient, while increasing the percentages of students in all other categories. The one exception to this pattern was 6th grade in which both Proficient and Distinguished percentages were reduced when using scores based on a

single rater. This pattern is likely due to the position of cut scores and the level of exact rater agreement.

Table 3. Percentage of Writing Proficiency Level Classifications Based on Original Writing Scale Score and on Single Rater Scores, by Rater and Grade Level

	Grade 5 (n=51,482)			Grade 6 (n=51,120)			Grade 8 (n=50,046)			Grade 10 (n=48,272)			Grade 11 (n=44,481)		
	O	R1	R2	O	R1	R2	O	R1	R2	O	R1	R2	O	R1	R2
N	19.05	25.41	25.73	17.86	36.59	36.64	11.19	15.09	15.14	12.17	16.02	16.20	19.12	24.50	24.69
A	49.40	58.70	58.67	43.06	50.04	50.13	45.48	47.31	47.50	45.77	47.75	47.85	35.49	48.01	48.29
P	29.61	12.00	11.89	35.23	10.27	10.17	34.17	24.04	23.92	35.50	25.91	25.74	39.66	18.64	18.33
D	1.95	3.89	3.71	3.85	3.09	3.06	9.16	13.56	13.44	6.56	10.31	10.22	5.73	8.84	8.68

Note: O= Original score; R1= Score based on Rater 1 only; R2= Score based on Rater 2 only.

Table 3 shows that most Kentucky students scored in the Apprentice category. In grades 5 and 6 more than 50% of students were classified as Apprentice, while between 45-48% of students were classified as Apprentice in the remaining grades. Ratings of '2' by all raters (overall scores of 8) would result in a classification of Apprentice for all grade levels. This table demonstrates why there was concern that the scoring might not be as representative of student performance as intended.

Tables 4-8 present the classifications students would have received if their overall scores were based on Rater 1 versus Rater 2. The diagonal from each table represents the classification agreement percentage between the two raters. The percentages off the diagonal represent students who would have received a different classification if their scores were based on a different rater. These tables represent only the Rater 1 versus Rater 2 classifications, rather than agreement with the actual classification.

Table 4. Rater Agreement for the Grade 5 ODW Assessment

		Rater 2				
		N	A	P	D	
Rater 1	N	18.57	6.76	.08	.01	
	A	7.07	45.42	5.55	.66	
	P	.08	5.70	4.77	1.45	
	D	.01	.79	1.49	1.59	
				Level of agreement		70.35%

Table 5. Rater Agreement for the Grade 6 ODW Assessment

		Rater 2				
		N	A	P	D	
Rater 1	N	28.55	7.95	.09	.01	
	A	.00	36.78	4.69	.58	
	P	.09	4.78	4.18	1.23	
	D	.01	.62	1.21	1.25	
				Level of agreement		70.76%

Table 6. Rater Agreement for the Grade 8 ODW Assessment

		Rater 2				
		N	A	P	D	
Rater 1	N	11.22	3.81	.06	.01	
	A	3.82	34.91	7.61	.97	
	P	.09	7.68	11.63	4.63	
	D	.01	1.10	4.62	7.83	
				Level of agreement		65.59%

Table 7. Rater Agreement for the Grade 10 ODW Assessment

		Rater 2				
		N	A	P	D	
Rater 1	N	12.39	3.59	.04	0	
	A	3.78	35.56	7.72	.70	
	P	.03	8.01	13.93	3.94	
	D	0	.69	4.05	5.58	
				Level of agreement		67.46%

Table 8. Rater Agreement for the Grade 11 ODW Assessment

		Rater 2				
		N	A	P	D	
Rater 1	N	18.96	5.48	.06	0	
	A	5.65	35.05	6.45	0.86	
	P	.08	6.81	8.69	3.06	
	D	0	.96	3.12	4.76	
				Level of agreement		67.46%

Levels of agreement by classification ranged from about 65% to 70% among raters. This is encouraging since the percentage of students from the Apprentice exact agreement cell represents only about 35-45% of the total exact agreement. This is an indication that when students score lower or higher than 2, raters often agree on those rating.

Next, the distance between the original performance categorizations and those based on single-rater scores was calculated for each student. Distance is defined as the number of categories away from the operational classification the student received compared to the classification based on a single rater. For example, if a student received a classification of Novice, but would have received a classification of Proficient if her rating were based only on Rater 2, the distance would be 2. Since the difference is in the positive direction (Proficient is higher than Novice), the difference would be classified as positive. Distances ranged from -3 to 2, with the largest percentage of performance categorizations being equivalent whether based on both raters or based on a single rater. Across the grade levels, less than half of one percent of ratings differed by a distance of 2 or more categories. Table 9 presents these results.

