

2010 Missouri End-of-Course Assessments Comparison of ARC Hand Scoring with Teacher Scores

Introduction

The Missouri End-of-Course (EOC) Assessments were created to adapt the State's testing needs of Missouri districts, schools, teachers, and students, while still meeting state and federal accountability requirements. The Missouri State Board of Education identified the following purposes for the Missouri EOC Assessments:

- Measuring and reflecting students' mastery toward post-secondary readiness
- Identifying students' strengths and weaknesses
- Communicating expectations for all students
- Serving as the basis for state and national accountability plans
- Evaluating programs

Course Level Expectations (CLEs) outline the ideas, concepts, and skills that form the foundation for an assessed EOC subject area, regardless of student grade level. Because a course such as Algebra I could be delivered in middle school or any grade level in secondary school, CLEs replaced the Grade Level Expectations (GLEs). Districts can offer courses with different titles that cover the same CLEs.

Each Algebra I, English II, and Biology EOC assessment includes two types of test items: selected-response items and performance events (PEs), which include writing prompts. The selected-response items present students with a question followed by four response options. The performance events require students to work through more complicated items. Performance events often allow for more than one approach to arrive at a correct response. The advantage of this type of assessment item is that it provides insight into a student's ability to apply knowledge and understanding in real-life situations.

The writing prompt, a special type of performance event that appears in the English II assessment, is an open-ended item that requires students to demonstrate their writing proficiency. Writing is scored holistically using a four-point scoring guide.

Hand-Scoring of PE Items for the Operational EOC Assessments

Student responses to the operational PE items were hand scored for state reporting and federal AYP determinations by professionally trained raters at the Assessment Resource Center (ARC). ARC provided these raters with extensive training using industry standards for qualification, including formal educational background and percent exact agreement criteria, to determine eligibility for scoring PEs on the EOC Assessments.

For the Spring 2010 administration of the Missouri EOC assessments, the following operational PE items were scored:

<u>EOC Assessment</u>	<u>Points Possible</u>
English II Writing Prompt ¹	1–4
Algebra I PE Item	0–4
Biology PE Items	
PE Item 1	0–1
PE Item 2	0–1
PE Item 3	0–2
PE Item 4	0–2
PE Item 5	0–2
PE Item 6	0–4
PE Item 7	0–1
PE Item 8	0–1
PE Item 9	0–1
PE Item 10	0–1
PE Item 11	0–1
PE Item 12	0–3
Total Biology PE Points	0–20

In addition to the PE scores provided by ARC, Missouri teachers were given the opportunity to score their students' responses to the PEs on the operational assessments. Teachers had access to scanned images of their students' responses to each PE through an internet-based teacher interface. Teachers were provided online training via the *Task Scoring Workshop*—a web-based, interactive tutorial. The tutorial included the same anchor and training papers that were used by the professional scorers at ARC. Teachers also had the option to use the resulting scores as a component of their students' final course grades.

The Task Scoring Workshop is an Internet-based tool that provides training on the scoring of open-ended test items. It is designed as a staff development tool that brings consistency to the scoring process for constructed response items. On-board tools, like scoring rubrics, anchor papers, annotations and practice papers help trainees practice the scoring process independently. Teachers receive continuous feedback on their progress, allowing them to fine tune their understanding of the scoring process.

The purpose of the current study is to provide a comparison of the official 2010 ARC PE scores with the scores provided by Missouri teachers. The two scores (i.e., the ARC score and the teacher score) were obtained through a process of matching barcode numbers from student response booklets. The resulting analyses provided in the next section focus on the percent agreement between ARC scores and teacher scores. The implications of the results for Missouri teachers and state policy makers are then considered.

¹The minimum score for a response in English II is 1 point. If a student does not respond to the WP in English II, he or she will receive a "condition code," equivalent to a score of 0. Condition codes are used when a paper is not scorable for one or more reasons.

