2023 No. 054

Validity Evidence for Combining Editing and Mechanics and On-Demand Writing

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

Prepared under: Contract #1900004339

Authors: Hye-Jeong Choi
Meng Fan
Emily R. Dickinson

Date: May 15, 2023

Validity Evidence for Combining Editing and Mechanics and On-Demand Writing

Table of Contents

Table of Contents (Continued)

List of Tables

List of Figures

Validity Evidence for Combining Editing and Mechanics and On-Demand Writing

Introduction

In spring 2022, the Kentucky Summative Assessment (KSA) for writing included both on-demand writing (ODW) and editing and mechanics (EM). This represents a change to Kentucky's writing assessment, which in recent years has included only on-demand writing prompts.

The Kentucky Department of Education (KDE) has developed separate Performance Level Descriptors (PLDs) for EM and ODW. Students are categorized into one of four performance levels (i.e., Novice, Apprentice, Proficient, and Distinguished [NAPD] categories) separately for EM and ODW based on their performance on those respective sets of test items. A student's combined writing performance level is then determined from the combination of their assigned EM and ODW performance levels.

Because these changes represent a major adjustment to how writing is assessed in Kentucky, it is important to evaluate their impact on the quality of student scores. The purpose of this study is to evaluate the reliability and construct validity of student scores derived from the new approach to measuring and reporting summative writing scores. This report summarizes the methods and results of the study.

Methods

Description of Data

KDE provided the Human Resources Research Organization (HumRRO) with data from the spring 2022 Kentucky Summative Assessments (KSA). Writing, assessed through two separate measures, is administered in Grades 5, 8, and 11. Student responses for ODW using four prompts. The first component of the writing assessment is used to assess students' writing skill on 5-6 traits: clarity and coherence, support, sourcing, organization, language/conventions, and counterclaims (Grades 8 and 11 only). At least two raters rated students' writing on a scale of 1-4 for each trait and 0 for un-scorable (i.e., non-response, insufficient amount to score, non-English, illegible, and off topic). When item-level scores were used for the analysis, we used the rater score for each trait. The second component of the writing assessment, EM, consists of 26 items measuring three constructs (i.e., Conventions of Standard English, Knowledge of Language, and Vocabulary Acquisition and Use). EM items include selected response and short-answer items. EM assessments included three forms for each grade, each with a total raw score of 36 points. Since items on separate EM forms differed, we treated different EM forms separately when item-level scores were used for analysis. Table 1 presents subscores or subdomains of EM and ODW for each grade.

Table 1. Subscore for Editing and Mechanics and On-Demand Writing Grade Level

Subject	Grade	Subscore/Trait	Abbreviation*
Editing and Mechanics	5, 8 & 11	Conventions of Standard English	EM1
		Knowledge of Language	EM2
		Vocabulary Acquisition and Use	EM3
On-Demand Writing	5	Clarity/Coherence	WR1
		Support	WR2
		Sourcing	WR3
		Organization	WR4
		Language/Conventions	WR5
	8 & 11	Clarity/Coherence	WR1
		Counterclaims	WR2
		Support	WR3
		Sourcing	WR4
		Organization	WR5
		Language/Conventions	WR6

* These abbreviations are used for tables in this report.

Before proceeding with the analysis, we cleaned the data. For this study, we included students who (1) took an online form, (2) had a valid ODW score from both rater 1 and rater 2, (3) had a valid student ID, and (4) took a non-accommodated form (that is, students who took Audio, Braille, or Large Print form were excluded). This approach reduced external factors unrelated to item content (e.g., test mode effects) and thus focuses on the impact of including EM items on the writing assessment. Table 2 presents the frequency of examinees for ODW prompts and EM forms by grade level in the final data.

Table 2. Frequency of examinees for EM Forms and ODW Prompts by Grade Level

Grade	ODW Prompt 1	ODW Prompt 2	ODW Prompt 3	ODW Prompt 4	EM Form 1	EM Form 2	EM Form 3	Total
5	11,404	10,247	10,244	10,539	14,697	13,789	13,948	42,434
8	11,553	10,711	10,657	10,876	15,181	14,132	14,484	43,797
11	9,622	9,416	9,280	9,313	12,989	12,404	12,238	37,631

Validity Evidence That Test Sections Represent One Writing Construct

A key component of a test validity argument is evidence supporting claims that the assessment items represent the intended measurement construct (Kane, 2006). Convergent validity can be measured by correlations between two tests of the same or similar constructs. Convergent validity is often used as validity evidence that a test measures its intended construct. Carlson and Herdman (2012) established thresholds for convergent validity, indicating that correlations above 0.50 were acceptable, and above 0.70 were recommended as indicators that two tests measured similar constructs. High correlations between the ODW and EM scores would indicate that the two writing subtests measured a similar construct and could be reasonably combined into a single score or reported performance category.

