

Proposal for the Removal of Constructed Response Items

The Missouri Department of Elementary and Secondary Education (DESE) requested that CTB investigate the possibility of removing a limited number of constructed response (CR) items from the Missouri Assessment Program (MAP) for the 2010 Spring Administration. In removing CR items, it is important to address the question of comparability: will scores from the proposed 2010 MAP be comparable to scores from the 2006 to 2009 MAP? In removing the CR items, it is of primary importance that the tests continue to measure the same construct as they have in past administrations with adequate reliability.

In order to create comparable test forms, test makers adhere to test blueprints that outline the proportion of points that should be used for each content standard. They also maintain strict psychometric guidelines that ensure alternate forms of a test are of equivalent difficulty. While the removal of CR items is not optimal, it does not necessarily undermine the comparability of test scores of future MAP forms to past MAP forms. It is important to examine the degree of the changes to determine if the underlying construct is being impacted. As part of this proposal, we will examine whether the changes to the 2010 MAP will affect the comparability of the proposed 2010 MAP to previous MAP forms.

This brief proposal will first outline CTB's plan to remove a limited number of CR items from the Communication Arts, Mathematics, and Science MAPs. This proposal will then address how the test blueprints will change in 2010 compared to previous administrations of the test. We will discuss the possible impact on the future reliability of the MAP when a limited number of CR items are removed. We will then discuss the impact to the test structure. Finally, we will draw conclusions regarding comparability of the proposed 2010 MAP to previous MAP forms.

Removing Constructed Response Items

CTB is providing two options for removing CR items. Table 1 shows the decrease in the number of CR items compared to the original test blueprint for each grade/content area under Option D and Option E as compared to the 2009 MAP. This table also shows the number of selected response (SR) items that will be added in place of the removed CR items. Because the Communication Arts items are passage based, it is not possible to add more SR items to the Communication Arts MAPs. It is also not advisable to add an extra passage to the Communication Arts MAP because the additional passage would only be measured by two SR items.

Option D and Option E both remove CR items; however, Option E removes the performance event (PE) from the Mathematics MAP and portions of the PE from Science. It also removes an additional CR item from the Communication Arts MAP. Option D is the more desirable of these two options as it makes the fewest changes from previous versions of MAP.

Option D

Under Option D, between three and five CR items will be removed from each of the grade-level Communication Arts and Mathematics MAP forms as shown in Table 1. In Mathematics, these CR items will be replaced with three SR items, which will be drawn from Missouri's item pool.

In Science, CTB proposes that nine CR items be removed from the Grade 5 test and replaced with 20 SR items as shown in Table 1. In Grade 8 Science, 10 CR items will be removed from the Grade 8 test and replaced with 20 SR items. For Science, the additional SR items will be drawn from CTB's TerraNova item pool.

Option E

Option E is an extension of Option D in that the performance event (PE) is removed from the Mathematics test and portions of the PE are removed from the Science test. Under Option E, an additional CR item is removed from the grade-level Communication Arts tests.

For Grades 4 and 8 Mathematics, the PE will be replaced by two additional SR items under Option E. CTB Content Editors advised that the removal of the Mathematics PE will not have an adverse effect on the test construct as it measures the same knowledge, skills, and abilities measured by other items.

In Science, CTB proposes deleting the portion of the PE measuring 'new investigation' because these items measure knowledge, skills, and abilities found in earlier items within that same PE. The 2009 PE consisted of 10 items in Grade 5 and 12 items in Grade 8. Under Option E, it would be reduced to six to seven items in Grade 5 and seven to nine items in Grade 8.

Unlike Mathematics and Science, the PE is retained in the Communication Arts MAP. It was determined that the removal of the Writing Prompt from the Communication Arts test would affect the underlying test construct to such a degree that a new construct would be measured. Instead of removing the Writing Prompt, an additional CR will be removed from each grade-level Communication Arts MAP.

