

# **Missouri**

Assessment Program  
*Grade-Level Assessments*

## **2011 Addendum** to the **2010 Technical Report**

Submitted to  
**Missouri Department of Elementary and Secondary Education**  
December 2011



Developed and published under contract with Missouri Department of Elementary and Secondary Education by CTB/McGraw-Hill LLC, 20 Ryan Ranch Road, Monterey, California 93940-5703. Copyright © 2011 by Missouri Department of Elementary and Secondary Education. All rights reserved. Only Missouri State educators and citizens may copy and/or download and print the document, located online at <http://www.dese.mo.gov/divimprove/asesstech/>. Any other use or reproduction of this document, in whole or in part, requires the prior written permission of Missouri Department of Elementary and Secondary Education.

## Table of Contents

CHAPTER 1: INTRODUCTION .....	1
CHAPTER 2: THE USES OF TEST SCORES .....	2
CHAPTER 3: TEST CONTENT DEVELOPMENT .....	3
3.7 Content and Process Standards .....	3
CHAPTER 4: TEST ADMINISTRATION .....	6
CHAPTER 5: CONSTRUCTED-RESPONSE SCORING .....	7
5.2 Inter-Rater Reliability .....	7
CHAPTER 6: OPERATIONAL DATA ANALYSES .....	11
6.2 Calibration Sample.....	11
6.3 Classical Item Statistics .....	11
6.2.2 Item-Level Statistics .....	11
6.4 Item Response Theory .....	12
6.3.1 Model Fit.....	12
CHAPTER 7: TEST RESULTS.....	17
7.1 Student Participation.....	17
7.2 Current Administration Data.....	17
7.3 Cross-year, Cross-sectional Comparisons .....	17
CHAPTER 8: ACHIEVEMENT-LEVEL SETTING .....	27
CHAPTER 9: EVIDENCE OF CONSTRUCT-RELATED VALIDITY .....	28
9.2 Reliability.....	28
9.2.1 Test Reliability.....	28
9.2.2 Standard Error of Measurement.....	28
9.2.3 Conditional Standard Error of Measurement.....	28
9.4 Analyses by Content Standard.....	28
9.4.1 Reliability of Content Standards.....	28
9.4.2 Correlations Among Content Standard Subscores.....	28
9.4.3 Standard Error of Measurement of Content Standards .....	29
CHAPTER 10: FAIRNESS.....	36
10.3 Evaluating Bias Through Impact Analysis .....	36
10.3.1 Reliability.....	36
10.3.2 Effect Size.....	36
Appendix A.....	A-1

## Table of Tables

Table 1. 1 Chapters, Sections, and Figures/Tables Updated for 2011 Addendum .....	1
Table 3. 4: MAP 2011 Content Standard Item/Point Distributions, Communication Arts	3
Table 3. 5: MAP 2011 GLE Strand Item/Point Distributions, Mathematics .....	3
Table 3. 6: MAP 2011 GLE Strand Item/Point Distributions, Science .....	4
Table 3. 7: MAP 2011 Number of Items/Points Measuring Process Strands, Communication Arts .....	4
Table 3. 8: MAP 2011 Number of Items/Points Measuring Process Strands, Mathematics .....	5
Table 3. 9: MAP 2011 Number of Items/Points Measuring Process Strands, Science .....	5
Table 5. 1: Inter-Rater Reliability, Communication Arts .....	8
Table 5. 2: Inter-Rater Reliability, Mathematics .....	9
Table 5. 3: Inter-Rater Reliability, Science .....	10
Table 6. 5: MAP Means, Standard Deviations for Raw Scores, <i>p</i> -Values, Item-Total Correlation ( $R_{it}$ ): Communication Arts 2011 .....	14
Table 6. 6: MAP Means, Standard Deviations for Raw Scores, <i>p</i> -Values, Item-Total Correlation ( $R_{it}$ ): Mathematics 2011 .....	14
Table 6. 7: MAP Means, Standard Deviations for Raw Scores, <i>p</i> -Values, Item-Total Correlation ( $R_{it}$ ): Science 2011 .....	14
Table 6. 8: Item Statistics: Grade 3.....	15
Table 6. 14: Item Fit Statistics for Misfitting Items.....	16
Table 7. 1: Participation Rates: All Students .....	19
Table 7. 2: Participation Rates: Males .....	19
Table 7. 3: Participation Rates: Females.....	19
Table 7. 4: Participation Rates: White .....	20
Table 7. 5: Participation Rates: Black.....	20
Table 7. 6: Participation Rates: Hispanic.....	20
Table 7. 7: Participation Rates: Asian/Pacific Islander .....	21
Table 7. 8: Participation Rates: Native American/Alaskan .....	21
Table 7. 9: Participation Rates: Students Receiving Accommodations.....	21
Table 7. 10: Summary Statistics for Communication Arts .....	22
Table 7. 11: Summary Statistics for Mathematics .....	22
Table 7. 12: Summary Statistics for Science .....	22
Table 7. 13: Comparison of State-Level Means, 2006 Through 2011 Census Data .....	23
Table 7. 14: Comparison of Percentage of Students in Each Achievement Level, Communication Arts 2006 Through 2011 Census Data .....	24
Table 7. 15: Comparison of Percentage of Students in Each Achievement Level, Mathematics 2006 Through 2011 Census Data .....	25
Table 7. 16: Comparison of Percentage of Students in Each Achievement Level, Science 2008 Through 2011 Census Data.....	26

Table 9. 1: Reliability in Communication Arts.....	30
Table 9. 2: Reliability in Mathematics.....	30
Table 9. 3: Reliability in Science.....	30
Table 9. 4: SEM by Subgroup.....	31
Table 9. 5: Conditional Standard Error of Measurement at the Basic, Proficient, & Advanced Cut Scores.....	32
Table 9. 11: Reliability (Diagonal) of Each Content Standard, Uncorrected Correlation Coefficient (below Diagonal), and Corrected Correlation Coefficient (above Diagonal) Among Content Standards: Communication Arts .....	33
Table 9. 12: Reliability (Diagonal) of Each Content Standard, Uncorrected Correlation Coefficient (below Diagonal), and Corrected Correlation Coefficient (above Diagonal) Among Content Standards: Mathematics .....	33
Table 9. 13: Reliability (Diagonal) of Each Content Standard, Uncorrected Correlation Coefficient (below Diagonal), and Corrected Correlation Coefficient (above Diagonal) Among Content Standards: Science.....	34
Table 9. 14: Mean, Standard Deviation, and SEM of Communication Arts Content Standards.....	35
Table 9. 15: Mean, Standard Deviation, and SEM of Mathematics Content Standards...	35
Table 9. 16: Mean, Standard Deviation, and SEM of Science Content Standards .....	35
Table 10. 4: Impact Analysis, Grade 3 .....	38
Table 10. 5: Impact Analysis, Grade 4 .....	38
Table 10. 6: Impact Analysis, Grade 5 .....	39
Table 10. 7: Impact Analysis, Grade 6 .....	40
Table 10. 8: Impact Analysis, Grade 7 .....	40
Table 10. 9: Impact Analysis, Grade 8 .....	41

## Table of Figures

Figure 6. 1: Item characteristic curve for Grade 3 Communication Arts, Session 3 Item 34.....	16
--	----

## CHAPTER 1: INTRODUCTION

The Missouri Assessment Program (MAP) is designed to measure students’ knowledge of Communication Arts, Mathematics, and Science. The 2011 MAP marked the sixth administration of grade-level Communication Arts and Mathematics MAP in Missouri. It was the fourth administration of the grade-span Science MAP at Grades 5 and 8. This addendum updates select results from the *Missouri Assessment Program Grade-Level Assessments Technical Report 2010*, and it should be used in conjunction with the *Missouri Assessment Program Grade-Level Assessments Technical Report 2010*.

For budgetary reasons, the Missouri Department of Elementary and Secondary Education (DESE) re-administered a previous form of the grade-level MAPs<sup>1</sup>. In addition, DESE did not administer the performance events (PEs) and writing prompts that had been part of previous administrations. Appendix A contains a special study conducted by CTB/McGraw-Hill that addressed the impact of the removal of the PEs/writing prompts.

Because the 2011 administration was based on a previous form, existing scoring tables were used for all tests except the Grade 3 Communication Arts test. The item parameters for Grade 3 Communication Arts were updated to include an item that had been suppressed in its previous administration. In this addendum, only classical and item response theory results will be presented for Grade 3 Communication Arts (see Chapter 6). State-level MAP results (see Chapters 7 and 10) are presented for all grades/content areas. Statistical analyses related to the evidence for construct validity were updated for those grades/content areas where the PEs/writing prompts were removed (see Chapters 3 and 9).

This chapter lists the figure and tables from 2010 that are updated in this addendum. For each figure or table that was updated, the associated text was also updated and included in this addendum. Table 1.1 lists the sections, tables, and figures updated by chapter.

**Table 1. 1 Chapters, Sections, and Figures/Tables Updated for 2011 Addendum**

Chapter	Section	Figure or Table
3: Test Content Development	3.7 Content and Process Standards	Tables 3.4–3.9
5: Constructed-Response Scoring	5.2 Inter-Rater Reliability	Tables 5.1–5.3
6: Operational Data Analysis	6.1 Calibration sample 6.2 Classical Item Statistics 6.3 Item Response Theory	Tables 6.5–6.8 Table 6.14 Figure 6.1
7: Test Results	7.1 Student Participation 7.2 Current Administration Data 7.3 Cross-year, Cross-sectional Comparisons	Tables 7.1–7.9 Tables 7.10–7.12 Tables 7.13–7.16
9: Evidence of Construct-Related Validity	9.2 Reliability 9.4 Analyses by Content Standard	Tables 9.1–9.5 Tables 9.11–9.16
10: Fairness	10.3 Evaluating Bias Through Impact Analyses	Tables 10.4–10.9

<sup>1</sup> Repeated use of the same test form is not recommended.

## **CHAPTER 2: THE USES OF TEST SCORES**

No updates to this chapter. Please see the 2010 MAP Technical Report.

## CHAPTER 3: TEST CONTENT DEVELOPMENT

### 3.7 Content and Process Standards

Table 3.4 provides the distribution of items and points on the 2011 MAP by Content Standard for Communication Arts. Tables 3.5 and 3.6 provide the same distribution by GLE strand for Mathematics and Science, respectively. (GLE strands are the reporting categories for these content domains; however, GLEs remain linked directly to the Content Standards.) Lastly, Tables 3.7 through 3.9 show the distribution of items and points by Process Strand for Communication Arts, Mathematics, and Science, respectively. Only those grades/content areas where performance events were removed following the 2010 administration are presented in Tables 3.4 through 3.9.