Because the ODW and EM scores are combined after score categories (NAPD) are determined separately for each subtest, polychoric correlations were also computed. Polychoric correlations are used to evaluate the relationship between the ODW performance level and EM performance level. Polychoric correlation is widely used to quantify the relationship between ordered categorical variables and it can be interpreted in the same way as a Pearson correlation.

Internal Consistency Reliability

Cronbach's α (Cronbach, 1951) was used to evaluate the reliability of the writing assessment. Cronbach's α is an internal consistency measure and has been widely used as a reliability measure. Internal consistency refers to the extent to which items on a test form are interrelated. Cronbach's α has a range of 0 to 1. The closer to 1, the more reliable the assessment. Established guidelines for interpreting Cronbach's α include Excellent ($\alpha>0.9$), good ($0.7<\alpha<0.9$), acceptable ($0.6<\alpha<0.7$), poor ($0.5<\alpha<0.6$), and unacceptable ($\alpha<0.5$) (George & Mallery, 2003, p. 231; Kline, 2000, p.13). To investigate the effect of adding EM to ODW in terms of internal consistency, we compared Cronbach's α for the ODW component only with Cronbach's α after combining the EM and ODW components.

Dimensionality Analysis

A confirmatory factor analysis (CFA) was used to evaluate the dimensionality of the writing assessment. In so doing, we compared one-factor, two-factor, and bi-factor models. The following goodness-of-fit indices were used to assess the model fit to compare these three models: Tucker Lewis index (TLI; >0.90 acceptable, >0.95 excellent; Tucker & Lewis, 1973), the comparative fit index (CFI: >0.90 acceptable, >.095 excellent; Bentler, 1990), and root mean square error of approximation (*RMSEA*; <0.08 acceptable, <0.05 excellent; Brown & Cudeck, 1993). The standardized root mean square residual (*SRMR*) was also used. *SRMR* is the square root of the difference between the residuals of the sample covariance matrix and the hypothesized covariance model. Hu & Bentler (1999) suggested a cutoff value close to 0.08 for *SRMR* for a good fit between the hypothesized model and the observed data. R package lavaan (Rosseel, 2012) were used for the CFA.

Results

Tables 3, 4, and 5 display descriptive statistics for writing scores in Grades 5, 8, and 11. Each table includes data for overall ODW, followed by each writing subscore, as well as the overall EM, followed by each EM subscore. The second through the fifth columns in the table present the mean raw score (RS Mean), the raw score standard deviation (RS SD), the raw score minimum (RS Min), and raw score maximum (RS Max), respectively. The sixth through the ninth columns provide the same information for the scale scores. The raw score ranges of the EM2 and EM3 subscores were much smaller than that of the EM1 subscore, particularly for Grades 5 and 8, by design.

Table 3. Descriptive Statistics of Raw and Scale Scores of Overall and Subscores/Traits: Grade 5 (N= 42,434)

	RS Mean	RS SD	RS Min	RS Max	SS Mean	SS S.D.	SS Min	SS Max
ODW	21.41	7.06	10	40	506.85	36.72	425	600
WR1	4.64	1.55	2	8	489.94	52.88	400	600
WR2	4.48	1.53	2	8	494.47	51.40	400	600
WR3	3.10	1.56	2	8	489.74	48.88	440	600
WR4	4.47	1.67	2	8	481.52	54.69	400	600
WR5	4.72	1.51	2	8	480.72	51.59	400	600
EM	20.47	5.87	1	36	500.64	37.26	401	600
EM1	12.53	3.80	1	22	517.55	33.18	404	600
EM2	3.15	1.81	0	8	483.29	53.39	400	600
EM3	4.79	1.91	0	9	505.86	67.05	400	600

Note. RS=Raw score; SS=Scale score; S.D.=Standard deviation; Min=Minimum; Max=Maximum; EM1=Conventions of Standard English; EM2=Knowledge of Language; EM3=Vocabulary Acquisition and Use; WR1=Clarity/Coherence; WR2=Support; WR3=Sourcing; WR4=Organization; WR5=Language/Conventions. The number of items for subscores are different among three forms, which means the maximum subscores are different among three forms. That is the reason why the sum of RS Max EM1, EM2, and EM3 is not 36.