Table 1. Number of SR, CR, and PE by Grade and Content Area

Grade	Content Area	Item		2009	Option D	Option E
		Type				
3	Communication Arts	SR		47	47	47
		CR		7	4	3
		PE		1	1	1
	Mathematics	SR		53	56	56
		CR		7	4	4
4	Communication Arts	SR		47	47	47
		CR		7	4	3
	Mathematics	SR		55	58	60
		CR		9	4	4
		PE		1	1	0
5	Communication Arts	SR		47	47	47
		CR		7	4	3
	Mathematics	SR		55	58	58
		CR		7	4	4
	Science	SR		25	45	45
		CR		21	12	12
PE		10	9	7-8		
6	Communication Arts	SR		47	47	47
		CR		7	4	3
	Mathematics	SR		54	57	57
		CR		7	4	4
7	Communication Arts	SR		51	51	47
		CR		7	4	3
		PE		1	1	1
	Mathematics	SR		55	58	58
		CR		7	4	4
8	Communication Arts	SR		51	51	51
		CR		7	4	3
	Mathematics	SR		54	57	59
		CR		9	4	4
		PE		1	1	0
	Science	SR		25	45	45
		CR		24	14	14
PE		12	10	7-9		

Test Blueprints

There are two aspects to consider when examining the test blueprints: (1) the transition from the Version 1.0 Grade-Level Expectations (GLEs) to the Version 2.0 GLEs; and (2) the distributions of points under Option D and Option E compared to the Version 2.0 test blueprint.

Transition to Version 2.0 GLEs

In 2010, DESE plans to transition to the Version 2.0 GLEs. All previous MAP forms have been based on the Version 1.0 GLEs. In response to this transition, CTB has proposed test blueprints based on the Version 2.0 GLEs (note that these blueprints have not been approved by DESE). The Version 2.0 blueprints were compared to the Version 1.0 blueprints to understand the comparability of the two sets of blueprints. Table 2 shows the percentage of Communication Arts items measuring each of the content standards (i.e., test blueprints) based on the Version 1.0 and Version 2.0 GLEs and the difference between the two test blueprints. Table 3 shows the Version 1.0 and Version 2.0 blueprints for Mathematics as well as the difference between the two. Finally, Table 4 shows the same information for Science. As is evident in Table 2, only minimal changes were made to the Communication Arts test blueprint in Grade 3 and no changes were made to Communication Arts test blueprints in the other grades. Table 4 shows that no changes were made to the Science blueprints.

In comparing the blueprints for the Version 1.0 and Version 2.0 GLEs, it is evident that the largest changes were made to the Mathematics blueprints, reflecting the fact that the Mathematics GLEs were changed more than the GLEs in the other content areas. In Mathematics, changes were made to the test blueprints in all grades. The largest difference between the Version 1.0 and Version 2.0 blueprints was in Grades 4 and 7. In Grade 4, 35% of the items will measure the *Numbers and Operations* strand using the Version 2.0 blueprint compared to 25% under the Version 1.0 blueprint. In Grade 7, 30% of the items will measure the *Algebraic Relationships* strand using the Version 2.0 blueprint compared to 20% under the Version 1.0 blueprint. CTB generally allows 10% variance in the percentage of items measuring a strand on alternate forms of the test, and the Version 2.0 blueprints meet this rule of thumb.

Distribution of Points under Options D and E

Tables 5 through 7 show projected percentages of items measuring the MAP GLE strands in 2010 under two test design options and the Version 2.0 blueprints for Communication Arts, Mathematics, and Science, respectively. Looking across Tables 5 through 7, the distributions of points associated with both Option D and E are within +/- 10 percentage points of the Version 2.0 test blueprint.