**Table 3. 4: MAP 2011 Content Standard Item/Point Distributions, Communication Arts**

Grade	Content Standard	TN NRT Items	SR Items	CR/PE Items	Total Items	SR Points	CR/PE Points	Total Points	% of Total Points
3	reading fiction/poetry/drama	23			23	23		23	37%
	reading nonfiction	7	7	4	18	14	8	22	35%
	speaking/writing standard English		15		15	15		15	24%
	writing formally & informally			2	2		2	2	3%
	Combined Reading from Standards 2 & 3	30	7	4	41	37	8	45	73%
	Total	30	22	6	58	52	10	62	100%
7	reading fiction/poetry/drama	13	7	4	24	20	8	28	42%
	reading nonfiction	20			20	20		20	30%
	speaking/writing standard English		16		16	16		16	24%
	writing formally & informally			2	2		2	2	3%
	Combined Reading from Standards 2 & 3	33	7	4	44	40	8	48	73%
	Total	33	23	6	62	56	10	66	100%

**Table 3. 5: MAP 2011 GLE Strand Item/Point Distributions, Mathematics**

Grade	GLE Strand	TN NRT Items	SR Items	CR/PE Items	Total Items	SR Points	CR/PE Points	Total Points	% of Total Points
4	Algebraic Relationships	5	7	1	13	12	2	14	22%
	Data and Probability	4	1	1	6	5	2	7	11%
	Geometric and Spatial Relationships	2	6	1	9	8	2	10	15%
	Measurement	3	7		10	10		10	15%
	Number and Operations	12	10	1	23	22	2	24	37%
	Total	26	31	4	61	57	8	65	100%
8	Algebraic Relationships	5	12	1	18	17	2	19	30%
	Data and Probability	4	3	1	8	7	2	9	14%
	Geometric and Spatial Relationships	4	10	1	15	14	2	16	25%
	Measurement	2	2	1	5	4	2	6	10%
	Number and Operations	13			13	13		13	21%
	Total	28	27	4	59	55	8	63	100%

**Table 3. 6: MAP 2011 GLE Strand Item/Point Distributions, Science**

Grade	GLE Strand	TN NRT Items	SR Items	CR/PE Items	Total Items	SR Points	CR/PE Points	Total Points	% of Total Points
5	characteristics of living organisms	2	4	1	7	6	2	8	12%
	Earth's processes	2	3	2	7	5	4	9	13%
	force and motion		2	3	5	2	6	8	12%
	interactions of organisms	3	2	2	7	5	4	9	13%
	matter and energy	6	1	2	9	7	4	11	16%
	scientific inquiry	6	2		8	8		8	12%
	technology and the environment	2	3	1	6	5	2	7	10%
	the universe	1	3	2	6	4	4	8	12%
	Total	22	20	13	55	42	26	68	100%
8	characteristics of living organisms	3		3	6	3	6	9	13%
	Earth's processes	5	1	2	8	6	4	10	14%
	force and motion	3	2	1	6	5	2	7	10%
	interactions of organisms	2	3	1	6	5	2	7	10%
	matter and energy	2	4	3	9	6	6	12	17%
	scientific inquiry	7	3		10	10		10	14%
	technology and the environment	1	3	2	6	4	4	8	11%
	the universe		4	2	6	4	4	8	11%
	Total	23	20	14	57	43	28	71	100%

**Table 3. 7: MAP 2011 Number of Items/Points Measuring Process Strands, Communication Arts**

Grade Level	Process Standard	NRT Items	SR Items	CR Items	Total Items	SR Points	CR Pts	Total Points
3	1.4		1		1	1		1
	1.5	9	2		11	11		11
	1.6	15	2	4	21	17	8	25
	2.1			2	2		2	2
	2.2		15		15	15		15
	2.4	1			1	1		1
	3.5	5	2		7	7		7
7	1.5	6	1		7	7		7
	1.6	21	2	1	24	23	2	25
	2.1		1	2	3	1	2	3
	2.2		15		15	15		15
	2.4	1	2		3	3		3
	3.1	1			1	1		1
	3.5	4	2	3	9	6	6	12

**Table 3. 8: MAP 2011 Number of Items/Points Measuring Process Strands, Mathematics**

Grade Level	Process Standard	NRT Items	SR Items	CR Items	Total Items	SR Points	CR Pts	Total Points
4	1.1		1		1	1		1
	1.2	1			1	1		1
	1.5		2		2	2		2
	1.6	5	12	2	19	17	4	21
	1.8			1	1		2	2
	1.10	11	5		16	16		16
	2.1			1	1		2	2
	3.1		4		4	4		4
	3.2	9	2		11	11		11
	3.3		4		4	4		4
	3.6		2	1	3	2	2	4
8	1.5			1	1		2	2
	1.6	3	8	2	13	11	4	15
	1.8	1	3		4	4		4
	1.10	4	1	1	6	5	2	7
	3.1	6	2		8	8		8
	3.2	4	4		8	8		8
	3.3	7	7		14	14		14
	3.5	1	1		2	2		2
	3.6	1	1		2	2		2
	3.8	1			1	1		1

**Table 3. 9: MAP 2011 Number of Items/Points Measuring Process Strands, Science**

Grade Level	Process Standard	NRT Items	SR Items	CR Items	Total Items	SR Points	CR Pts	Total Points
5	1.3	2			2	2		2
	1.5	5	1	1	7	6	2	8
	1.6	3	4	7	14	7	14	21
	1.10	12	15	5	32	27	10	37
8	1.3	1			1	1		1
	1.5	3			3	3		3
	1.6	3	4	6	13	7	12	19
	1.8		1		1	1		1
	1.10	16	15	7	38	31	14	45
	3.8			1	1		2	2

## **CHAPTER 4: TEST ADMINISTRATION**

No updates to this chapter. Please see the 2010 MAP Technical Report.

## CHAPTER 5: CONSTRUCTED-RESPONSE SCORING

### 5.2 Inter-Rater Reliability

Approximately 5% of the papers in Communication Arts, Mathematics, and Science were scored independently by a second reader. The statistics for the inter-rater reliability were calculated for all items at all grades. To determine the reliability of scoring, the percentage of perfect agreement and adjacent agreement between the two readers was examined.

For each item, a weighted Kappa was calculated to reflect the level of improvement beyond the chance level in the consistency of scoring. These weighted Kappa values are presented in Tables 5.1 to 5.3. To aid in the interpretation of Kappa, the following cutoffs have been suggested (Landis & Koch, 1977; Altman, 1991):

<u>Kappa Value</u>	<u>Strength of Agreement</u>
0	None
<0.20	Poor
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Good
0.81 – 1.00	Very Good

All Communication Arts, Mathematics, and Science items show good inter-rater agreement. As shown in Table 5.1, raters demonstrated at least 93% perfect and adjacent agreement for all Communication Arts items. Except for two items, the strength of the inter-rater agreement may be interpreted as good or very good as indicated by the weighted Kappa values. One Grade 6 item (Session 1, Item 5) and one Grade 8 item (Session 1, Item 4) had weighted Kappa values that indicate only moderate agreement between the raters.

As shown in Table 5.2, raters demonstrated at or above 99% perfect and adjacent agreement for all Mathematics items. The weighted Kappa values indicate that there was very good inter-rater agreement for all Mathematics items.

As shown in Table 5.3, raters demonstrated at or above 97% perfect and adjacent agreement for all Science items. The weighted Kappa statistic indicates good or very good inter-rater agreement for all Science items.

**Table 5. 1: Inter-Rater Reliability, Communication Arts**

<b>Grade</b>	<b>Session</b>	<b>Item #</b>	<b># Points</b>	<b>% Perfect</b>	<b>% Adjacent</b>	<b>% Perfect &amp; Adjacent*</b>	<b>Weighted Kappa</b>
<b>3</b>	1	3	2	89	10	99	0.85
	1	4	2	72	26	98	0.70
	1	5	2	83	15	98	0.83
	1	6A	2	87	13	100	0.87
	1	6B	1	99	1	100	0.95
	1	6C	1	98	2	100	0.77
<b>4</b>	1	3	2	82	16	98	0.85
	1	4	2	86	13	99	0.86
	1	5	2	90	8	98	0.88
	1	6A	2	81	17	98	0.78
	1	6B	1	98	2	100	0.88
	1	6C	1	98	2	100	0.71
<b>5</b>	1	3	2	70	28	98	0.69
	1	4A	2	73	25	98	0.71
	1	4B	2	96	3	100	0.81
	1	5	2	79	17	96	0.77
	1	6	2	79	20	99	0.73
<b>6</b>	1	3	2	91	9	100	0.78
	1	4	2	78	21	99	0.71
	1	5	2	64	30	93	0.55
	1	6A	2	80	19	99	0.84
	1	6B	1	96	4	100	0.84
<b>7</b>	1	3	2	75	14	90	0.61
	1	4	2	84	15	99	0.67
	1	5	2	75	23	98	0.75
	1	6A	2	89	11	99	0.83
	1	6B	1	98	2	100	0.92
	1	6C	1	97	3	100	0.76
<b>8</b>	1	3	2	77	22	99	0.81
	1	4	2	65	28	93	0.59
	1	5	2	69	28	97	0.66
	1	6A	2	68	30	98	0.70
	1	6B	1	96	3	100	0.85
	1	6C	1	97	3	100	0.84

\* The percent perfect & adjacent may not add up to 100 for 1-point items due to the percent discrepant. The percent discrepant includes the cases where one rater assigned a score and the other rater assigned a condition code. With 2- or more point items, it also refers to the cases where the assigned score varied by more than 1 point.

**Table 5. 2: Inter-Rater Reliability, Mathematics**

<b>Grade</b>	<b>Session</b>	<b>Item #</b>	<b># Points</b>	<b>% Perfect</b>	<b>% Adjacent</b>	<b>% Perfect &amp; Adjacent*</b>	<b>Weighted Kappa</b>
<b>3</b>	3	1	2	95	5	100	0.91
	3	2	2	95	4	99	0.96
	3	3	2	93	7	100	0.95
	3	4	2	96	4	99	0.97
<b>4</b>	3	1	2	96	4	100	0.95
	3	2	2	93	7	100	0.95
	3	3	2	93	7	100	0.90
	3	4	2	92	8	100	0.93
<b>5</b>	3	1	2	94	6	99	0.96
	3	2	2	93	7	99	0.95
	3	3	2	97	3	100	0.98
	3	4	2	94	6	100	0.94
<b>6</b>	3	1	2	92	8	100	0.95
	3	2	2	96	4	100	0.97
	3	3	2	90	10	100	0.88
	3	4	2	95	5	100	0.95
<b>7</b>	3	1	2	97	2	100	0.97
	3	2	2	98	2	100	0.98
	3	3	2	96	3	99	0.96
	3	4	2	95	5	100	0.94
<b>8</b>	3	1	2	92	7	99	0.95
	3	2	2	97	3	100	0.98
	3	3	2	97	3	100	0.98
	3	4	2	78	22	100	0.81

\* The percent perfect & adjacent may not add up to 100 for 1-point items due to the percent discrepant. The percent discrepant includes the cases where one rater assigned a score and the other rater assigned a condition code. With 2- or more point items, it also refers to the cases where the assigned score varied by more than 1 point.