Table 4. Descriptive Statistics of Raw and Scale Scores of Overall and Subscores /Traits: Grade 8 (N= 43,797)

	RS Mean	RS S.D.	RS Min	RS Max	SS Mean	SS S.D.	SS Min	SS Max
ODW	23.84	8.78	12	48	496.08	39.79	430	600
WR1	4.29	1.53	2	8	465.20	54.90	400	600
Wr2	3.45	1.53	2	8	489.24	47.25	428	600
WR3	4.01	1.54	2	8	469.07	55.86	400	600
WR4	3.77	1.63	2	8	480.09	52.75	412	600
WR5	4.09	1.57	2	8	467.14	53.60	400	600
WR6	4.23	1.55	2	8	467.03	54.52	400	600
EM	19.59	6.37	0	36	517.34	40.95	400	600
EM1	9.69	3.26	0	20	488.20	29.57	400	600
EM2	2.49	2.19	0	6	468.12	63.81	400	592
EM3	7.41	2.33	0	12	504.68	46.70	400	594

Note. RS=Raw score; SS=Scale score; S.D.=Standard deviation; Min=Minimum; Max=Maximum; EM1=Conventions of Standard English; EM2=Knowledge of Language; EM3=Vocabulary Acquisition and Use; WR1=Clarity/Coherence; WR2=Counterclaims.WR3=Support; WR4=Sourcing; WR5=Organization; WR6=Language/Conventions. The number of items for subscores are different among three forms, which means the maximum subscores are different among three forms. That is the reason why the sum of RS Max EM1, EM2, and EM3 is not 36.

Table 5. Descriptive Statistics of Raw and Scale Scores of Overall and Subscores /Traits: Grade 11 (N= 37,631)

	RS Mean	RS S.D.	RS Min	RS Max	SS Mean	SS S.D.	SS Min	SS Max
ODW	25.04	9.16	12	48	495.93	38.97	424	600
WR1	4.50	1.61	2	8	475.02	56.80	400	600
WR2	3.70	1.61	2	8	491.65	50.34	429	600
WR3	4.33	1.62	2	8	480.62	56.94	400	600
WR4	3.53	1.79	2	8	504.22	38.04	464	600
WR5	4.36	1.64	2	8	477.77	56.32	400	600
WR6	4.63	1.54	2	8	475.75	56.59	400	600
EM	21.76	6.54	0	36	527.98	38.85	400	600
EM1	9.05	4.06	0	18	497.81	52.78	400	591
EM2	5.89	2.60	0	13	506.28	39.03	400	600
EM3	6.83	2.24	0	13	528.95	43.08	400	599

Note. RS= Raw score; SS= Scale score; S.D.= Standard deviation; Min= Minimum; Max= Maximum; EM1=Conventions of Standard English; EM2=Knowledge of Language; EM3=Vocabulary Acquisition and Use; WR1=Clarity/Coherence; WR2=Counterclaims.WR3=Support; WR4=Sourcing; WR5=Organization; WR6=Language/Conventions. The number of items for subscores are different among three forms, which means the maximum subscores are different among three forms. That is the reason why the sum of RS Max EM1, EM2, and EM3 is not 36.

Evidence That Test Sections Represent One Writing Construct

Tables 6, 7, and 8 present the correlations between ODW and EM for overall scale scores and scale subscores. The bold correlations are those within the same test. The italic correlations are those between OD and EM. For overall scores, the correlations between ODW and EM were all above 0.60 except for grade 5 (ODW vs. SS EM: 0.506) which indicates ODW and EM total scores measured the similar construct (acceptable level for convergent validity coefficient). The correlations between ODW and its subscores are consistently above .70 except for grade 5. The correlations between WR3 and other subscores at grade 5 are ranging from 0.548 to 0.574.

Table 6. Correlations Between ODW and EM Overall and Subscores: Grade 5 (N= 42,434)

	ODW	WR1	WR2	WR3	WR4	WR5	EM	EM1	EM2	EM3
ODW	1.000									.
WR1	**0.814**	1.000								
WR2	**0.821**	**0.920**	1.000							
WR3	**0.574**	**0.559**	**0.564**	1.000						
WR4	**0.817**	**0.894**	**0.895**	**0.548**	1.000					
WR5	**0.810**	**0.894**	**0.880**	**0.553**	**0.910**	1.000				
EM	*0.506*	*0.564*	*0.555*	*0.407*	*0.575*	*0.587*	1.000			
EM1	*0.455*	*0.511*	*0.507*	*0.365*	*0.523*	*0.532*	**0.848**	1.000		
EM2	*0.333*	*0.380*	*0.369*	*0.269*	*0.389*	*0.395*	**0.658**	**0.491**	1.000	
EM3	*0.221*	*0.269*	*0.266*	*0.179*	*0.265*	*0.272*	**0.422**	**0.414**	**0.435**	1.000

Note. EM1=Conventions of Standard English; EM2=Knowledge of Language; EM3=Vocabulary Acquisition and Use; WR1=Clarity/Coherence; WR2=Support; WR3=Sourcing; WR4=Organization; WR5=Language/Conventions.