Table 2. Percentage of Points Measuring Each Content Standard for the Test Blueprints based on the Version 1.0 and Version 2.0 GLEs and the Difference (DIF) between Version 2.0 minus Version 1.0 Percentages, Communication Arts*

Grade	Blueprint	GLE Strand		
		Reading (fiction and non-fiction)	Writing Formally	Writing Standard English
3	Version 1.0 GLE	70%	9%	22%
	Version 2.0 GLE	68%	10%	22%
	DIF (V 2.0 – V 1.0)	-2%	1%	0%
4	Version 1.0 GLE	82%	3%	15%
	Version 2.0 GLE	82%	3%	15%
	DIF (V 2.0 – V 1.0)	0%	0%	0%
5	Version 1.0 GLE	79%	3%	18%
	Version 2.0 GLE	79%	3%	18%
	DIF (V 2.0 – V 1.0)	0%	0%	0%
6	Version 1.0 GLE	78%	2%	20%
	Version 2.0 GLE	78%	2%	20%
	DIF (V 2.0 – V 1.0)	0%	0%	0%
7	Version 1.0 GLE	68%	10%	22%
	Version 2.0 GLE	68%	10%	22%
	DIF (V 2.0 – V 1.0)	0%	0%	0%
8	Version 1.0 GLE	77%	1%	22%
	Version 2.0 GLE	77%	1%	22%
	DIF (V 2.0 – V 1.0)	0%	0%	0%

*Numbers may not add to 100 due to rounding

Table 3. Percentage of Points Measuring Each Content Standard for the Test Blueprints based on the Version 1.0 and Version 2.0 GLEs and the Difference (DIF) between Version 2.0 minus Version 1.0 Percentages, Mathematics*

Gd.	Blueprint	GLE Strand				
		Algebraic Relationships	Data and Probability	Geometric and Spatial Relationships	Measurement	Number and Operations
3	Version 1.0 GLE	20%	10%	20%	15%	35%
	Version 2.0 GLE	22%	10%	18%	17%	33%
	DIF (V 2.0 – V 1.0)	2%	0%	-2%	2%	-2%
4	Version 1.0 GLE	20%	15%	20%	20%	25%
	Version 2.0 GLE	20%	10%	15%	20%	35%
	DIF (V 2.0 – V 1.0)	0%	-5%	-5%	0%	10%
5	Version 1.0 GLE	20%	20%	20%	20%	20%
	Version 2.0 GLE	22%	16%	17%	17%	28%
	DIF (V 2.0 – V 1.0)	2%	-4%	-3%	-3%	8%
6	Version 1.0 GLE	20%	20%	20%	20%	20%
	Version 2.0 GLE	19%	24%	14%	14%	29%
	DIF (V 2.0 – V 1.0)	-1%	4%	-6%	-6%	9%
7	Version 1.0 GLE	20%	20%	20%	20%	20%
	Version 2.0 GLE	30%	17%	18%	13%	22%
	DIF (V 2.0 – V 1.0)	10%	-3%	-2%	-7%	2%
8	Version 1.0 GLE	30%	20%	20%	15%	15%
	Version 2.0 GLE	35%	20%	25%	10%	10%
	DIF (V 2.0 – V 1.0)	5%	0%	5%	-5%	-5%

*Numbers may not add to 100 due to rounding

Table 4. Percentage of Points Measuring Each Content Standard for the Test Blueprints based on the Version 1.0 and Version 2.0 GLEs and the Difference (DIF) between Version 2.0 minus Version 1.0 Percentages, Science*

Gd.	Blueprint	GLE Strand							
		1	2	3	4	5	6	7	8
5	Version 1.0 GLE	13%	10%	10%	11%	12%	11%	25%	8%
	Version 2.0 GLE	13%	10%	10%	11%	12%	11%	25%	8%
	DIF (V 2.0 – V 1.0)	0%	0%	0%	0%	0%	0%	0%	0%
8	Version 1.0 GLE	13%	8%	12%	9%	13%	10%	28%	7%
	Version 2.0 GLE	13%	8%	12%	9%	13%	10%	28%	7%
	DIF (V 2.0 – V 1.0)	0%	0%	0%	0%	0%	0%	0%	0%