Results

For each EOC assessment, the ARC score and the teacher score were compared by examining (a) the percent agreement at each score point; (b) the percent exact agreement overall and exact plus adjacent agreement overall; (c) the difference between mean PE scores overall; and (d) the Pearson correlation between the two sets of scores.

English II

The results for English II are presented in Tables 1 through 3. Table 1 provides a cross-tabulation of ARC scores with Missouri teacher scores. Table 2 provides the frequency distribution of the difference between the teacher scores and the ARC scores. Table 3 provides the mean ARC score and the mean teacher score, including the Pearson correlation between the two sets of scores.

Examination of Tables 1 and 2 shows the overall exact agreement between ARC scores with Missouri teacher scores to be 51.3%. The overall exact plus adjacent agreement was found to be 97.2%. While these percentages are encouraging, especially with respect to the use of EOC student scores for the purpose of assigning course grades, they are lower than the generally accepted standard in the industry. Typically, when professional raters are scoring a constructed-response item with a 0 to 4 scoring rubric, exact agreement is expected to be approximately 80.0% and exact plus adjacent agreement is expected to be 99.0%. Still, given that teachers self-trained using an online tutorial with no supervision or qualifying required, the agreement rates seem reasonable given the intended use of the scores.

Table 1: English II Cross-Tabulation of ARC Scores with Missouri Teacher Scores

Subject	ARC Item Score	Row Label	Teacher Item Score				
			0	1	2	3	4
English II	0	<i>N</i>	80	0	0	1	0
		% of 0-Point ARC Scores	98.77	0.00	0.00	1.23	0.00
		% of Total Scores	0.35	0.00	0.00	0.00	0.00
	1	<i>N</i>	21	207	44	4	0
		% of 1-Point ARC Scores	7.61	75	15.94	1.45	0.00
		% of Total Scores	0.09	0.91	0.19	0.02	0.00
	2	<i>N</i>	15	622	1166	293	16
		% of 2-Point ARC Scores	0.71	29.45	55.21	13.87	0.76
		% of Total Scores	0.07	2.72	5.10	1.28	0.07
	3	<i>N</i>	35	433	4885	8450	3624
		% of 3-Point ARC Scores	0.20	2.48	28.03	48.49	20.80
		% of Total Scores	0.15	1.90	21.38	36.99	15.86
	4	<i>N</i>	7	9	131	991	1812
		% of 4-Point ARC Scores	0.24	0.31	4.44	33.59	61.42
		% of Total Scores	0.03	0.04	0.57	4.34	7.93

Table 2: English II Distribution of the Difference between Teacher Scores and ARC Scores

MO EOC, Spring 2010 PE Item Scoring ARC Total - Teacher Total			
Subject	ARC Total - Teacher Total	N	Percent
English II	-3	1	0.00
	-2	20	0.09
	-1	3961	17.34
	0	11715	51.28
	1	6519	28.53
	2	579	2.53
	3	44	0.19
	4	7	0.03

The difference between English II mean scores provided in Table 3 and the frequency distribution in Table 2 shows that ARC raters were more lenient than Missouri teachers. The ARC mean score was 3.00, where the mean score for teachers was found to be 2.83. As indicated by standard deviations of 0.55 and 0.88, respectively, the teacher distribution was slightly more variable than the distribution of ARC’s raters. This is consistent with the hypothesis that the more careful training and monitoring of professional scorers can lead to more accuracy and precision in scoring. Finally, the correlation between the two sets of scores was .53, which represents a moderate to moderately-high correlation coefficient.