Table 7. Correlations Between ODW and EM Overall and Subscores: Grade 8 (N= 43,797)

	ODW	WR1	WR2	WR3	WR4	WR5	WR6	EM	EM1	EM2	EM3
ODW	1.000										
WR1	**0.917**	1.000									
WR2	**0.835**	**0.810**	1.000								
WR3	**0.924**	**0.913**	**0.811**	1.000							
WR4	**0.886**	**0.850**	**0.810**	**0.855**	1.000						
WR5	**0.922**	**0.917**	**0.809**	**0.910**	**0.840**	1.000					
WR6	**0.910**	**0.912**	**0.802**	**0.892**	**0.835**	**0.912**	1.000				
EM	*0.605*	*0.579*	*0.544*	*0.570*	*0.548*	*0.579*	*0.601*	1.000			
EM1	*0.524*	*0.503*	*0.473*	*0.494*	*0.477*	*0.504*	*0.528*	**0.807**	1.000		
EM2	*0.442*	*0.423*	*0.405*	*0.417*	*0.406*	*0.422*	*0.438*	**0.704**	**0.577**	1.000	
EM3	*0.389*	*0.371*	*0.350*	*0.367*	*0.352*	*0.370*	*0.386*	**0.622**	**0.633**	**0.515**	1.000

Note. EM1=Conventions of Standard English; EM2=Knowledge of Language; EM3=Vocabulary Acquisition and Use; WR1=Clarity/Coherence; WR2=Support; WR3=Sourcing; WR4=Organization; WR5=Language/Conventions; WR6=Language/Conventions.

Table 8. Correlations Between ODW and EM Overall and Subscores: Grade 11 (N= 37,631)

	ODW	WR1	WR2	WR3	WR4	WR5	WR6	EM	EM1	EM2	EM3
ODW	1.000										
WR1	**0.935**	1.000									
WR2	**0.883**	**0.851**	1.000								
WR3	**0.933**	**0.938**	**0.853**	1.000							
WR4	**0.839**	**0.779**	**0.759**	**0.781**	1.000						
WR5	**0.933**	**0.940**	**0.851**	**0.943**	**0.770**	1.000					
WR6	**0.913**	**0.910**	**0.825**	**0.895**	**0.759**	**0.916**	1.000				
EM	*0.602*	*0.594*	*0.555*	*0.602*	*0.499*	*0.611*	*0.630*	1.000			
EM1	*0.542*	*0.530*	*0.503*	*0.535*	*0.460*	*0.546*	*0.571*	**0.799**	1.000		
EM2	*0.429*	*0.420*	*0.400*	*0.423*	*0.364*	*0.431*	*0.446*	**0.672**	**0.492**	1.000	
EM3	*0.229*	*0.226*	*0.211*	*0.229*	*0.182*	*0.232*	*0.234*	**0.349**	**0.272**	**0.575**	1.000

Note. EM1=Conventions of Standard English; EM2=Knowledge of Language; EM3=Vocabulary Acquisition and Use; WR1=Clarity/Coherence; WR2=Support; WR3=Sourcing; WR4=Organization; WR5=Language/Conventions; WR6=Language/Conventions.

We also evaluated the relationship between performance levels for ODW and EM as the overall writing performance level was determined by combining these two performance levels in which the ODW performance level was more weighted if they differed. Table 9 presents distributions of these two performance levels for each grade. The values on the diagonal within each grade level in Table 9 reflect the number of students classified in the same proficiency level by the EM and ODW assessments. By dividing by the total number of students at each grade level, we get the percentage of students consistently classified by the two measures. The percentage of students consistently classified are 31%, 49%, and 51% for Grades 5, 8, and 11, respectively. Polychoric correlations between ODW and EM performance levels were 0.53, 0.62, and 0.62 for Grades 5, 8, and 11, respectively. These correlations indicate that the strength of the relationship between ODW and EM performance levels is acceptable. These results support that ODW and EM are converging on a common construct. Grade 5 students tended to have higher performance levels for EM compared to ODW, while Grade 11 students tended to have more similar performance levels on both EM and ODW.