*Numbers may not add to 100 due to rounding

Table 5. Percentage of Points Measuring Each Strand under the Version 2.0 GLE Test Blueprint, Option D, and Option E as well as the difference between the Version 2.0 GLE Blueprint and Option D, and Version 2.0 GLE minus Option E, Communication Arts*

Grade	Blueprint	GLE Strand		
		Reading	Writing Formally	Writing Standard English
3	V2.0 GLE	68%	10%	22%
	Option D	64%	10%	26%
	V2.0-Option D	4%	0%	-4%
	Option E	63%	11%	26%
	V2.0-Option E	5%	-1%	-4%
4	V2.0 GLE	82%	3%	15%
	Option D	82%	4%	15%
	V2.0-Option D	0%	-1%	0%
	Option E	81%	4%	15%
	V2.0-Option E	1%	-1%	0%
5	V2.0 GLE	79%	3%	18%
	Option D	79%	4%	18%
	V2.0-Option D	0%	-1%	0%
	Option E	77%	4%	19%
	V2.0-Option E	2%	-1%	-1%
6	V2.0 GLE	78%	2%	20%
	Option D	78%	2%	20%
	V2.0-Option D	0%	0%	0%
	Option E	77%	2%	21%
	V2.0-Option E	1%	0%	-1%
7	V2.0 GLE	68%	10%	22%
	Option D	69%	10%	21%
	V2.0-Option D	-1%	0%	1%
	Option E	68%	10%	22%
	V2.0-Option E	0%	0%	0%
8	V2.0 GLE	77%	1%	22%
	Option D	76%	2%	22%
	V2.0-Option D	1%	-1%	0%
	Option E	75%	2%	23%
	V2.0-Option E	2%	-1%	-1%

*Numbers may not add to 100 due to rounding

Table 6. Percentage of Points Measuring Each Strand under the Version 2.0 GLE Test Blueprint, Option D, and Option E as well as the difference between the Version 2.0 GLE Blueprint and Option D, and Version 2.0 GLE minus Option E, Mathematics*

Gd	Blueprint	GLE Strand				
		Algebraic Relationships	Data and Probability	Geometric and Spatial Relationships	Measurement	Number and Operations
3	V2.0 GLE	22%	10%	18%	17%	33%
	Option D/E	22%	11%	17%	17%	33%
	V2.0-Opt D/E	0%	-1%	1%	0%	0%
4	V2.0 GLE	20%	10%	15%	20%	35%
	Option D	21%	11%	14%	20%	33%
	V2.0-Option D	-1%	-1%	1%	0%	2%
	Option E	22%	12%	12%	21%	34%
	V2.0 Option E	-2%	-2%	3%	-1%	1%
5	V2.0 GLE	22%	16%	17%	17%	28%
	Option D/E	21%	17%	17%	17%	29%
	V2.0-Opt D/E	1%	-1%	0%	0%	-1%
6	V2.0 GLE	19%	24%	14%	14%	29%
	Option D/E	20%	24%	14%	14%	29%
	V2.0-Opt D/E	-1%	0%	0%	0%	0%
7	V2.0 GLE	30%	17%	18%	13%	22%
	Option D/E	30%	16%	18%	13%	22%
	V2.0-Opt D/E	0%	1%	0%	0%	0%
8	V2.0 GLE	35%	20%	25%	10%	10%
	Option D	33%	17%	22%	10%	17%
	V2.0-Option D	2%	3%	3%	0%	-7%
	Option E	28%	18%	25%	10%	18%
	V2.0-Option E	7%	2%	0%	0%	-8%

*Numbers may not add to 100 due to rounding

Table 7. Percentage of Points Measuring Each Strand under the Version 2.0 GLE Test Blueprint, Option D, and Option E as well as the difference between the Version 2.0 GLE Blueprint and Option D, and Version 2.0 GLE minus Option E, Science*