Table 3: English II Summary Statistics for ARC Scores and Teacher Scores

MO EOC, Spring 2010 PE Scoring Summary Statistics					
Subject	Scorers	N	Mean	SD	Correlation
English II	ARC	22846	3.00	.55	.53
	Teachers	22846	2.83	.88	

Algebra I

The results for Algebra I are presented in Tables 4 through 6. Table 4 provides a cross-tabulation of ARC scores with Missouri teacher scores. Table 5 provides the frequency distribution of the difference between the teacher scores and the ARC scores. Table 6 provides the mean ARC score and the mean teacher score, including the Pearson correlation between the two sets of scores. Examination of Tables 4 and 5 shows the overall exact agreement between ARC scores with Missouri teacher scores to be 52.9%. The overall exact plus adjacent agreement was 91.5%. Similar to the English II results,

these percentages seem reasonable, especially with respect to the intended use of the EOC teacher scores.

Table 4: Algebra I Cross-Tabulation of ARC Scores with Missouri Teacher Scores

Subject	ARC Item Score	Row Label	Teacher Item Score				
			0	1	2	3	4
Algebra I	0	<i>N</i>	1832	1405	398	71	17
		% of 0-Point ARC Scores	49.21	37.74	10.69	1.91	0.46
		% of Total Scores	9.07	6.96	1.97	0.35	0.08
	1	<i>N</i>	316	2189	1395	508	60
		% of 1-Point ARC Scores	7.07	48.99	31.22	11.37	1.34
		% of Total Scores	1.57	10.84	6.91	2.52	0.3
	2	<i>N</i>	16	496	2250	1298	484
		% of 2-Point ARC Scores	0.35	10.92	49.52	28.57	10.65
		% of Total Scores	0.08	2.46	11.14	6.43	2.4
	3	<i>N</i>	4	99	1603	1759	999
% of 3-Point ARC Scores		0.09	2.22	35.91	39.4	22.38	
% of Total Scores		0.02	0.49	7.94	8.71	4.95	
4	<i>N</i>	2	2	57	284	2647	
	% of 4-Point ARC Scores	0.07	0.07	1.91	9.49	88.47	
	% of Total Scores	0.01	0.01	0.28	1.41	13.11	

The difference between the Algebra I mean scores provided in Table 6 and the frequency distribution in Table 5 shows that unlike the scoring of the English II writing task, Missouri teachers were more lenient than ARC raters when they scored the Algebra I PE. In Algebra I, the ARC mean score was 1.93 where the mean score for teachers was found to be 2.19. The correlation between the two sets of scores was .79 which represents a moderately-high correlation coefficient. This would be consistent with the possibility that the scoring guide for Algebra I might have been less subjective (less call for rater judgment) than the scoring guide for English II.

Table 5: Algebra I Distribution of the Difference between Teacher Scores and ARC Scores

MO EOC, Spring 2010 PE Scoring Teacher Total - ARC Total			
Subject	Teacher Total- ARC Total	N	Percent
Algebra I	-4	17	0.08
	-3	131	0.65
	-2	1390	6.88
	-1	5097	25.24
	0	10677	52.88
	1	2699	13.37
	2	172	0.85
	3	6	0.03
	4	2	0.01

Table 6: Algebra I Summary Statistics for ARC Scores and Teacher Scores

MO EOC, Spring 2010 PE Scoring Summary Statistics					
Subject	Scorers	N	Mean	SD	Correlation
Algebra I	ARC	20191	1.93	1.33	.79
	Teachers	20191	2.19	1.28	

Biology

The results for Biology are presented in Tables 7 through 10. Because there are multiple parts or items for Biology, the results for this content area are presented differently than the other two content areas. Table 7 provides the frequency distribution of the difference between the teacher item scores and the ARC item scores. Table 8 provides the frequency distribution of the difference between the teacher total score and the ARC total score. Recall that the total Biology PE score range is from 0 to 20. Table 9 provides a cross-tabulation of ARC scores with Missouri teacher scores for each of the 12 Biology PE items. Table 10 provides the total mean ARC score and the total mean teacher score, including the Pearson correlation between the two sets of scores.