Table 9. Cross-table for Performance level of ODW and EM for each grade

		ODW N	ODW A	ODW P	ODW D
Grade 5	EM N	432 (0.97)	203 (0.46)	17 (0.04)	0 (0.00)
	EM A	6442 (14.46)	4679 (10.50)	1219 (2.74)	61 (0.14)
	EM P	5589 (12.55)	9694 (21.76)	6911 (15.51)	1321 (2.97)
	EM D	520 (1.17)	2353 (5.28)	3269 (7.34)	1838 (4.13)
Grade 8	EM N	367 (0.79)	589 (1.26)	23 (0.05)	3 (0.01)
	EM A	3083 (6.60)	11438 (24.47)	2631 (5.63)	554 (1.19)
	EM P	674 (1.44)	8696 (18.60)	7240 (15.49)	3876 (8.29)
	EM D	39 (0.08)	1170 (2.50)	2573 (5.50)	3789 (8.11)
Grade 11	EM N	249 (0.60)	411 (1.00)	16 (0.04)	3 (0.01)
	EM A	2187 (5.30)	8107 (19.66)	2173 (5.27)	184 (0.45)
	EM P	576 (1.40)	7350 (17.82)	10361 (25.12)	3141 (7.62)
	EM D	67 (0.16)	773 (1.87)	3130 (7.59)	2515 (6.10)

Note. N = Novice; A = Apprentice; P = Proficient; D = Distinguished. Numbers in parentheses are percentages of the total number of students at that grade level.

Reliability Analyses

Table 10 presents Cronbach's α for ODW and EM. Besides, it shows Cronbach's α with and without EM on the writing assessment. When EM items were added to ODW, Cronbach's α decreased slightly. However, Cronbach's α was still above 0.90, indicating excellent internal consistency (Tavakol & Dennick, 2011).

In general, when items are added to a test, Cronbach's α increases as Cronbach's α is a function of the number of items and the average correlation between items. The average correlations between ODW were relatively high (0.72 for Grade 5); however, the average correlations between EM items were relatively low (0.41 for Grade 5 for EM Form 1). This caused the reduction in Cronbach's α. See Appendix A for the full item correlation table for Grade 5 EM Form 1.

Table 10. Comparison of Cronbach's α with and Without EM in Writing Assessment

Grade	ODW Only	EM_F1 Only	EM_F2 Only	EM_F3 Only	ODW + EM_F1	ODW + EM_F2	ODW + EM_F3
5	0.961	0.818	0.779	0.789	0.922	0.912	0.914
8	0.976	0.815	0.818	0.818	0.936	0.936	0.936
11	0.977	0.838	0.833	0.862	0.942	0.940	0.945