**Table 5. 3: Inter-Rater Reliability, Science**

<b>Grade</b>	<b>Session</b>	<b>Item #</b>	<b># Points</b>	<b>% Perfect</b>	<b>% Adjacent</b>	<b>% Perfect &amp; Adjacent*</b>	<b>Weighted Kappa</b>
<b>5</b>	1	1	2	97	3	100	0.97
	1	2	2	80	20	100	0.83
	1	3	2	92	8	100	0.93
	1	4	2	82	15	97	0.83
	1	5	2	83	15	98	0.84
	1	6	2	85	14	100	0.88
	1	7	2	86	14	99	0.87
	1	8	2	77	21	98	0.80
	1	9	2	90	10	100	0.91
	1	10	2	87	12	100	0.88
	1	11	2	95	5	100	0.96
	1	12	2	85	14	99	0.82
	1	13	2	96	4	100	0.96
<b>8</b>	1	1	2	96	4	100	0.97
	1	2	2	85	14	99	0.87
	1	3	2	93	7	100	0.94
	1	4	2	86	12	98	0.88
	1	5	2	98	2	100	0.98
	1	6	2	88	12	100	0.88
	1	7	2	82	17	99	0.74
	1	8	2	78	22	99	0.73
	1	9	2	92	8	100	0.89
	1	10	2	90	10	99	0.91
	1	11	2	86	13	99	0.83
	1	12	2	94	6	100	0.91
	1	13	2	93	7	100	0.87
	1	14	2	92	7	100	0.75

\* The percent perfect & adjacent may not add up to 100 for 1-point items due to the percent discrepant. The percent discrepant includes the cases where one rater assigned a score and the other rater assigned a condition code. With 2- or more point items, it also refers to the cases where the assigned score varied by more than 1 point.

## CHAPTER 6: OPERATIONAL DATA ANALYSES

### 6.2 Calibration Sample

In this section, we describe the calibration sample in adherence to Standard 1.5 of the AERA, APA, & NCME (1999) Standards. Standard 1.5 states:

The composition of any sample of examinees from which validity evidence is obtained should be described in as much detail as is practical, including major relevant sociodemographic and developmental characteristics.

In 2011, the grade-level calibration samples were comprised of at least 98% of the total student population for that grade.

### 6.3 Classical Item Statistics

In this section, we present summary test statistics for each grade/content area of the MAP. This is followed by item-level statistics for Grade 3 Communication Arts.

Tables 6.5 through 6.7 present the number of items and score points on each test, as well as the mean and standard deviation of the raw scores,  $p$ -values, and item-total correlations (also known as item discrimination values) for each grade level of Communication Arts, Mathematics, and Science, respectively.

The mean  $p$ -value is the average of all item  $p$ -values of a specific grade/content area. The mean item-total correlation ( $R_{it}$ ) is the average of all item biserial correlations of a specific grade/content area. The  $p$ -value and item-total correlation are explained in the next section.

#### 6.2.2 Item-Level Statistics

Table 6.8 presents the item statistics for each item for Grade 3 Communication Arts. For all other grades/content areas, please see the *Missouri Assessment Program Grade-Level Assessments Technical Report 2010*. The tables include test session, item booklet number and part (if applicable),  $p$ -values, item-total correlations ( $R_{it}$ ), and omit rates for each item by grade/content area.

*p-value*: The  $p$ -value is a measure of item difficulty. For a selected-response item, the  $p$ -value is calculated from the number of students who correctly responded to an item divided by the total number of students who attempted the item. The value is reported as a proportion. For a constructed-response item, the  $p$ -value is calculated from the average score for the item divided by the maximum points possible and is also reported as a proportion.

In terms of  $p$ -values, test scores tend to be more precise when their average  $p$ -values are in the mid 0.50s to low 0.70s. However, in building a criterion-referenced test, it is important to select items on the basis of content rather than on purely statistical criteria. As demonstrated in Table 6.5, the average  $p$ -values associated with the Communication Arts MAP range from .70 (Grade 8) to .78 (Grade 4). The average  $p$ -values associated with the Mathematics MAP (Table 6.6) range from .59 (Grade 8) to .81 (Grade 3). The average  $p$ -values associated with the Science MAP (Table 6.7) range from .62 (Grade 8) to .66 (Grade 5).

It is important that one examines the range of  $p$ -values and not just the average  $p$ -value to determine whether a test measures well. It is desirable for the test to measure well throughout the range of skills present at a given grade. That is, it is important that the items measure the performance of both low-scoring and high-scoring students, as well as students in the center of the distribution. Having a range of  $p$ -values also helps to prevent floor and/or ceiling effects so that the test does not have large numbers of students at the minimum or maximum possible scores. The Grade 3 Communication Arts MAP has items with  $p$ -values ranging from the low 0.40s to the 0.90s (see Table 6.8).

*Item-Total Correlations:* An item-total correlation is the correlation between an item and the total test score, where the item score is included in the total score. It indicates how well an item differentiates between low- and high-achieving students. In general, items with correlations below .20 are said to be poorly discriminating. The majority of the items in the MAP had item-test correlations above this threshold. Any item with an item-total correlation below the .20 threshold was further analyzed to assure that the item was correctly keyed.

*Omit Rates:* The omit rate for each item indicates the percentage of students who did not answer the item. Omit rates can be used to examine possible speededness issues on tests. A test may be speeded if students do not have adequate time to answer all questions on the test. As a rule of thumb, an item is said to have a high omit rate if more than 5% of students failed to respond to the item. The results in Table 6.8 show that no Grade 3 Communication Arts items had high omit rates.

## 6.4 Item Response Theory

### 6.3.1 Model Fit

One Grade 3 Communication Arts operational item was flagged for poor fit. Table 6.14 shows the chi-square statistic and the Z-statistic for the flagged item. The average percent across ten cells of observed percentage correct and predicted percentage correct is also provided. The difference between the observed and predicted percentages provides an indication of how well the modeled response curves reflect the empirical curves.

The flagged item was examined more closely by studying its item characteristic curve (ICC). The ICC models the relationship between the examinees' performance on an item and the examinees' underlying ability. In almost all cases for which model misfit occurs,

relatively few students occupy these scale score ranges, which are at the lower and upper tails of the distribution. Poor fit may occur in one region of the underlying ability distribution when there are relatively few students at that particular point in the distribution. The model tends to show good model-data fit for the flagged items in the middle of the theta distribution where the majority of students perform.

Figure 6.1 shows the ICCs for the misfitting MAP item. The smooth line in this figure represents the predicted relationship between examinee performance on the item and examinee ability, and the jagged line represents the observed relationship. Large differences between the two lines indicate poor fit. This figure also shows the distribution of theta scores, so that the fit between observed and predicted performance at different ability levels can be interpreted in light of the overall distribution of examinees. Figure 6.1 presents the ICC for Session 3, Item 34 (SR item), on the Grade 3 Communication Arts test. As shown, there is poor fit at the lower end of the ability range.

**Table 6. 5: MAP Means, Standard Deviations for Raw Scores,  $p$ -Values, Item-Total Correlation ( $R_{it}$ ): Communication Arts 2011**

Grade	Total Items	Total Points	Mean Raw Score (SD)	Mean $p$ -value (SD)	Mean $R_{it}$ (SD)
3	58	62	46.75 (9.95)	0.76 (0.14)	0.37 (0.08)
4	58	62	47.58 (10.46)	0.78 (0.15)	0.41 (0.10)
5	56	61	43.96 (9.99)	0.74 (0.16)	0.38 (0.09)
6	56	60	43.18 (10.17)	0.72 (0.14)	0.37 (0.09)
7	62	66	47.18 (10.29)	0.72 (0.17)	0.34 (0.11)
8	60	64	44.10 (11.03)	0.70 (0.16)	0.37 (0.09)

**Table 6. 6: MAP Means, Standard Deviations for Raw Scores,  $p$ -Values, Item-Total Correlation ( $R_{it}$ ): Mathematics 2011**

Grade	Total Items	Total Points	Mean Raw Score (SD)	Mean $p$ -value (SD)	Mean $R_{it}$ (SD)
3	55	59	46.41 (9.53)	0.81 (0.14)	0.40 (0.08)
4	61	65	48.33 (11.01)	0.76 (0.13)	0.38 (0.09)
5	58	62	45.00 (11.10)	0.73 (0.14)	0.38 (0.11)
6	58	62	43.75 (11.54)	0.72 (0.15)	0.40 (0.08)
7	61	65	41.73 (12.09)	0.66 (0.17)	0.39 (0.09)
8	59	63	36.65 (12.71)	0.59 (0.17)	0.40 (0.11)

**Table 6. 7: MAP Means, Standard Deviations for Raw Scores,  $p$ -Values, Item-Total Correlation ( $R_{it}$ ): Science 2011**

Grade	Total Items	Total Points	Mean Raw Score (SD)	Mean $p$ -value (SD)	Mean $R_{it}$ (SD)
5	55	68	41.95 (11.35)	0.66 (0.19)	0.35 (0.10)
8	57	71	39.51 (11.79)	0.62 (0.23)	0.37 (0.10)

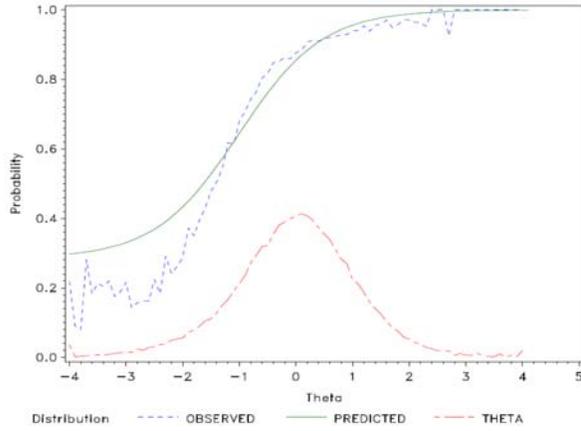
**Table 6. 8: Item Statistics: Grade 3**

Communication Arts									
Session	Item	<i>p</i> -Value	R <sub>it</sub>	Omit Rate	Session	Item	<i>p</i> -Value	R <sub>it</sub>	Omit Rate
1	1	0.73	0.24	0.05	3	22	0.73	0.45	0.28
1	2	0.73	0.35	0.07	3	23	0.63	0.39	0.51
1	3	0.78	0.45	0.27	3	24	0.40	0.21	0.22
1	4	0.68	0.48	0.52	3	25	0.88	0.45	0.35
1	5	0.66	0.46	0.42	3	26	0.86	0.47	0.54
1	6A	0.66	0.34	0.49	3	27	0.64	0.47	0.64
1	6B	0.85	0.42	0.49	3	28	0.75	0.32	0.82
1	6C	0.96	0.34	0.49	3	29	0.97	0.36	0.22
1	7	0.84	0.29	0.38	3	30	0.76	0.45	0.29
1	8	0.83	0.43	0.43	3	31	0.86	0.45	0.38
1	9	0.62	0.32	0.40	3	32	0.60	0.29	0.69
1	10	0.70	0.26	0.40	3	33	0.88	0.40	0.36
1	11	0.46	0.20	0.40	3	34	0.81	0.43	0.55
1	12	0.42	0.20	0.40	3	35	0.45	0.24	0.74
3	1	0.98	0.32	0.05	3	36	0.64	0.36	0.99
3	2	0.95	0.44	0.12	3	37	0.93	0.37	0.66
3	3	0.90	0.34	0.34	3	38	0.89	0.50	0.86
3	4	0.81	0.43	0.38	3	39	0.77	0.37	0.97
3	5	0.92	0.40	0.60	4	1	0.85	0.42	0.10
3	6	0.90	0.18	0.12	4	2	0.77	0.40	0.20
3	7	0.67	0.19	0.15	4	3	0.60	0.40	0.89
3	8	0.89	0.43	0.36	4	4	0.79	0.46	0.22
3	9	0.65	0.38	0.69	4	5	0.57	0.42	0.41
3	10	0.91	0.46	0.67					
3	11	0.86	0.41	2.26					
3	12	0.70	0.29	2.56					
3	13	0.82	0.44	0.16					
3	14	0.92	0.44	0.24					
3	15	0.81	0.50	0.44					
3	16	0.61	0.36	0.67					
3	17	0.80	0.34	0.42					
3	18	0.69	0.40	0.61					
3	19	0.86	0.40	0.86					
3	20	0.81	0.32	0.15					
3	21	0.78	0.34	0.23					

**Table 6. 14: Item Fit Statistics for Misfitting Items**

Content	Grade	Session	Item	Chi-Square	DF	Total N	Z	Observed	Predicted	Obs-Pred
CA	3	3	34	814.74	7	65033	215.88	0.81	0.81	-0.01

**Figure 6. 1: Item characteristic curve for Grade 3 Communication Arts, Session 3 Item 34**



## CHAPTER 7: TEST RESULTS

### 7.1 Student Participation

The following are subgroups reported during the administration of the MAP (other demographic information is collected separately and merged into the MAP data after CTB/McGraw-Hill sends DESE the General Research File):

- Gender: Female and Male
- Race and Ethnicity: White, Black, Hispanic, Asian/Pacific Islander, and Native American/Alaskan
- Accommodations: students receiving testing accommodations

For the purposes of this report, participation rate is defined as the percentage of students who received a valid scale score given the total number of students who received a test book. These participation rates are summarized in Tables 7.1 through 7.9. The tables show both the number of students classified as accountable and the percentage of students classified as reportable. Reportable students include all students with a valid scale score. Accountable students include all students for whom a test book was submitted. These include students who should have received a MAP scale score, but did not take the test and could not be assigned a scale score.