Combining the exact agreement percentage from across all 12 PE items (Table 7), the overall exact agreement between ARC scores with Missouri teacher scores is 70.7%. The overall exact plus adjacent agreement is 94.0%. These percentages are generally higher than the other content areas because most of the Biology PE items are scored 0 to 1, 0 to 2, or 0 to 3. Similar to the other content areas, these percentages seem reasonable, especially with respect to the intended use of the EOC teacher scores.

Table 7: Biology Distribution of the Difference between Teacher Scores and ARC Scores for PEs

MO EOC, Spring 2010 PE Scoring Teacher Item Score - ARC Item Score			
Subject	Teacher Score- ARC Score	N	Percent
Biology	-4	12	0
	-3	254	0.1
	-2	1639	0.62
	-1	17792	6.73
	0	186880	70.69
	1	43732	16.54
	2	12070	4.57
	3	1907	0.72
	4	84	0.03

Table 8: Biology Distribution of the Difference between Teacher Scores and ARC Scores for the Total PE Score

MO EOC, Spring 2010 PE Total Scoring Teacher Total - ARC Total			
Subject	Teacher Total- ARC Total	N	Percent
Biology	-19	1	0.00
	-17	1	0.00
	-16	2	0.01
	-15	4	0.02
	-14	3	0.01
	-13	11	0.05
	-12	29	0.13
	-11	63	0.28
	-10	118	0.53
	-9	251	1.13
	-8	441	1.98
	-7	802	3.60
	-6	1215	5.45
	-5	1838	8.24
	-4	2487	11.15
	-3	2920	13.09
	-2	3179	14.25
	-1	2974	13.33

Table 8: Biology Distribution of the Difference between Teacher Scores and ARC Scores for the Total PE Score (continued)

MO EOC, Spring 2010 PE Total Scoring Teacher Total - ARC Total			
Subject	Teacher Total- ARC Total	N	Percent
Biology continued	0	2568	11.51
	1	1675	7.51
	2	886	3.97
	3	484	2.17
	4	205	0.92
	5	79	0.35
	6	38	0.17
	7	17	0.08
	8	3	0.01
	9	2	0.01
	10	2	0.01
	11	2	0.01
	12	1	0.00
	13	2	0.01

Table 9: Biology Cross-Tabulation of ARC Scores with Missouri Teacher Scores

Subject	Item	ARC Item Score	Row Label	Teacher Item Score				
				0	1	2	3	4
Biology	1	0	<i>N</i> % of 0-Point ARC Scores % of Total Scores	8417 82.34 37.83	1805 17.66 8.11	/	/	/
		1	<i>N</i> % of 1-Point ARC Scores % of Total Scores	527 4.38 2.37	11498 95.62 51.68	/	/	/
		2	<i>N</i> % of 2-Point ARC Scores % of Total Scores	11441 93.66 51.55	774 6.34 3.49	/	/	/
	2	0	<i>N</i> % of 0-Point ARC Scores % of Total Scores	221 2.21 1.00	9758 97.79 43.97	/	/	/
		1	<i>N</i> % of 1-Point ARC Scores % of Total Scores	4611 34.38 20.98	4764 35.52 21.67	4037 30.10 18.37	/	/
		2	<i>N</i> % of 2-Point ARC Scores % of Total Scores	411 7.19 1.87	1972 34.51 8.97	3332 58.30 15.16	/	/
	3	0	<i>N</i> % of 0-Point ARC Scores % of Total Scores	65 2.28 0.30	577 20.21 2.62	2213 77.51 10.07	/	/
		1	<i>N</i> % of 1-Point ARC Scores % of Total Scores	2312 52.73 10.44	1351 30.81 6.1	722 16.46 3.26	/	/
		2	<i>N</i> % of 2-Point ARC Scores % of Total Scores	549 5.26 2.48	5705 54.71 25.76	4174 40.03 18.85	/	/
	4	0	<i>N</i> % of 0-Point ARC Scores % of Total Scores	182 2.48 0.82	1781 24.28 8.04	5372 73.24 24.26	/	/
		1	<i>N</i> % of 1-Point ARC Scores % of Total Scores	2708 45.47 12.33	2013 33.80 9.17	1234 20.72 5.62	/	/
		2	<i>N</i> % of 2-Point ARC Scores % of Total Scores	988 10.26 4.50	4603 47.78 20.96	4042 41.96 18.41	/	/
	5	0	<i>N</i> % of 0-Point ARC Scores % of Total Scores	169 2.65 0.77	1436 22.54 6.54	4766 74.81 21.70	/	/
		1	<i>N</i> % of 1-Point ARC Scores % of Total Scores				/	/
		2	<i>N</i> % of 2-Point ARC Scores % of Total Scores				/	/