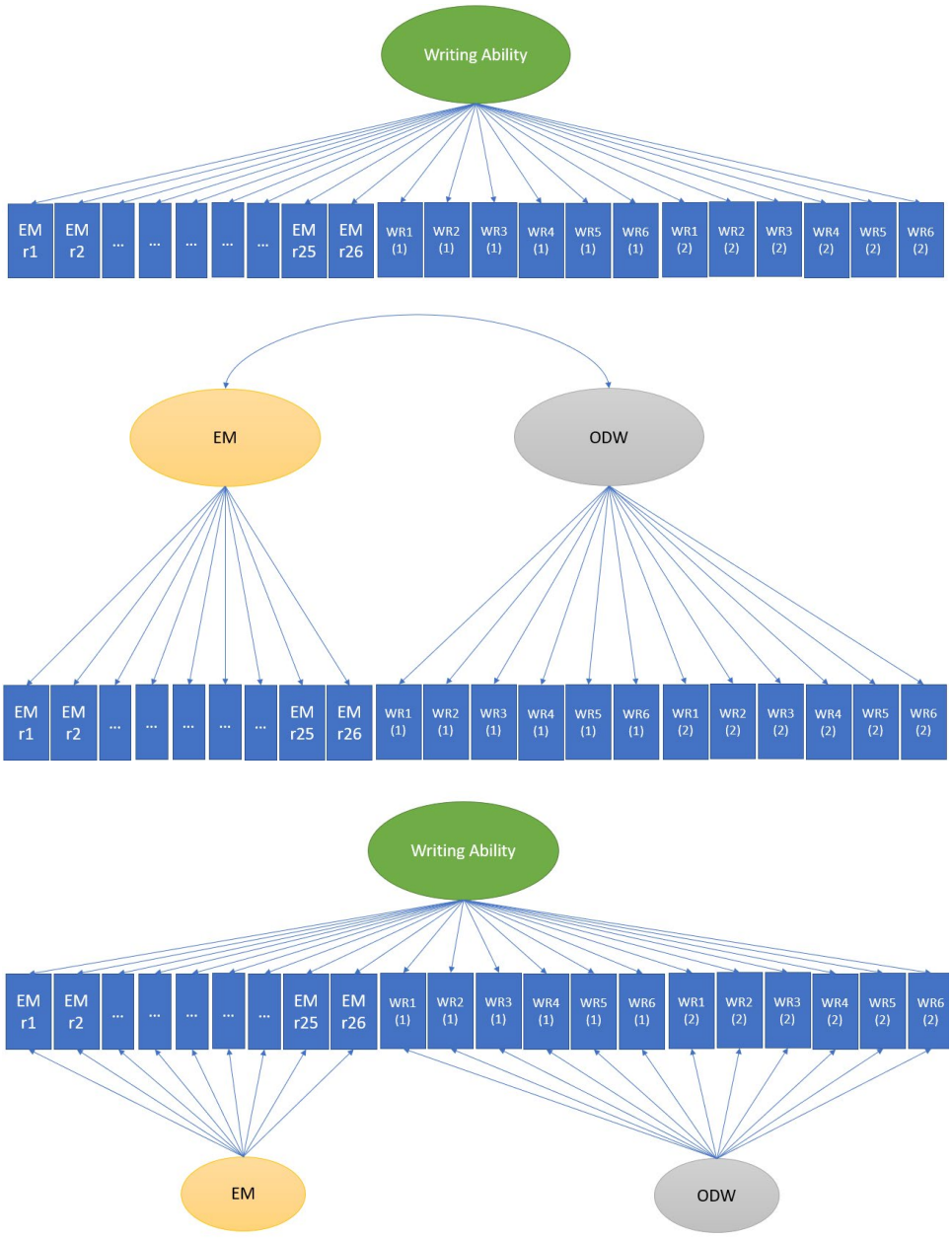
Note. EM_F1, EM_F2, and EM_F3 indicate the EM form.

Dimensionality Analysis

Confirmatory Factor Analysis (CFA)

We compared three different models. Figure 1 depicts one-factor, two-factor, and bi-factor models. EM consists of three subscores measured by various items. ODW consists of five (for Grade 5) or six (for Grades 8 and 11) traits measured by two raters. Note in the figure, (1) indicates rater 1 and (2) indicates rater 2.

Figure 1. Schematic representation of three confirmatory factor models: One-factor, two-factor, and bi-factor model



Tables 11, 12, and 13 display the model fit statistics for the one-factor, two-factor, and bi-factor models, respectively. For all three models, both CFI and TLI statistics across the forms and grades were above 0.95, suggesting excellent model fit. Both *RMSEA* and *SRMR* suggested the two-factor model fit excellently whereas the one-factor model fit acceptably. It is clear *RMSEA* and *SRMR* suggested the two-factor model than the one-factor model, and CFI and TLI suggested both the one- and two-factor models fit well to the data. For the two-factor model, the correlations between the two factors were greater than 0.6 indicating they are highly correlated to each other. All fit indices suggest the bi-factor model fits the best for this data supporting a common general factor presents as well as two specific factors.

It is not uncommon for *RMSEA* (an absolute fit index) and CFI (a relative fit index) to disagree in terms of goodness of fit. That is because they evaluate the magnitude of the model's fit function value from different perspectives. There are a couple of possible reasons to explain this discrepancy among fit indices. First, *RMSEA* tends to reward more complex models with large samples, whereas the CFI is less influenced by sample size, and it penalizes complex models (Peugh & Feldon, 2020). The data are, in fact, very large (Ns were greater than 10,000) and the two-factor model is a more complex model than the one-factor model. The other possible explanation is the method effect for the two-factor model. Method effect refers to the effect that is attributable to the measurement method rather than to the construct of interest (Podsakoff et al., 2003). For example, negatively and positively worded items can create a method effect. In our study, item response types can create a method effect. That is, EM items scores were from multiple-choice items or short-answer items, whereas ODW were trait scores for constructed response items.