### 7.2 Current Administration Data

The Communication Arts and Mathematics MAP assessments were administered to students in Grades 3 through 8. The Science MAP assessments were administered to students in Grades 5 and 8.

Tables 7.10 through 7.12 provide a summary of the scale scores based on the state population for the 2011 administration of the MAP. In compliance with AERA, APA, & NCME (1999) Standard 13.19, these tables present the number of students, mean and standard deviation of scale scores, and scale scores at specific percentile points. Standard 13.19 states:

In educational settings, when average or summary scores for groups of students are reported, they should be supplemented with additional information about the sample size and shape or dispersion of score distributions.

### 7.3 Cross-year, Cross-sectional Comparisons

It is often desirable to examine the scores of students across time. The data in this section compare student performance on the MAP using census data from 2006 through 2011. It should be noted that beginning in 2008, invalidated students were assigned to the LOSS

and to the Below Basic achievement level. Prior to 2008, invalidated students did not receive a scale score.

Table 7.13 shows the state-level means for all grades from 2006 through 2011 for Communication Arts and Mathematics and from 2008 through 2011 for Science. The Science MAP was administered for the first time in 2008. As shown in Table 7.13, the mean scale scores in all grades and content areas increased from 2010 to 2011.

Table 7.14 shows the percentage of students in each achievement level in 2006 through 2011 on the Communication Arts test. The percentages at or above Proficient increased from 2010 to 2011 for all grades.

Table 7.15 shows the percentage of students in each achievement level in 2006 through 2011 on the Mathematics test. As compared to 2010, increases in the percentage of students at or above Proficient were observed in all grades in 2011 except for Grade 8 where the percentage of students at or above Proficient decreased slightly.

Table 7.16 shows the percentage of students in each achievement level in 2008 through 2011 on the Science test. In Grades 5 and 8, the percentage of students at or above Proficient increased from 2010 to 2011.

**Table 7. 1: Participation Rates: All Students**

Grade	Accountable in Comm. Arts	Percent Reportable in Comm. Arts	Accountable in Mathematics	Percent Reportable in Mathematics	Accountable in Science	Percent Reportable in Science
3	66487	99.6%	66487	99.7%		
4	67049	99.6%	67049	99.7%		
5	67461	99.4%	67461	99.5%	67461	99.6%
6	66633	99.7%	66633	99.8%		
7	67517	99.6%	67517	99.7%		
8	66205	99.5%	66205	99.6%	66205	99.4%

**Table 7. 2: Participation Rates: Males**

Grade	Accountable in Comm. Arts	Percent Reportable in Comm. Arts	Accountable in Mathematics	Percent Reportable in Mathematics	Accountable in Science	Percent Reportable in Science
3	34051	99.5%	34051	99.6%		
4	33956	99.6%	33956	99.7%		
5	34638	99.4%	34638	99.5%	34638	99.6%
6	34103	99.7%	34103	99.7%		
7	34650	99.6%	34650	99.6%		
8	33678	99.5%	33678	99.6%	33678	99.3%

**Table 7. 3: Participation Rates: Females**

Grade	Accountable in Comm. Arts	Percent Reportable in Comm. Arts	Accountable in Mathematics	Percent Reportable in Mathematics	Accountable in Science	Percent Reportable in Science
3	32299	99.6%	32299	99.7%		
4	32952	99.6%	32952	99.8%		
5	32687	99.4%	32687	99.5%	32687	99.6%
6	32467	99.7%	32467	99.8%		
7	32790	99.7%	32790	99.7%		
8	32382	99.6%	32382	99.7%	32382	99.6%

**Table 7. 4: Participation Rates: White**

Grade	Accountable in Comm. Arts	Percent Reportable in Comm. Arts	Accountable in Mathematics	Percent Reportable in Mathematics	Accountable in Science	Percent Reportable in Science
3	48699	99.8%	48699	99.8%		
4	49710	99.9%	49710	99.9%		
5	50130	99.8%	50130	99.8%	50130	99.8%
6	49867	99.8%	49867	99.8%		
7	50491	99.8%	50491	99.8%		
8	50001	99.7%	50001	99.7%	50001	99.6%

**Table 7. 5: Participation Rates: Black**

Grade	Accountable in Comm. Arts	Percent Reportable in Comm. Arts	Accountable in Mathematics	Percent Reportable in Mathematics	Accountable in Science	Percent Reportable in Science
3	11566	99.2%	11566	99.1%		
4	11379	99.0%	11379	99.4%		
5	11658	98.6%	11658	98.5%	11658	99.0%
6	11395	99.5%	11395	99.5%		
7	11670	99.3%	11670	99.3%		
8	11118	99.3%	11118	99.4%	11118	98.8%

**Table 7. 6: Participation Rates: Hispanic**

Grade	Accountable in Comm. Arts	Percent Reportable in Comm. Arts	Accountable in Mathematics	Percent Reportable in Mathematics	Accountable in Science	Percent Reportable in Science
3	3541	98.4%	3541	99.1%		
4	3289	98.4%	3289	99.4%		
5	3138	97.8%	3138	98.6%	3138	99.2%
6	2948	99.5%	2948	99.6%		
7	2945	99.0%	2945	99.5%		
8	2644	99.0%	2644	99.5%	2644	99.4%

**Table 7. 7: Participation Rates: Asian/Pacific Islander**

Grade	Accountable in Comm. Arts	Percent Reportable in Comm. Arts	Accountable in Mathematics	Percent Reportable in Mathematics	Accountable in Science	Percent Reportable in Science
3	1374	96.6%	1374	99.9%		
4	1345	96.1%	1345	99.9%		
5	1336	95.5%	1336	99.3%	1336	99.6%
6	1169	98.0%	1169	99.7%		
7	1193	97.4%	1193	99.7%		
8	1193	98.0%	1193	99.7%	1193	99.7%

**Table 7. 8: Participation Rates: Native American/Alaskan**

Grade	Accountable in Comm. Arts	Percent Reportable in Comm. Arts	Accountable in Mathematics	Percent Reportable in Mathematics	Accountable in Science	Percent Reportable in Science
3	307	99.7%	307	100.0%		
4	289	100.0%	289	99.7%		
5	288	99.3%	288	99.3%	288	99.3%
6	335	100.0%	335	100.0%		
7	335	99.7%	335	99.7%		
8	325	99.4%	325	99.4%	325	99.1%

**Table 7. 9: Participation Rates: Students Receiving Accommodations**

Grade	Accountable in Comm. Arts	Percent Reportable in Comm. Arts	Accountable in Mathematics	Percent Reportable in Mathematics	Accountable in Science	Percent Reportable in Science
3	6412	99.8%	6734	99.9%		
4	7095	99.9%	7420	99.9%		
5	7343	99.8%	7652	99.9%	7327	99.9%
6	6876	99.8%	7160	99.9%		
7	7038	99.8%	7198	99.8%		
8	6562	99.7%	6778	99.7%	6531	99.4%

**Table 7. 10: Summary Statistics for Communication Arts**

Grade	N	Mean	Std. Dev.	Scale Scores by Percentiles				
				10	25	50	75	90
3	66,196	641.19	36.52	598	621	643	663	683
4	66,748	662.18	38.23	616	640	663	685	707
5	67,052	673.68	34.85	632	654	676	695	713
6	66,443	675.02	32.81	637	657	676	695	712
7	67,257	680.56	36.61	636	660	683	704	723
8	65,905	695.11	34.10	655	677	697	716	734

**Table 7. 11: Summary Statistics for Mathematics**

Grade	N	Mean	Std. Dev.	Scale Scores by Percentiles				
				10	25	50	75	90
3	66,258	627.03	39.69	581	604	627	649	669
4	66,881	649.68	34.87	608	630	651	670	688
5	67,124	669.05	42.47	618	644	670	694	717
6	66,476	684.95	39.80	635	661	687	710	731
7	67,294	687.53	40.73	638	665	690	713	735
8	65,956	708.40	40.12	659	684	710	734	756

**Table 7. 12: Summary Statistics for Science**

Grade	N	Mean	Std. Dev.	Scale Scores by Percentiles				
				10	25	50	75	90
5	67,196	666.04	33.43	625	648	669	688	704
8	65,828	700.05	30.98	660	682	703	721	736

**Table 7. 13: Comparison of State-Level Means, 2006 Through 2011 Census Data**

Grade	Year	Communication Arts			Mathematics			Science		
		N	Mean SS	S.D. SS	N	Mean SS	S.D. SS	N	Mean SS	S.D. SS
3	2006	64,486	639.86	36.84	64,763	621.59	39.11			
	2007	66,347	639.58	38.04	66,640	622.40	38.72			
	2008	66,179	637.60	37.54	66,258	621.65	36.92			
	2009	67,163	637.43	38.18	67,232	621.67	36.76			
	2010	66,751	640.27	36.63	66,814	624.89	39.28			
	2011	66,196	641.19	36.52	66,258	627.03	39.69			
4	2006	65,179	654.55	38.56	65,306	643.88	37.07			
	2007	65,274	656.11	39.51	65,363	644.47	36.56			
	2008	66,873	655.61	33.63	66,944	644.18	34.19			
	2009	66,490	656.77	33.41	66,587	644.20	33.89			
	2010	67,301	661.34	38.95	67,394	647.59	34.01			
	2011	66,748	662.18	38.23	66,881	649.68	34.87			
5	2006	66,007	668.18	37.09	66,123	660.06	39.99			
	2007	65,461	671.01	37.14	65,498	663.21	41.50			
	2008	65,544	671.48	33.71	65,636	661.43	40.73	65,586	661.64	31.52
	2009	67,083	671.58	32.84	67,155	662.07	40.52	67,118	662.22	30.40
	2010	66,500	673.65	35.33	66,580	667.70	41.74	66,558	664.76	32.48
	2011	67,052	673.68	34.85	67,124	669.05	42.48	67,196	666.04	33.43
6	2006	66,948	666.85	33.70	67,017	673.30	39.80			
	2007	66,247	667.99	34.63	66,332	676.31	41.75			
	2008	65,672	671.27	33.50	65,716	678.46	41.13			
	2009	65,716	671.67	33.04	65,755	678.87	39.56			
	2010	67,260	674.18	33.12	67,315	683.36	39.48			
	2011	66,443	675.02	32.81	66,476	684.95	39.80			
7	2006	70,290	671.63	37.06	70,698	675.38	41.27			
	2007	67,167	672.11	36.26	67,554	677.41	42.62			
	2008	66,701	675.87	35.08	66,727	681.15	41.38			
	2009	66,316	677.68	34.75	66,330	683.63	40.72			
	2010	66,034	678.85	36.25	66,052	686.51	40.28			
	2011	67,257	680.56	36.61	67,294	687.53	40.73			
8	2006	72,483	686.85	37.87	72,542	697.73	40.37			
	2007	70,187	686.90	37.54	70,204	698.33	41.98			
	2008	67,278	691.05	33.57	67,312	701.30	39.40	67,209	694.36	30.67
	2009	66,741	692.56	33.31	66,770	703.60	38.63	66,702	695.65	30.94
	2010	66,139	694.28	34.01	66,166	707.98	40.04	66,101	698.28	31.07
	2011	65,905	695.11	34.10	65,956	708.40	40.12	65,828	700.05	30.98