Subject	Item	ARC Item Score	Row Label	Teacher Item Score				
				0	1	2	3	4
Biology continued	6	0	<i>N</i>	1676	895	499	262	84
			% of 0-Point ARC Scores	49.06	26.20	14.61	7.67	2.46
			% of Total Scores	7.58	4.05	2.26	1.18	0.38
		1	<i>N</i>	1988	1917	1254	988	489
			% of 1-Point ARC Scores	29.96	28.89	18.90	14.89	7.37
	2	% of Total Scores	8.99	8.67	5.67	4.47	2.21	
		<i>N</i>	195	498	961	1542	1136	
	3	% of 2-Point ARC Scores	4.50	11.50	22.18	35.60	26.22	
		% of Total Scores	0.88	2.25	4.34	6.97	5.14	
	4	<i>N</i>	45	170	575	1661	2072	
		% of 3-Point ARC Scores	0.99	3.76	12.71	36.72	45.81	
	5	% of Total Scores	0.20	0.77	2.60	7.51	9.37	
		<i>N</i>	12	30	146	742	2284	
	6	% of 4-Point ARC Scores	0.37	0.93	4.54	23.09	71.06	
		% of Total Scores	0.05	0.14	0.66	3.35	10.33	
	7	0	<i>N</i>	11026	1380			
			% of 0-Point ARC Scores	88.88	11.12			
	8	1	% of Total Scores	50.66	6.34			
			<i>N</i>	471	8886			
	9	0	% of 1-Point ARC Scores	5.03	94.97			
% of Total Scores			2.16	40.83				
10	1	<i>N</i>	8065	1058				
		% of 0-Point ARC Scores	88.40	12.60				
11	0	% of Total Scores	36.7	4.82				
		<i>N</i>	1206	11644				
12	1	% of 1-Point ARC Scores	9.39	90.61				
		% of Total Scores	5.49	52.99				
13	0	<i>N</i>	6775	1919				
		% of 0-Point ARC Scores	77.93	22.07				
14	1	% of Total Scores	30.70	8.70				
		<i>N</i>	1211	12165				
15	0	% of 1-Point ARC Scores	9.05	90.95				
		% of Total Scores	5.49	55.12				
16	1	<i>N</i>	3570	2033				
		% of 0-Point ARC Scores	63.72	36.28				
17	0	% of Total Scores	16.15	9.20				
		<i>N</i>	1067	15437				
18	1	% of 1-Point ARC Scores	6.47	93.53				
		% of Total Scores	4.83	69.83				
19	0	<i>N</i>	4341	4097				
		% of 0-Point ARC Scores	51.45	48.55				
20	1	% of Total Scores	19.67	18.57				