Table 11. Model Fit Statistics for One-Factor Model

Grade	Assessment	CFI	TLI	RMSEA	Lower	Upper	SRMR
5	ODW+EM_F1	.994	.993	.083	.083	.084	.086
	ODW+EM_F2	.994	.994	.076	.076	.077	.084
	ODW+EM_F3	.994	.993	.080	.080	.081	.084
8	ODW+EM_F1	.996	.996	.072	.072	.073	.075
	ODW+EM_F2	.996	.995	.070	.069	.070	.079
	ODW+EM_F3	.996	.996	.069	.068	.069	.073
11	ODW+EM_F1	.997	.997	.072	.071	.073	.076
	ODW+EM_F2	.997	.997	.070	.070	.071	.076
	ODW+EM_F3	.996	.995	.083	.082	.084	.091

Note. EM_F1, EM_F2, and EM_F3 indicate the EM form combined with ODW; Lower=*RMSEA* 90% lower confidence interval; Upper=*RMSEA* 90% upper confidence interval.

Table 12. Model Fit Statistics for Two-Factor Model

Grade	Assessment	CFI	TLI	RMSEA	Lower	Upper	SRMR	Corr
5	ODW+EM_F1	.997	.996	.061	.060	.061	.047	.678
	ODW+EM_F2	.997	.996	.058	.058	.059	.048	.665
	ODW+EM_F3	.996	.996	.063	.063	.064	.052	.674
8	ODW+EM_F1	.998	.998	.050	.049	.050	.045	.697
	ODW+EM_F2	.998	.998	.047	.047	.048	.046	.700
	ODW+EM_F3	.998	.998	.047	.047	.048	.040	.700
11	ODW+EM_F1	.999	.999	.045	.045	.046	.029	.693
	ODW+EM_F2	.999	.999	.045	.044	.046	.028	.663
	ODW+EM_F3	.999	.999	.047	.046	.047	.031	.655

Note. EM_F1, EM_F2, and EM_F3 indicate the EM form combined with ODW; Lower=*RMSEA* 90% lower confidence interval; Upper=*RMSEA* 90% upper confidence interval; Corr=correlation between two factors.

Table 13. Model Fit Statistics for Bi-factor Model

Grade	Assessment	CFI	TLI	RMSEA	Lower	Upper	SRMR
5	ODW+EM_F1	.999	.999	.037	.036	.038	.039
	ODW+EM_F2	.999	.999	.035	.035	.036	.040
	ODW+EM_F3	.999	.998	.040	.040	.041	.045
8	ODW+EM_F1	.999	.999	.034	.034	.035	.042
	ODW+EM_F2	.999	.998	.041	.040	.041	.042
	ODW+EM_F3	.999	.999	.032	.032	.033	.039
11	ODW+EM_F1	.999	.999	.031	.031	.032	.026
	ODW+EM_F2	1.000	.999	.029	.029	.030	.024
	ODW+EM_F3	.999	.999	.032	.032	.033	.028

Note. EM_F1, EM_F2, and EM_F3 indicate the EM form combined with ODW; Lower=*RMSEA* 90% lower confidence interval; Upper=*RMSEA* 90% upper confidence interval.

Discussion

This study evaluated the impact of adding EM items to the current ODW assessment. We found evidence to support the reliability and validity of the combined writing assessment, though with some mixed results. Cronbach's α indicated very strong internal consistency. Though internal consistency reliability dropped slightly when EM items were added, the overall reliability was still quite high. The reduced reliability was a function of very low correlations among many EM items.

The correlations between EM and ODW overall scores were moderate. The correlations within ODW are consistently high except for WR3 of Grade 5. In terms of the correlations within EM, the correlations between EM3 and total scores or other subscores are consistently weak. This is not surprising since the subscores are based on a smaller subset of items that are intended to measure different aspects of the larger constructs.

A comparison of the distributions of ODW and EM performance levels showed that at the lower grades, students tended to perform better on EM than on ODW. This may reflect grade-appropriate developmental differences in the EM and ODW constructs.

CFA results strongly support the bi-factor model, suggesting that the set of EM and ODW items reflect separate EM and ODW factors, as well as an overall writing ability construct.

Overall, the results from this study support Kentucky's current approach to assessing writing that combines editing and mechanics items with on-demand writing prompts. Currently, EM and ODW scale scores are estimated separately, and only combined after student performance levels on the two tests have been assigned. KDE should consider exploring scoring the writing assessment within a bi-factor model framework, as this may better capture the structure of the writing construct.

References

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246.

Brown, M. W., & Cudeck, R. (1993). Alternative ways of assessing model ®t. In K. A. Bollen, & J. S. Long, *Testing structural equation models* (pp. 136±162). Newbury Park, CA: Sage

Carlson, K. D., & Herdman, A. O. (2012). Understanding the impact of convergent validity on research results. *Organizational Research Methods, 15*(1), 17–32.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

George, D & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Allyn & Bacon.