Grade	Blueprint	GLE Strand							
		1	2	3	4	5	6	7	8
5	V2.0 GLE	13%	10%	10%	11%	12%	11%	25%	8%
	Option D	13%	10%	10%	11%	12%	11%	25%	8%
	V2.0-Option D	0%	0%	0%	0%	0%	0%	0%	0%
	Option E	12%	10%	10%	11%	12%	11%	25%	9%
	V2.0-Option E	1%	0%	0%	0%	0%	0%	0%	-1%
8	V2.0 GLE	13%	8%	12%	9%	13%	10%	28%	7%
	Option D	13%	8%	12%	9%	13%	10%	28%	7%
	V2.0-Option D	0%	0%	0%	0%	0%	0%	0%	0%
	Option E	12%	8%	12%	10%	12%	11%	29%	7%
	V2.0-Option E	1%	0%	0%	-1%	1%	-1%	-1%	0%

*Numbers may not add to 100 due to rounding

Test Reliability

The relationship between test reliability and test length is well known: longer tests are more reliable. In order to understand the impact that the removal of CR items will have on test reliability, CTB examined the impact the removal of CR items has on the reliability of the 2009 test. Table 8 shows the reliability of each grade/content area test comprised solely of multiple-choice items. These reliability coefficients are estimated from an early return sample comprised of approximately 20% of the Missouri data.

Table 8. Reliability of MAP using only SR items

Content Area	Grade	Number of SR Items	Current Alpha
Communication Arts	3	47	0.90
	4	47	0.91
	5	47	0.90
	6	47	0.89
	7	51	0.90
	8	51	0.88
Mathematics	3	53	0.91
	4	55	0.91
	5	55	0.91
	6	54	0.91
	7	55	0.91
	8	54	0.91
Science	5	22	0.75
	8	23	0.78

As can be seen in Table 8, all tests have acceptable reliability estimates except for the Grades 5 and 8 Science tests which are low.

Using the information in Table 8, we estimated the number of SR items that would need to be added to the Grades 5 and 8 Science tests to have a reliability of .85. To estimate the number of SR items needed, we used Spearman-Brown's prediction formula:

$$N = \frac{\rho_{xx'}^*(1 - \rho_{xx'})}{\rho_{xx'}(1 - \rho_{xx'}^*)}$$

where N indicates the proportion of items, given the current test length, that should be added to the overall test length, $\rho_{xx'}^*$ is the desired reliability, and $\rho_{xx'}$ is the observed reliability of the current test. Using this formula, it was determined that, at least, 20 SR items would need to be added to the Grade 5 Science test and 15 SR items would need to be added to the Grade 8 Science test in order to have a reliability of approximately .85. It is assumed that the additional CR items would further increase the test reliability of the Grade 5 and Grade 8 Science MAP.

Test Structure

In this section, we examine the test structure used in 2009 compared to the projected structure of the 2010 test. Tables 9 to 11 show the estimated testing times for each session of the MAP in 2009 and 2010 (under Options D and E) for Communication Arts, Mathematics, and Science, respectively. As observed in each table, the 2010 test will be much shorter than the 2009 test. This is due to the removal of operational CR items and the embedded field test.

The 2010 test will be structured into three or four sessions; however, many of these sessions will be shorter than the sessions that Missouri students took in 2009. While this may be noticeable to Missouri teachers who will now administer shorter tests, it is anticipated that the shorter testing times will be welcomed by Missouri educators.

Table 9. Approximate Testing Times in Minutes by Session for the 2009 MAP, Option D, and Option E and Difference in Total Testing Time, Communication Arts