Subject	Item	ARC Item Score	Row Label	Teacher Item Score					
				0	1	2	3	4	
Biology continued	1	<i>N</i>		891	12735				
		% of 1-Point ARC Scores		6.54	93.46				
		% of Total Scores		4.04	57.72				
	12	0	<i>N</i>		3604	2505	1956	1156	
			% of 0-Point ARC Scores		39.08	27.17	21.21	12.54	
			% of Total Scores		16.58	11.52	9.00	5.32	
		1	<i>N</i>		582	1196	1711	1498	
			% of 1-Point ARC Scores		11.67	23.98	34.31	30.04	
	2	<i>N</i>		190	480	833	1011		
		% of 2-Point ARC Scores		7.56	19.09	33.13	40.21		
	3	<i>N</i>		179	522	1591	2728		
		% of 3-Point ARC Scores		3.57	10.40	31.69	54.34		
		% of Total Scores		0.82	2.40	7.32	12.55		

Table 10: Biology Summary Statistics for ARC Scores and Teacher Scores

MO EOC, Spring 2010 PE Scoring Summary Statistics					
Subject	Scorers	N	Mean	SD	Correlation
Biology	ARC	22303	9.55	4.52	.82
	Teachers	22303	11.87	4.95	

The difference between the Biology total mean scores provided in Table 10 shows that Missouri teachers were more lenient than ARC raters. The ARC total mean score was 9.55 and the teacher mean score was 11.87. The correlation between the two sets of scores was .82 which represents a relatively high correlation coefficient.

Possible Implications of the Results

The following bullet points outline some possible implications of the results of the 2010 comparison of scoring of PEs between ARC raters and Missouri teachers.

- While teachers in each content area used the teacher interface to score PE item responses for over 20,000 students, the responses for approximately 45,000 other students per content area were not scored by teachers. Because not all students in the state had their responses scored by their teacher, and not all Missouri teachers used the teacher interface for scoring, the sample of teachers who scored items and were compared to ARC raters in the present study could be biased and not representative of the entire population of Missouri teachers. It is conceivable that the teachers who did participate were more convinced of the importance of using the scores as part of course grades, more comfortable with the task of scoring by rubric, and otherwise more engaged than other Missouri teachers.
- While the percentages for exact agreement and exact plus adjacent agreement were reasonable, especially with respect to the current use of EOC student scores for the purpose of assigning course grades, they were found to be lower than the generally accepted standard in the industry. DESE may want to further evaluate the online training for teachers to determine possible improvements or other activities that could enhance teacher scoring.
- The correlations between the two sets of scores for Algebra I and Biology were moderately high to high, suggesting a strong relationship between how ARC raters approached scoring and how Missouri teachers approached the same scoring task. The higher correlation for Biology could be due in part to the more structured scoring of the items and the larger number of score points. The English II correlation was considerably lower which could reflect differences in the clarity of the scoring rubric, or possible issues with the amount of teacher training required for a writing sample as opposed to the PE items in Algebra I or Biology.

Summary

In summary, we believe these results to be very encouraging. While the results do not suggest that teachers, who self-select for participation in the teacher scoring, are trained remotely, and score unmonitored, can provide an acceptable level of scoring to replace professional scorers, the results do suggest that the teachers in this study, on the whole, were very conscientious and able to provide results that were valid for local use.

Further evidence for this last point is provided in Table 11 which gives a side-by-side comparison of statistics between the 2009 and 2010 ARC scores and teacher scores. With only one exception, i.e., the Pearson correlation for English II, teacher scoring showed improvement from 2009 to 2010. For example, the percent exact agreement, percent exact plus adjacent agreement and the correlation were all higher in 2010 (with the one exception). Riverside suggests continued monitoring of agreement rates between teachers and professional scorers.

Table 11: Comparison between 2009 and 2010 ARC Scores and Teacher Scores

Subject	Statistic	Year	
		2009	2010
English II	% Exact Agreement	49.5	51.3
	% Exact plus Adjacent	95.0	97.2
	Pearson Correlation	.58	.53
Algebra I	% Exact Agreement	49.1	52.9
	% Exact plus Adjacent	84.5	91.5
	Pearson Correlation	.70	.79
Biology	% Exact Agreement	66.6	70.7
	% Exact plus Adjacent	93.4	94.0
	Pearson Correlation	.79	.82