**Table 7. 14: Comparison of Percentage of Students in Each Achievement Level, Communication Arts  
2006 Through 2011 Census Data**

Grade	Year	N	No Level	Below Basic	Basic	Proficient	Advanced	Prof & Adv
3	2006	65,344	1.3	8.8	47.5	25.7	16.7	42.4
	2007	67,259	1.4	9.4	46.6	25.8	16.8	42.6
	2008	66,357	0.3	9.3	50.2	25.2	15.1	40.3
	2009	67,357	0.3	9.6	49.8	25.1	15.2	40.3
	2010	66,947	0.3	8.2	48.4	26.9	16.2	43.1
	2011	66,487	0.4	7.6	48.4	27.0	16.6	43.6
4	2006	65,849	1.0	10.6	44.5	28.8	15.0	43.8
	2007	65,982	1.1	10.5	43.4	28.2	16.8	45.1
	2008	67,049	0.3	8.0	46.7	33.4	11.7	45.1
	2009	66,709	0.3	7.6	45.8	33.6	12.7	46.3
	2010	67,510	0.3	8.6	40.2	31.2	19.7	50.9
	2011	67,049	0.4	8.2	39.5	31.6	20.2	51.9
5	2006	66,704	1.0	9.1	44.8	29.6	15.4	45.0
	2007	66,098	1.0	8.3	42.9	29.8	18.0	47.8
	2008	65,734	0.3	6.4	45.1	32.2	15.9	48.1
	2009	67,307	0.3	6.3	44.6	33.9	14.9	48.8
	2010	66,730	0.3	7.1	41.5	32.1	18.9	51.0
	2011	67,461	0.6	6.9	41.4	32.4	18.7	51.1
6	2006	67,709	1.1	11.9	44.8	31.6	10.6	42.2
	2007	67,045	1.2	11.2	44	31.8	11.7	43.6
	2008	65,830	0.2	9.0	43.5	34	13.4	47.4
	2009	65,908	0.3	8.6	43.4	33.8	13.9	47.7
	2010	67,476	0.3	7.8	42.3	33.9	15.7	49.6
	2011	66,633	0.3	7.3	41.9	34.3	16.2	50.5
7	2006	71,632	1.9	13.7	41.8	30.5	12.2	42.7
	2007	68,404	1.8	13.1	40.7	32.8	11.6	44.4
	2008	66,923	0.3	10.0	40.7	36.1	12.9	49.0
	2009	66,531	0.3	8.7	40.3	37.2	13.6	50.8
	2010	66,279	0.4	9.8	38.1	35.2	16.5	51.7
	2011	67,517	0.4	9.0	36.9	36.0	17.8	53.8
8	2006	73,516	1.4	9.1	48.0	26.6	15.0	41.5
	2007	71,200	1.4	8.7	48.3	26.9	14.6	41.6
	2008	67,574	0.4	5.7	45.8	33.1	15.0	48.1
	2009	67,077	0.5	5.3	44.5	33.4	16.3	49.7
	2010	66,463	0.5	4.9	42.8	34.3	17.4	51.8
	2011	66,205	0.5	4.6	42.5	33.9	18.5	52.5

**Table 7. 15: Comparison of Percentage of Students in Each Achievement Level, Mathematics 2006 Through 2011 Census Data**

Grade	Year	N	No Level	Below Basic	Basic	Proficient	Advanced	Prof & Adv
3	2006	65,325	0.9	7.2	48.7	33.3	10.0	43.3
	2007	67,257	0.9	7.2	46.9	35.0	10.0	45.0
	2008	66,357	0.1	6.5	49.6	35.0	8.8	43.8
	2009	67,357	0.2	6.8	48.5	35.6	8.8	44.4
	2010	66,947	0.2	6.2	46.6	37.0	10.1	47.1
	2011	66,487	0.3	5.6	44.7	38.1	11.3	49.4
4	2006	65,845	0.8	8.3	47.5	34.4	9.0	43.4
	2007	65,975	0.9	8.1	46.5	35.2	9.3	44.5
	2008	67,049	0.2	7.6	48.0	36.0	8.2	44.2
	2009	66,709	0.2	7.3	48.2	36.6	7.8	44.4
	2010	67,510	0.2	6.1	45.4	39.3	9.1	48.4
	2011	67,049	0.3	5.6	43.7	39.9	10.5	50.5
5	2006	66,703	0.9	8.1	47.8	32.7	10.6	43.3
	2007	66,075	0.9	7.6	44.9	33.1	13.4	46.6
	2008	65,734	0.1	7.5	46.5	34.4	11.4	45.8
	2009	67,307	0.2	7.5	45.1	35.6	11.6	47.2
	2010	66,730	0.2	6.2	41.9	36.7	15.1	51.7
	2011	67,461	0.5	6.1	40.9	36.3	16.2	52.5
6	2006	67,706	1.0	11.1	44.1	34.4	9.5	43.9
	2007	67,039	1.1	11.1	40.0	35.5	12.3	47.8
	2008	65,830	0.2	9.5	39.6	37.8	12.9	50.7
	2009	65,908	0.2	8.9	40.7	37.5	12.6	50.1
	2010	67,476	0.2	7.8	36.6	40.3	15.0	55.4
	2011	66,633	0.2	7.5	35.4	40.5	16.4	56.9
7	2006	71,575	1.2	17.4	38.5	32.7	10.2	42.9
	2007	68,405	1.2	16.7	37.1	33.2	11.7	44.9
	2008	66,923	0.3	13.9	36.3	36.7	12.8	49.5
	2009	66,531	0.3	12.5	35.2	37.6	14.3	51.9
	2010	66,279	0.3	10.8	34.3	38.8	15.7	54.5
	2011	67,517	0.3	10.5	33.5	39.2	16.6	55.8
8	2006	73,523	1.3	21.1	37.8	27.6	12.2	39.8
	2007	71,190	1.4	21.4	36.6	26.6	14.0	40.6
	2008	67,574	0.4	18.0	37.7	29.9	13.9	43.8
	2009	67,077	0.5	16.4	36.8	31.5	14.9	46.4
	2010	66,463	0.4	14.9	33.3	32.1	19.2	51.3
	2011	66,205	0.4	15.0	33.9	31.0	19.8	50.8

**Table 7. 16: Comparison of Percentage of Students in Each Achievement Level, Science 2008 Through 2011 Census Data**

<b>Grade</b>	<b>Year</b>	<b>N</b>	<b>No Level</b>	<b>Below Basic</b>	<b>Basic</b>	<b>Proficient</b>	<b>Advanced</b>	<b>Prof &amp; Adv</b>
<b>5</b>	2008	65,734	0.2	11.2	44.0	29.6	14.9	44.5
	2009	67,307	0.3	10.6	44.1	30.3	14.8	45.1
	2010	66,730	0.3	10.4	40.5	29.6	19.3	48.9
	2011	67,461	0.4	10.0	39.1	29.5	21.0	50.5
<b>8</b>	2008	67,574	0.5	19.3	37.0	36.7	6.5	43.2
	2009	67,077	0.6	18.2	36.5	37.2	7.6	44.8
	2010	66,463	0.5	16.4	35.1	38.4	9.6	48.0
	2011	66,205	0.6	15.7	33.7	38.6	11.4	50.0

## **CHAPTER 8: ACHIEVEMENT-LEVEL SETTING**

No changes to this chapter. Please see the 2010 MAP Technical Report.

## CHAPTER 9: EVIDENCE OF CONSTRUCT-RELATED VALIDITY

### 9.2 Reliability

#### 9.2.1. Test Reliability

The reliability coefficients for the MAP are reported in Tables 9.1, 9.2, and 9.3 for Communication Arts, Mathematics, and Science, respectively. These reliability coefficients were computed using the census data. All reliability statistics are 0.90 or greater for all tests indicating acceptable reliability. The reliability statistics by subgroup are reported and discussed in Chapter 10.

#### 9.2.2. Standard Error of Measurement

The overall Standard Error of Measurement (SEM) is expressed in scale score units and is a test-level statistic. The SEM is summarized in Table 9.4 with respect to all students and each subgroup.

#### 9.2.3. Conditional Standard Error of Measurement

The CSEM of each cut score is reported in Table 9.5 for the grades/content areas where PE items were removed. Note that the CSEMs vary in magnitude across the entire range of student ability estimates (i.e., scale scores) and are smaller in the middle of the score distribution and higher at the tails. This pattern is seen for all MAP CSEMs and is to be expected when IRT methods are used. The CSEMs at the three cut scores that define the performance levels are presented in Table 9.5 and range from 7 to 14 scale score points.

### 9.4 Analyses by Content Standard

#### 9.4.1. Reliability of Content Standards

Cronbach's (1951) coefficient alpha was computed for each of the Content Standards by grade/content area using the census data. Tables 9.11 through 9.13 report the reliability statistics along the diagonal of each matrix for each grade/content area where PE items were removed. Reliability indices, such as Cronbach's coefficient alpha, are a function of the number of test items. It is expected that coefficient alpha would be low for a Content Standard assessed by a small number of items (e.g., Writing Formally and Informally).

#### 9.4.2. Correlations Among Content Standard Subscores

In this section, we measure the strength of the interrelationships among the Content Standards by computing correlation between the Content Standards. Tables 9.11 through 9.13 report the uncorrected Pearson product-moment (PPM) correlation coefficients, as well as the PPM corrected for attenuation (CAPP), in addition to the reliability coefficients described above. The PPM among the Content Standard subscores is presented below the diagonal portion of the matrix, the CAPP is presented above the

diagonal portion of the matrix, and the reliability coefficients are shown on the diagonal in each table.

The uncorrected PPM in Tables 9.11 through 9.13 should be interpreted in the context of the reliability coefficient. In general, we expect to see lower PPM coefficients between variables that are less reliable. Overall, the PPM coefficients show that performance on one Content Standard is moderately to strongly related to performance on another Content Standard within the same content area.

#### **9.4.3. Standard Error of Measurement of Content Standards**

The SEM associated with each of the Content Standards is reported in Tables 9.14 through 9.16 for Communication Arts, Mathematics, and Science, respectively for each grade/content area where PE items were removed. These SEMs are reported in the percent correct metric as Content Standards are reported in that metric.