Grade	Session	2009	Option D	Option E	Option D-2009	Option E-2009
3	1	45-55	45-55	40-50		
	2	60-90	60-90	60-90		
	3-Part 1 (5 min. break after)	28	30	30		
	3-Part 2	24	n/a	n/a		
	3-Part 3	6-11	n/a	n/a		
	4	50-65	25	25		
	Total Time	273	200	195	-73	-78
4, 5	1	45-55	45-55	40-50		
	2-Part 1 (5 min. break after)	28	30	30		
	2-Part 2	24	n/a	n/a		
	2-Part 3	11	n/a	n/a		
	3	50-65	25	25		
	Total Time	183	110	105	-73	-78
6, 8	1	45-55	45-55	40-50		
	2-Part 1 (5 min. break after)	27	30	30		
	2-Part 2	26	n/a	n/a		
	2-Part 3	11	n/a	n/a		
	3	50-65	25	25		
	Total Time	184	110	105	-74	-79
7	1	45-55	45-55	40-50		
	2	60-90	60-90	60-90		
	3-Part 1 (5 min. break after)	27	30	30		
	3-Part 2	26	n/a	n/a		
	3-Part 3	11	n/a	n/a		
	4	50-65	25	25		
	Total Time	274	200	195	-74	-79

Table 10. Approximate Testing Times in Minutes by Session for the 2009 MAP, Option D, and Option E and Difference in Total Testing Time, Mathematics

Grade	Session	2009	Option D	Option E	Option D-2009	Option E-2009
4, 8	1-Part 1 (5 min. break after)	40-55	25-35	30-35		
	1-Part 2	15-20	15-20	n/a		
	2-Part 1 (5 min. break after)	10	10	10		
	2-Part 2	30	30	30		
	2-Part 3	15	n/a	n/a		
	2-Part 4	10	n/a	n/a		
	3	50-70	25-35	25-35		
	Total Time	210	130	110	-80	-100
3, 5, 6, 7	1	40-55	25-35	25-35		
	2-Part 1 (5 min. break after)	10	10	10		
	2-Part 2	30	30	30		
	2-Part 3	15	n/a	n/a		
	2-Part 4	10	n/a	n/a		
	3	35-45	25-35	25-35		
	Total Time	165	110	110	-55	-55

Table 11. Approximate Testing Times in Minutes by Session for the 2009 MAP, Option D, and Option E and Difference in Total Testing Time, Science

Grade	Session	2009	Option D	Option E	Option D-2009	Option E-2009
5, 8	1	65-85	50-70	50-70		
	2-Part 1 (5 min. break after)	25	25	25		
	2-Part 2	40-55	40-55	40-55		
	2-Part 3	5	n/a	n/a		
	3	90-105	55-70	40-55		
	Total Time	275	220	205	-55	-70

Conclusions

The removal of a limited number of CR items on the MAP will not undermine the comparability of 2010 MAP scores to previous forms of the MAP. The proportion of items measuring each underlying content standard, while not identical, remains comparable across years. In addition, the MAP grade/content area tests will continue to be reliable instruments from which to generate student scores.

While the test will be noticeably shorter for students and teachers, both groups will continue to see that the MAP is divided into three or four test sessions. Furthermore, the removal of the

embedded field test explains, to a large degree, why the 2010 forms will be shorter than previous years.

In choosing between Option D and Option E, it is most desirable to use Option D. The removal of the PE on Math will be noticeable to Missouri teachers who are accustomed to their students spending 20 to 40 minutes on a single test item. While we do not think that the removal of this item would undermine the comparability of scores, it may cause the teachers to raise concerns about test validity.

It must be observed that CR items measure student knowledge, skills, and abilities in an intrinsically different manner than do SR items. While both item types require that students demonstrate knowledge, skills, and abilities, educators often perceive that CRs are “more valid” indicators of student knowledge than are SR items. Even so, the removal of a limited number of CR items will not undermine the psychometric characteristics (such as test reliability) of the test. The removal of the CR items is a question of construct validity, and it appears from the analyses presented here that the same construct would still be measured.

Based on the preliminary analyses done here, either Option D or Option E is acceptable. Both remove CR items from the test, and both should result in 2010 test scores that are comparable to previous forms of the MAP. Again, Option D remains more desirable because it leaves the PE items intact.

Once DESE has determined which option it wishes to use for the construction of the 2010 MAP, CTB will undertake further analyses to assess the impact that the removal of the CR items will have on the psychometric characteristics of the MAP as well as the percent of students in each MAP achievement level.