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), 1-55.

Kane, M. (2006). Content-related validity evidence in test development. *Handbook of test development, 1*, 131-153.

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). Routledge

Peugh, J., & Feldon, D. F. (2020). “How well does your structural equation model fit your data?”: Is Marcoulides and Yuan’s equivalence test the answer? *CBE—Life Sciences Education, 19*(3), es5.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, *88*(5), 879.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of statistical software, 48*, 1-36

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1–10.

Appendix A: Item Correlations for EM and trait correlation for ODW (Grade 5)

Table A-1. Item Correlations for Grade 5 EM Form 1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	1.00	.12	.15	.16	.16	.18	.21	.21	.14	.15	.11	.11	.16	.18	.12	.17	.08	.16	.13	.06	.04	.17	.16	.12	.23	.07
2		1.00	.12	.12	.15	.14	.14	.16	.12	.12	.09	.09	.13	.12	.11	.13	.08	.11	.10	.04	.03	.12	.12	.11	.16	.07
3			1.00	.19	.19	.24	.16	.33	.29	.15	.16	.19	.22	.25	.19	.17	.17	.18	.20	.10	-.01	.31	.23	.16	.20	.04
4				1.00	.20	.22	.22	.21	.20	.41	.13	.19	.22	.17	.14	.20	.12	.19	.29	.09	.12	.20	.18	.16	.22	.09
5					1.00	.23	.16	.25	.20	.19	.15	.15	.23	.21	.15	.26	.11	.19	.15	.08	.06	.22	.20	.16	.21	.07
6						1.00	.18	.26	.22	.20	.16	.18	.29	.21	.20	.22	.11	.24	.19	.11	.05	.24	.25	.17	.23	.06
7							1.00	.21	.14	.19	.11	.11	.19	.16	.09	.18	.08	.17	.16	.06	.09	.15	.16	.12	.20	.10
8								1.00	.24	.16	.21	.17	.25	.37	.23	.22	.12	.19	.21	.13	-.01	.40	.31	.20	.24	.05
9									1.00	.18	.14	.31	.19	.16	.19	.17	.27	.16	.15	.09	.01	.21	.21	.16	.21	.05
10										1.00	.13	.15	.21	.13	.15	.20	.11	.18	.24	.08	.11	.13	.18	.15	.22	.09
11											1.00	.11	.15	.16	.15	.16	.08	.14	.13	.07	.00	.17	.20	.11	.16	.04
12												1.00	.17	.11	.15	.15	.21	.14	.13	.07	.02	.15	.16	.13	.18	.05
13													1.00	.22	.17	.23	.10	.38	.21	.10	.09	.22	.22	.17	.22	.08
14														1.00	.15	.28	.08	.20	.19	.10	-.01	.33	.21	.14	.18	.05
15															1.00	.14	.13	.12	.13	.10	-.03	.22	.24	.21	.21	.04
16																1.00	.09	.21	.17	.08	.06	.18	.20	.17	.22	.07
17																	1.00	.09	.09	.04	.00	.11	.10	.10	.13	.05
18																		1.00	.18	.08	.08	.17	.18	.13	.18	.09
19																			1.00	.07	.02	.23	.18	.13	.16	.02
20																				1.00	.00	.12	.12	.09	.09	.03
21																					1.00	-.03	.01	.02	.04	.07
22																						1.00	.29	.19	.22	.04
23																							1.00	.21	.25	.04
24																								1.00	.21	.05
25																									1.00	.08
26																										1.00

Table A-2. Correlation for Trait score for Grade 5 ODW

	WR1 (1)	WR2 (1)	WR3 (1)	WR4 (1)	WR5 (1)	WR1 (2)	WR2 (2)	WR3 (2)	WR4 (2)	WR5 (2)
WR1 (1)	1.00	.87	.55	.85	.85	.80	.79	.47	.79	.78
WR2 (1)		1.00	.56	.85	.83	.78	.81	.48	.79	.77
WR3 (1)			1.00	.55	.54	.52	.52	.83	.52	.52
WR4 (1)				1.00	.87	.77	.78	.47	.83	.78
WR5 (1)					1.00	.77	.77	.47	.79	.81
WR1 (2)						1.00	.90	.50	.87	.88
WR2 (2)							1.00	.51	.86	.85
WR3 (2)								1.00	.50	.50
WR4 (2)									1.00	.87
WR5 (2)										1.00