**Table 9. 1: Reliability in Communication Arts**

Grade	Number of Items	Number of Score Points	Cronbach's Alpha
3	58	62	0.91
4	58	62	0.92
5	56	61	0.91
6	56	60	0.91
7	62	66	0.90
8	60	64	0.91

**Table 9. 2: Reliability in Mathematics**

Grade	Number of Items	Number of Score Points	Cronbach's Alpha
3	55	59	0.91
4	61	65	0.92
5	58	62	0.91
6	58	62	0.92
7	61	65	0.92
8	59	63	0.92

**Table 9. 3: Reliability in Science**

Grade	Number of Items	Number of Score Points	Cronbach's Alpha
5	55	68	0.90
8	57	71	0.91

**Table 9. 4: SEM by Subgroup**

<b>Grade</b>	<b>Category</b>	<b>Group</b>	<b>CA SEM</b>	<b>MA SEM</b>	<b>SC SEM</b>
<b>3</b>	<b>Overall</b>		10.96	11.91	
	<b>Ethnicity</b>	White (not Hispanic)	11.38	12.06	
		Black (not Hispanic)	10.89	10.87	
		Hispanic	10.72	10.65	
		Asian/Pacific Islander	12.13	13.10	
		Native American	10.73	11.43	
<b>Gender</b>	Male	11.25	11.38		
	Female	11.02	11.71		
<b>Accommodations</b>	No	11.00	12.05		
	Yes	12.54	10.61		
<b>4</b>	<b>Overall</b>		10.81	9.86	
	<b>Ethnicity</b>	White (not Hispanic)	11.01	9.81	
		Black (not Hispanic)	10.11	10.19	
		Hispanic	10.30	9.67	
		Asian/Pacific Islander	11.95	11.00	
		Native American	10.45	10.34	
<b>Gender</b>	Male	10.23	10.03		
	Female	11.07	9.65		
<b>Accommodations</b>	No	11.01	9.87		
	Yes	11.46	10.34		
<b>5</b>	<b>Overall</b>		10.45	12.74	10.57
	<b>Ethnicity</b>	White (not Hispanic)	10.40	12.65	10.48
		Black (not Hispanic)	10.86	12.81	12.18
		Hispanic	10.50	12.12	11.14
		Asian/Pacific Islander	11.03	14.17	11.49
		Native American	9.47	12.18	10.36
<b>Gender</b>	Male	10.77	12.37	10.92	
	Female	10.50	12.30	10.63	
<b>Accommodations</b>	No	10.22	12.60	10.29	
	Yes	13.02	12.74	12.74	
<b>6</b>	<b>Overall</b>		9.84	11.26	
	<b>Ethnicity</b>	White (not Hispanic)	10.00	11.18	
		Black (not Hispanic)	10.12	10.65	
		Hispanic	10.20	10.35	
		Asian/Pacific Islander	11.32	12.86	
		Native American	9.51	11.12	
<b>Gender</b>	Male	9.98	10.82		
	Female	10.00	10.90		
<b>Accommodations</b>	No	10.20	11.06		
	Yes	11.94	11.31		

**Table 9. 4: SEM by Subgroup (Cont'd)**

Grade	Category	Group	CA SEM	MA SEM	SC SEM
7	Overall		11.58	11.52	
	Ethnicity	White (not Hispanic)	11.51	10.76	
		Black (not Hispanic)	12.25	12.53	
		Hispanic	11.72	11.47	
		Asian/Pacific Islander	11.78	11.65	
		Native American	11.47	12.00	
	Gender	Male	11.98	11.21	
Female		11.26	10.98		
Accommodations	No	11.22	11.17		
	Yes	13.86	14.44		
8	Overall		10.23	11.35	9.29
	Ethnicity	White (not Hispanic)	10.04	10.59	9.21
		Black (not Hispanic)	10.82	12.62	10.03
		Hispanic	9.97	11.70	9.45
		Asian/Pacific Islander	11.55	11.73	9.69
		Native American	10.05	11.63	9.23
	Gender	Male	10.78	10.95	9.17
Female		9.88	10.94	9.27	
Accommodations	No	9.86	10.50	9.03	
	Yes	13.75	15.09	10.76	

**Table 9. 5: Conditional Standard Error of Measurement at the Basic, Proficient, & Advanced Cut Scores**

Content Area	Grade	Basic		Proficient		Advanced	
		Cut Score	CSEM	Cut Score	CSEM	Cut Score	CSEM
Communication Arts	3	592	8	648	10	673	14
	7	634	10	680	9	712	12
Mathematics	4	596	9	651	8	688	13
	8	670	11	710	8	741	8
Science	5	626	10	669	8	692	9
	8	671	9	703	7	735	8

**Table 9. 11: Reliability (Diagonal) of Each Content Standard, Uncorrected Correlation Coefficient (below Diagonal), and Corrected Correlation Coefficient (above Diagonal) Among Content Standards: Communication Arts**

Grade	No.	Content Standard	Number of Items	1	2	3	4	5
3	1	Speaking/Writing Standard English	15	0.72	0.94	0.91		0.93
	2	Reading Fiction/Poetry/Drama	23	0.70	0.78	0.97		1.12
	3	Reading Nonfiction	18	0.69	0.77	0.80		1.13
	4	Writing Formally/Informally	NR*					
	5	Combined Reading	41	0.74	0.93	0.95		0.88
7	1	Speaking/Writing Standard English	16	0.65	0.91	0.91		0.92
	2	Reading Fiction/Poetry/Drama	24	0.65	0.78	0.97		1.15
	3	Reading Nonfiction	20	0.66	0.77	0.80		1.11
	4	Writing Formally/Informally	NR*					
	5	Combined Reading	44	0.69	0.95	0.93		0.88

\*NR=Not Reported

**Table 9. 12: Reliability (Diagonal) of Each Content Standard, Uncorrected Correlation Coefficient (below Diagonal), and Corrected Correlation Coefficient (above Diagonal) Among Content Standards: Mathematics**

Grade	No.	Content Standard	Number of Items	1	2	3	4	5
4	1	Number and Operations	23	0.84	1.00	0.91	0.99	0.95
	2	Algebraic Relationships	13	0.75	0.66	0.95	1.00	0.98
	3	Geometric and Spatial Relationships	9	0.64	0.60	0.60	0.95	0.96
	4	Measurement	10	0.74	0.66	0.61	0.67	0.95
	5	Data and Probability	6	0.63	0.58	0.54	0.56	0.52
8	1	Number and Operations	13	0.73	0.94	0.96	1.01	0.99
	2	Algebraic Relationships	18	0.74	0.83	0.98	1.02	1.00
	3	Geometric and Spatial Relationships	15	0.67	0.73	0.68	1.05	1.01
	4	Measurement	6	0.65	0.69	0.65	0.56	1.05
	5	Data and Probability	8	0.66	0.71	0.65	0.61	0.61

**Table 9. 13: Reliability (Diagonal) of Each Content Standard, Uncorrected Correlation Coefficient (below Diagonal), and Corrected Correlation Coefficient (above Diagonal) Among Content Standards: Science**

Grade	No.	Content Standard	Number of Items	1	2	3	4	5	6	7	8
5	1	Matter and Energy	9	0.58	1.04	0.89	1.00	1.00	0.95	0.90	0.95
	2	Force and Motion	5	0.53	0.45	0.98	1.05	1.05	1.01	1.00	1.09
	3	Characteristics of Living Organisms	7	0.49	0.48	0.52	1.03	0.96	0.94	0.98	1.04
	4	Interactions of Organisms	7	0.58	0.54	0.57	0.59	1.03	0.99	1.00	1.07
	5	Earth's Processes	7	0.60	0.55	0.54	0.62	0.61	1.00	0.97	1.02
	6	The Universe	6	0.55	0.51	0.51	0.57	0.59	0.57	0.93	0.98
	7	Scientific Inquiry	8	0.47	0.46	0.48	0.52	0.52	0.48	0.47	1.02
	8	Technology and the Environment	6	0.44	0.44	0.46	0.50	0.48	0.45	0.43	0.37
8	1	Matter and Energy	9	0.66	1.10	1.04	1.02	1.02	0.96	0.99	0.99
	2	Force and Motion	6	0.54	0.37	1.06	1.04	1.05	1.02	1.05	1.02
	3	Characteristics of Living Organisms	6	0.60	0.45	0.50	1.03	1.05	0.92	0.98	1.02
	4	Interactions of Organisms	6	0.62	0.47	0.55	0.56	1.04	0.93	0.97	1.02
	5	Earth's Processes	8	0.63	0.49	0.56	0.59	0.58	0.92	1.01	1.05
	6	The Universe	6	0.60	0.48	0.51	0.54	0.54	0.60	0.85	0.86
	7	Scientific Inquiry	10	0.63	0.50	0.54	0.57	0.60	0.51	0.61	1.01
	8	Technology and the Environment	6	0.60	0.47	0.54	0.57	0.60	0.50	0.59	0.56

**Table 9. 14: Mean, Standard Deviation, and SEM of Communication Arts Content Standards**

Grade	Content Standard	Mean	Std. Deviation	SEM
3	1	73.44	18.87	9.99
	2	77.29	16.27	7.63
	3	72.84	19.65	8.79
	5	75.11	16.86	5.84
7	1	62.48	17.73	10.49
	2	73.70	16.40	7.69
	3	73.72	19.50	8.72
	5	73.71	16.64	5.76

**Table 9. 15: Mean, Standard Deviation, and SEM of Mathematics Content Standards**

Grade	Content Standard	Mean	Std. Deviation	SEM
4	1	75.01	19.00	7.60
	2	73.13	19.23	11.21
	3	75.88	18.91	11.96
	4	69.92	22.16	12.73
	5	78.43	20.55	14.24
8	1	68.03	22.03	11.45
	2	52.59	24.41	10.06
	3	57.47	19.72	11.16
	4	44.04	26.74	17.74
	5	61.48	22.86	14.28

**Table 9. 16: Mean, Standard Deviation, and SEM of Science Content Standards**

Grade	Content Standard	Mean	Std. Deviation	SEM
5	1	53.28	21.04	13.64
	2	58.15	22.41	16.62
	3	74.06	20.47	14.18
	4	60.08	23.41	14.99
	5	53.59	25.40	15.86
	6	60.74	22.65	14.85
	7	76.12	19.13	13.93
	8	61.02	20.30	16.11
8	1	52.10	20.92	12.20
	2	54.36	22.27	17.68
	3	46.74	18.19	12.86
	4	55.17	25.80	17.11
	5	61.26	21.51	13.94
	6	39.29	22.70	14.36
	7	77.70	19.25	12.02
	8	54.26	20.51	13.60

## CHAPTER 10: FAIRNESS

### 10.3 Evaluating Bias Through Impact Analysis

The impact of achievement testing on minorities can be determined and reported in the form of average scores and also in terms of test score reliability. Tables 10.4 through 10.9 present the number of students, scale score means, standard deviations, effect size (Cohen's  $d$ ), and test form reliability statistics (Coefficient Alpha, see Chapter 9) for various subgroups of interest.

#### 10.3.1 Reliability

Tables 10.4 through 10.9 show the test reliability for the various subgroups of interest. This analysis shows that the test reliability is of acceptable magnitude for all the subgroups.