Table 9. Percentage of Calculated Differences between Original NAPD Categorizations and Categorizations Based on Single Rater, by Rater and Grade Level

	Grade 5		Grade 6		Grade 8		Grade 10		Grade 11	
Distance	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
-3	.04	.04	.02	.02	.04	.05	.04	.04	.03	.03
-2	.28	.29	.37	.36	.27	.29	.30	.29	.38	.41
-1	22.45	23.00	45.10	45.29	11.61	11.84	11.43	11.59	23.57	24.23
0	74.24	73.88	53.79	53.68	81.16	81.06	82.11	82.34	71.79	71.21
1	2.92	2.72	.71	.65	6.78	6.61	6.08	5.72	4.20	4.10
2	.06	.06	0	0	.14	.16	.05	.03	.03	.02

Note: R1= Score based on Rater 1 only; R2= Score based on Rater 2 only.

Table 9 demonstrates that if classification of students were based on a single rater, most of the time that classification would match the classification based on two separate raters. It also shows that most unmatched classifications were in adjacent categories. It was extremely rare for the raters to disagree in such a way that it would alter classification by 2 or more NAPD categories.

Finally, we calculated convergent validity correlations based on the single-rater scores. If the scores derived from a single rater had substantially lower correlations than correlations from combined raters, then this would provide evidence that combining multiple raters resulted in more valid writing scores. Table 10 presents the newly calculated correlations, along with the original correlations presented in Table 2 (adjusted for rater severity and unadjusted). The four sets of correlations are all very similar, suggesting that the system of adding raters' scores does not necessarily produce more valid writing scores.

Table 10. Correlations between Original Writing Scores and Single-Rater Writing Scores with Reading and Math Scale Scores

	Grade				
	5	6	8	10	11
	Unadjusted Scale Score (from Table 2)				
Reading	.581	.642	.642	.698	.508
Math	.541	.604	.591	.488	.404
	Adjusted Scale Score (from Table 2)				
Reading	.542	.616	.604	.669	.489
Math	.504	.577	.554	.462*	.384
	Rater 1 Only				
Reading	.534	.598	.601	.667	.503
Math	.501	.561	.552	.449	.376
	Rater 2 Only				
Reading	.535	.601	.602	.665	.515
Math	.497	.564	.551	.455	.377

Discussion and Conclusions

While convergent validity coefficients provide evidence of the validity of K-PREP on-demand writing scores, there remain a number of concerns about the current scoring system that need to be addressed. The levels of agreement are similar to other reports of rater consensus in the application of scoring rubrics;however, it has been argued that 70% agreement is too low in the context of large-scale assessments (Jonsson & Svingby, 2007). Also of concern is the compression of the scale that results from raters rarely assigning a score of '4.' Finally, the use of two raters results in higher convergent validity coefficients than a single rater, but only by between 0 and 0.04. This increase represents, at most, accounting for an additional 0.16% of the variance in the writing score.

It is also surprising that correcting the writing scores for rater severity did not result in better correlations than the uncorrected scores, and in fact, the corrected correlations were somewhat weaker across grades. The corrections made for rater severity, however, happened even though raters may have been retrained to account for severity during the scoring process. It is unknown whether the adjustment would result in better validity coefficients if the training process did not interact with the correction. Based on the 2012 convergent validity evidence, adjusting scores for severity using the method applied was not supported.

The following changes are recommended for improving the current K-PREP on-demand writing scoring system based on these investigations:

- Improve the overall scoring system by requiring higher rates of inter-rater agreement and ensuring the use of all possible score points. Improvements might include reevaluating the scoring rubric to identify any modifications that would clarify the appropriate scoring decisions. Training staff should also verify that there are high-quality training papers representing each possible score point and that qualifying sets adequately cover the full scale as well. Additional quality checks could also be built into the monitoring of live scoring, such as identifying raters who tend to assign a particular score much more frequently than others.
- Change the approach to the calculation of scores so that it does not rely on adding separate ratings to get an overall score. Under the current system, scores assigned by separate raters are added together to create a 16-point scale. A wider scale can be useful from a statistical perspective by addressing problems associated with range restriction. However, reporting on the 16-point scale under the current system is misleading; a given student can truly only earn up to 8 possible score points on the two items to which s/he responds. Basing scores on a single rater will also produce scores with roughly the same convergent validity evidence.
- Conduct similar analyses to those in this report comparing 2012 writing results with 2013 writing results, where students are administered writing assessments in adjacent years (e.g. between Grades 5 and 6 and Grades 10 and 11). The test-retest reliability correlations will help verify and extend the results presented here and will allow for better informed decisions from KDE.
- Re-examine the scoring rubric for scores level '2.' Because so many students score at this level, it may be possible to generate another reliable score point and extend the scale if

some differentiation can be made among the ‘2’ level writing samples. This process may entail having writing content experts reevaluate the rubric and some of the writing samples, but if the scale is not extended, we could see anomalous shifts in the percentage of students scoring in any category based on minor improvements in writing, or we could also mask improvements in writing performance because of the limits of the scale.

References

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity, and educational consequences. *Educational Research Review, 2*, 130-144.

National Governors Association Center for Best Practices, Council of Chief State School Officers (2010). *Common Core State Standards English Language Arts.* National Governors Association Center for Best Practices, Council of Chief State School Officers, Washington D.C.