#### 10.3.2 Effect Size

One way to evaluate the magnitude of the differences is to calculate the effect size. Cohen's  $d$  was used to calculate the effect size. Cohen's  $d$  is given by the formula:

$$d = \frac{\bar{x}_a - \bar{x}_b}{\sqrt{\frac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2}{(n_a + n_b) - 2}}},$$

where  $\bar{x}_a$  is the mean score of group A,  $\bar{x}_b$  is the mean score of group B,  $s_a^2$  is the variance of group A,  $s_b^2$  is the variance of group B,  $n_a$  is the number of students in group A, and  $n_b$  is the number of students in group B.

Cohen's  $d$ , then, expresses the difference in group means in terms of the standard deviation. For example, if  $d=.34$  for two groups, then it may be interpreted that the mean difference between the two groups is .34 of the pooled standard deviation. Cohen (1988) offered guidelines for interpreting the meaning of the  $d$  statistic:  $d=.20$  is a small-effect size,  $d=.50$  is a medium-effect size, and  $d=.80$  is a large-effect size.

Using Cohen's (1988) guidelines, certain trends become apparent in Tables 10.4 through 10.9. On the Communication Arts test in all grades, there are small differences in mean test scores between Females and Males where Females outperform Males. On the Communication Arts, Mathematics, and Science tests in all grades, there is a large difference between the mean test scores of accommodated and non-accommodated students where accommodated students underperform non-accommodated students.<sup>2</sup>

There is a medium difference in mean Communication Arts test scores of Black students compared to White students where Black students underperform White students in all

---

<sup>2</sup> Accommodated students include English-language learners who receive accommodations.

grades. There is a small difference between the mean test scores of Hispanic and White students where Hispanics underperform White students on Communication Arts in all grades except Grade 3 where there is a medium difference. Similarly, there is a small difference between the mean test scores of Native Americans and White students where Native Americans underperform White students on Communication Arts in all grades except Grade 5. There is a small difference in the mean Communication Arts test scores where Asian/Pacific Islander students outperform White students in all grades except Grades 3 and 8.

There is a medium difference between the mean Mathematics tests scores of Black and White students where Black students underperform White students in all grades, except Grades 6 and 8 where there is a large difference between mean test scores. There is a small difference in mean Mathematics test scores of Hispanic students compared to White students in Grades 3 through 8 where Hispanic students underperform White students. There is a small difference between the mean test scores of Native American students and White students where Native American students underperform White students in all grades except for Grade 5. Finally, there is a small difference between the mean Mathematics test scores of Asian/Pacific Islander students and White students where Asian/Pacific Islander students outperform White students in all grades, except Grade 6 where there is a medium difference between mean test scores.

There is a large difference between the mean Science test scores of Black students and White students in Grades 5 and 8 where Black students underperform White students. There is a medium difference between mean Science test scores of Hispanic students and White students in Grades 5 and 8 where Hispanic students underperform White students. There is a small difference between the mean Science test scores of Native American students and White students in Grade 8 where Native American students underperform White students.

**Table 10. 4: Impact Analysis, Grade 3**

Content Area	Category	Group	N	Mean	Std. Dev.	Effect Size	Coefficient Alpha
Communication Arts	Ethnicity	White (not Hispanic)	48607	646.32	34.32		0.89
		Black (not Hispanic)	11475	622.42	38.50	0.68	0.92
		Hispanic	3486	628.71	35.74	0.51	0.91
		Asian/Pacific Islander	1327	651.75	38.36	-0.16	0.90
		Native American	306	638.60	32.34	0.23	0.89
	Gender	Male	33897	637.13	37.50		0.91
		Female	32166	645.55	34.85	-0.23	0.90
	Accommodations	No	59795	645.46	33.18		0.89
		Yes	6401	601.29	41.80	1.30	0.91
Mathematics	Ethnicity	White (not Hispanic)	48612	632.57	38.14		0.90
		Black (not Hispanic)	11461	605.20	38.44	0.72	0.92
		Hispanic	3508	616.66	35.50	0.42	0.91
		Asian/Pacific Islander	1372	644.23	43.67	-0.30	0.91
		Native American	307	623.98	38.11	0.23	0.91
	Gender	Male	33916	626.80	40.25		0.92
		Female	32207	627.37	39.03	-0.01	0.91
	Accommodations	No	59552	630.82	38.10		0.90
		Yes	6706	593.29	37.52	0.99	0.92

**Table 10. 5: Impact Analysis, Grade 4**

Content Area	Category	Group	N	Mean	Std. Dev.	Effect Size	Coefficient Alpha
Communication Arts	Ethnicity	White (not Hispanic)	49639	666.89	36.70		0.91
		Black (not Hispanic)	11266	643.72	38.22	0.63	0.93
		Hispanic	3235	650.82	36.42	0.44	0.92
		Asian/Pacific Islander	1293	676.35	42.26	-0.26	0.92
		Native American	289	657.47	36.96	0.26	0.92
	Gender	Male	33804	656.96	38.65		0.93
		Female	32812	667.70	36.91	-0.28	0.91
	Accommodations	No	59664	667.09	34.82		0.90
		Yes	7084	620.83	40.53	1.30	0.92
Mathematics	Ethnicity	White (not Hispanic)	49642	654.18	32.72		0.91
		Black (not Hispanic)	11307	630.38	36.04	0.71	0.92
		Hispanic	3268	642.77	32.24	0.35	0.91
		Asian/Pacific Islander	1343	668.67	41.56	-0.44	0.93
		Native American	288	645.09	36.56	0.28	0.92
	Gender	Male	33858	649.29	35.47		0.92
		Female	32886	650.25	34.10	-0.03	0.92
	Accommodations	No	59496	653.31	32.89		0.91
		Yes	7385	620.46	36.56	0.99	0.92

**Table 10. 6: Impact Analysis, Grade 5**

<b>Content Area</b>	<b>Category</b>	<b>Group</b>	<b>N</b>	<b>Mean</b>	<b>Std. Dev.</b>	<b>Effect Size</b>	<b>Coefficient Alpha</b>
<b>Communication Arts</b>	<b>Ethnicity</b>	White (not Hispanic)	50026	678.04	32.89		0.90
		Black (not Hispanic)	11496	655.99	36.19	0.66	0.91
		Hispanic	3068	664.38	35.01	0.41	0.91
		Asian/Pacific Islander	1276	686.98	39.01	-0.27	0.92
		Native American	286	673.17	33.50	0.15	0.92
	<b>Gender</b>	Male	34419	670.10	35.91		0.91
		Female	32506	677.55	33.21	-0.22	0.90
	<b>Accommodations</b>	No	59724	678.43	30.81		0.89
		Yes	7328	635.03	41.18	1.35	0.90
<b>Mathematics</b>	<b>Ethnicity</b>	White (not Hispanic)	50032	674.88	39.99		0.90
		Black (not Hispanic)	11483	644.32	42.69	0.75	0.91
		Hispanic	3095	659.07	40.41	0.40	0.91
		Asian/Pacific Islander	1326	690.79	50.12	-0.40	0.92
		Native American	286	668.60	40.61	0.16	0.91
	<b>Gender</b>	Male	34466	668.87	43.75		0.92
		Female	32528	669.36	41.01	-0.01	0.91
	<b>Accommodations</b>	No	59515	673.99	39.86		0.90
		Yes	7609	630.35	42.45	1.09	0.91
<b>Science</b>	<b>Ethnicity</b>	White (not Hispanic)	50019	672.64	29.06		0.87
		Black (not Hispanic)	11542	640.07	36.73	1.06	0.89
		Hispanic	3114	654.99	33.58	0.60	0.89
		Asian/Pacific Islander	1331	672.42	38.30	0.01	0.91
		Native American	286	667.16	31.23	0.19	0.89
	<b>Gender</b>	Male	34500	667.09	34.54		0.90
		Female	32563	665.05	32.05	0.06	0.89
	<b>Accommodations</b>	No	59921	669.51	31.02		0.89
		Yes	7275	637.44	38.41	1.01	0.89

**Table 10. 7: Impact Analysis, Grade 6**

Content Area	Category	Group	N	Mean	Std. Dev.	Effect Size	Coefficient Alpha
Communication Arts	Ethnicity	White (not Hispanic)	49785	679.07	31.62		0.90
		Black (not Hispanic)	11335	658.43	32.00	0.65	0.90
		Hispanic	2933	666.48	30.74	0.40	0.89
		Asian/Pacific Islander	1146	688.38	37.73	-0.29	0.91
		Native American	335	669.89	33.64	0.29	0.92
	Gender	Male	34010	670.58	33.26		0.91
		Female	32376	679.73	31.62	-0.28	0.90
	Accommodations	No	59581	679.37	29.46		0.88
		Yes	6862	637.26	36.01	1.39	0.89
Mathematics	Ethnicity	White (not Hispanic)	49782	690.61	37.27		0.91
		Black (not Hispanic)	11340	660.35	40.25	0.80	0.93
		Hispanic	2937	676.55	36.58	0.38	0.92
		Asian/Pacific Islander	1165	709.37	45.48	-0.50	0.92
		Native American	335	677.15	42.02	0.36	0.93
	Gender	Male	34015	684.26	40.91		0.93
		Female	32400	685.72	38.54	-0.04	0.92
	Accommodations	No	59353	689.84	36.87		0.91
		Yes	7123	644.21	39.98	1.23	0.92

**Table 10. 8: Impact Analysis, Grade 7**

Content Area	Category	Group	N	Mean	Std. Dev.	Effect Size	Coefficient Alpha
Communication Arts	Ethnicity	White (not Hispanic)	50374	685.45	34.69		0.89
		Black (not Hispanic)	11593	660.52	36.92	0.71	0.89
		Hispanic	2917	671.25	35.33	0.41	0.89
		Asian/Pacific Islander	1162	693.43	44.53	-0.23	0.93
		Native American	334	677.58	34.58	0.23	0.89
	Gender	Male	34504	674.32	37.87		0.90
		Female	32680	687.19	33.96	-0.36	0.89
	Accommodations	No	60237	685.77	32.39		0.88
		Yes	7020	635.83	40.01	1.50	0.88
Mathematics	Ethnicity	White (not Hispanic)	50375	693.35	38.03		0.92
		Black (not Hispanic)	11586	662.60	41.78	0.79	0.91
		Hispanic	2931	678.14	38.22	0.40	0.91
		Asian/Pacific Islander	1190	710.47	47.55	-0.45	0.94
		Native American	334	681.49	37.94	0.31	0.90
	Gender	Male	34528	686.00	42.38		0.93
		Female	32691	689.20	38.81	-0.08	0.92
	Accommodations	No	60152	692.83	37.22		0.91
		Yes	7142	642.92	41.70	1.32	0.88

**Table 10. 9: Impact Analysis, Grade 8**

<b>Content Area</b>	<b>Category</b>	<b>Group</b>	<b>N</b>	<b>Mean</b>	<b>Std. Dev.</b>	<b>Effect Size</b>	<b>Coefficient Alpha</b>
<b>Communication Arts</b>	<b>Ethnicity</b>	White (not Hispanic)	49847	699.62	31.76		0.90
		Black (not Hispanic)	11041	675.60	36.05	0.74	0.91
		Hispanic	2618	687.59	33.23	0.38	0.91
		Asian/Pacific Islander	1169	705.40	43.67	-0.18	0.93
		Native American	323	693.25	33.50	0.20	0.91
	<b>Gender</b>	Male	33513	690.33	35.93		0.91
		Female	32256	700.17	31.23	-0.29	0.90
	<b>Accommodations</b>	No	59367	699.76	29.73		0.89
		Yes	6538	652.88	41.46	1.51	0.89
<b>Mathematics</b>	<b>Ethnicity</b>	White (not Hispanic)	49858	714.19	37.44		0.92
		Black (not Hispanic)	11048	682.69	39.89	0.83	0.90
		Hispanic	2632	698.75	39.00	0.41	0.91
		Asian/Pacific Islander	1189	730.57	47.89	-0.43	0.94
		Native American	323	703.85	41.10	0.28	0.92
	<b>Gender</b>	Male	33536	708.14	41.39		0.93
		Female	32288	708.77	38.69	-0.02	0.92
	<b>Accommodations</b>	No	59245	713.22	37.14		0.92
		Yes	6711	665.85	40.32	1.26	0.86
<b>Science</b>	<b>Ethnicity</b>	White (not Hispanic)	49791	705.92	27.76		0.89
		Black (not Hispanic)	10990	675.22	31.71	1.08	0.90
		Hispanic	2629	690.14	29.87	0.57	0.90
		Asian/Pacific Islander	1190	707.50	34.26	-0.06	0.92
		Native American	322	699.32	29.19	0.24	0.90
	<b>Gender</b>	Male	33453	701.34	32.40		0.92
		Female	32240	698.80	29.31	0.08	0.90
	<b>Accommodations</b>	No	59379	703.51	28.57		0.90
		Yes	6449	668.23	34.03	1.21	0.90

# **Appendix A**

## Special Study on the Removal of Performance Events from the Grade-Level Missouri Assessment Program

Traditionally the Grade-Level Missouri Assessment Program (MAP) has consisted of selected-response (SR) items, constructed-response (CR) items, and performance events (PEs). Recent budget constraints have forced the Department of Elementary and Secondary Education (DESE) to temporarily discontinue the administration of PEs. To that end, this report examines the effect of removing PEs on the scale scores of Missouri students. The report was guided by three underlying questions:

1. What is the immediate effect of removing PEs?
2. Will the 2011 MAP scale scores be comparable to the 2010 MAP scale scores?
3. Is a new standard setting warranted given the removal of PEs?

### *Performance Events*

Before exploring these three questions, a review of the PE is necessary. The PE is a special type of CR item that requires students to more fully explain or examine a concept than the traditional CR item. Table 1 shows the number of PEs and PE points in each grade and content area as well as the total number of items and score points. PEs were administered only in Grades 3 and 7 Communication Arts, Grades 4 and 8 Mathematics, and Grades 5 and 8 Science.

In Communication Arts, the PEs were writing prompts that were administered in Grades 3 and 7. Students were given 60 to 90 minutes to respond to the writing prompt. The writing prompt was worth four score points.

In Mathematics, the PE was a single item worth four score points and was administered in Grades 4 and 8. Students were given 15 to 20 minutes to respond to the Mathematics PE.

The Science PE comprised the last session of the Grades 5 and 8 Science tests. As shown in Table 1, the PE in Grade 5 consisted of 8 items worth 14 points. The PE in Grade 8 consisted of 9 items worth 15 points. The items for each Science PEs were associated with a single stimulus. The students were given 55 to 70 minutes to respond to the Science PEs.

**Table 1. Number of Performance Events and Number of Points**

	<i>Number of PE Items</i>	<i>Number of PE Points</i>	<i>Number of Total Items</i>	<i>Number of Score Total Points</i>
G3 CA	1	4	56	63
G4 MA	1	4	62	69
G5 SC	8	14	63	82
G7 CA	1	4	63	70
G8 MA	1	4	61	68
G8 SC	9	15	66	86

### *Effect of Removing Performance Events*

To some extent, the effect of removing the PEs is known. Removing the PE

- decreases testing time
- decreases the number of score points in some reporting categories
- affects reliability statistics
- decreases measurement precision
- changes the percentage of students in each achievement level

These issues can and will be examined using data from the 2010 test with and without the PEs.

There are several other areas where it is harder to predict what will occur when PEs are removed. These include (but are not limited to)

- student motivation
- student test fatigue
- alignment of curriculum to assessment
- educator motivation to teach skills measured by PEs
- educator perception of validity of MAP scores without PEs

These issues can only be addressed through surveying students and teachers, which is beyond the scope of this paper.

### *Comparability of Scores*

The comparability of scale scores must also be examined when PEs are removed. This question asks whether the construct being measured is equivalent with and without the PEs. If the construct is not equivalent, then it will not be appropriate to compare the 2010 MAP scores to the 2011 MAP scores in those grades/content areas that have PEs. This question will be explored in this report.

### *Necessity of Standard Setting*

Whenever a test is changed, it is necessary to ask if those changes warrant that cut scores be reset or, at least, re-examined and adjusted. Within a testing program's life cycle, cut scores are usually changed with the introduction of new underlying content standards, new test blueprints, and/or a new testing program. Even then, most states are reluctant to change cut scores if there has not been a large change between the old and new content standards and/or test blueprints. States sometimes change/adjust cuts when students have outgrown the current cut scores. For example, if 100% of students were Proficient, then this may warrant an adjustment of cut scores so that teachers and students can work toward higher goals.

In Missouri, cut scores were established for the grade-span program in the late 1990s. These cut scores were in place for the entire life span of that program. New cut scores were introduced in 2006 for Communication Arts and Mathematics when Missouri changed to the grade-level MAP tests. New cut scores were introduced for Science in 2008.

The stability of the cut scores is important to the stability of the testing program itself. If cut scores are changed frequently, then this may create a perception that the state sponsoring agency

(DESE) is chasing a desired outcome. This means that the need for resetting (or re-examining) cut scores must also be taken in the larger context of the anticipated lifespan of the current grade-level testing program.

This question will be examined again in the Discussion section of this report.

## Methodology

Data from the 2010 administration of the grade-level MAPs were used to examine the guiding questions of this study. In this section, we will first discuss the data followed by an explanation of the methodology used to address the first two guiding questions (see page 1). Throughout this report, the data and results from the MAP with PEs are referred to as the “original” condition, and the data and results from the MAP without PEs are called the “noPE” condition.

### *Data*

The data were taken from the 2010 administration of the grade-level MAPs in grades/content areas where PEs were administered. The scale scores based on all test items including PEs were already available from the Spring 2010 administration.

### *Effect of Removing Performance Events*

To examine the effect of removing PEs on the psychometric properties of the test, it was first necessary to re-calibrate the 2010 data without the PEs.

### *Calibration*

The calibration sample was re-calibrated without the PEs using the same 3-parameter logistic/2-parameter partial credit (3PL/2PPC) IRT models. The data were linked (through Stocking and Lord equating) to the MAP scale using the same anchor set as the original calibration. Then data were scored using item-pattern scoring, which applied the IRT parameters derived from the new calibration.

### *Matched Data*

To compare the scale scores with PEs to those without PEs, it was necessary to match the data from the calibration sample to the original scored data in the General Research File (GRF). This file contains all MO students who were administered MAP in 2010.

The data in the GRF were delivered to DESE who then applied cleaning rules, such as removing duplicate students. For this reason, it was necessary to apply cleaning rules to the GRF prior to matching the re-calibrated and re-scored data.

Cleaning Rules. The following students were removed from the GRF prior to matching the re-calibrated data:

- Students with duplicate MOSIS ID numbers
- Students with duplicate first name, last name, and birthdays

- Students whose achievement level was “0”
- Students who were invalidated
- Students who did not have a valid attempt

Table 2 shows the total number of students in the GRF, the number removed, the total number in the study, and the overall percent of data that matched. As shown in Table 2, 99% of the GRF data were used.

**Table 2. Total Number of Students in GRF, Students Removed, Students in Study, and Percent of Data Matched.**

	<i>Total in GRT</i>	<i>Total Removed</i>	<i>Total in Study</i>	<i>% Data Matched</i>
G3 CA	66947	541	66406	99.2%
G4 MA	67510	424	67086	99.4%
G5 SC	66730	466	66264	99.3%
G7 CA	66279	535	65744	99.2%
G8 MA	66463	626	65837	99.1%
G8 SC	66463	686	65777	99.0%

### *Descriptive Analysis*

Correlation. The correlation of the original raw score to the noPE raw score was computed as was the correlation between the original raw score and the PE score.

Reliability. The reliability was computed for both the original and noPE tests using Chronbach’s alpha.

Scale Scores. The mean and standard deviation of the original scale scores and the noPE scale scores was computed.

Achievement Levels. Students were assigned to achievement levels based on their original and noPE scale scores. The frequency distributions of both sets of achievement levels were compiled.

Cross tabulations were used to compare the original achievement levels (AL\_original) to the noPE achievement levels (AL\_noPE). The percentage of students who did not change achievement level was computed.

The achievement levels were also dichotomized into above and below Proficient. The percent perfect agreement was also computed for the dichotomized and multilevel cross tabulations.

Conditional Standard Error of Measurement (CSEM). The CSEM was computed for the lowest obtainable scale score (LOSS), all cut points, and the highest obtainable scale score (HOSS) for both the original and noPE conditions.

## Score Comparability

The test construct was explored for all grades/content areas with PEs. Because the Science tests have more PE items and points, more in-depth analyses were undertaken for these two tests.

### *Test Construct*

The percentage of items measuring each Content Standard was compiled before and after the removal of PEs. As a rule of thumb, CTB/McGraw-Hill allows the percentage of items covering each Content Standard to vary by up to 10 percentage points between test forms.

### *Residuals Analyses*

The item score residual distributions were analyzed for the SR and CR items. If the residual patterns are similar across the two conditions (original versus noPE), we can be confident that the construct is being measured in a similar way.

In these analyses, the residual was defined as

$$\text{Residual} = \text{observed score} - \text{expected score}.$$

To find the expected score for each person-item combination, the response probability was calculated using the 3PL/2PPC model. The expected score on each SR item was defined as the probability of a correct response, and the expected score on each CR item was defined as

$$\Sigma (x P(x)),$$

where  $x$  is the score point and  $P(x)$  is the probability of the student obtaining that score point. The residuals analyses were first conducted using the full data set and followed by an analysis on the matched data set. Both signed and absolute residuals were computed.

QQ Plots. Quantile-quantile (QQ) plots of the residuals were created for the original versus noPE conditions. The QQ plot allows one to graphically compare the similarity of the distributions of residuals for the two conditions. To construct the QQ plot, the value of the original group's residual at the first percentile was plotted relative to the value of the noPE group's residual at the first percentile. Then this was done again at the second percentile, third percentile, and so forth. If the residuals are distributed in a like manner ( $x = y$ ), then the plotted values will fall along a 45-degree line. If the plot is arced or s-shaped, then this shows graphically that the distribution of residuals is not the same and further investigation is needed. In the QQ plots of the residuals, the  $x$ -axis represents the residual for the original condition and the  $y$ -axis represents the residual for the noPE condition.

### *Principal Components Analyses*

Principal Components Analyses (PCA) with varimax rotation was conducted for both Science tests. This analysis allows one to explore the underlying structure of the Science test and determine whether the PEs measure the construct differently than the remaining CR and SR items.

### Results

In this section, the results of the descriptive analyses are presented followed by the comparability analyses.

### *Descriptive Analysis*

Table 3 shows the correlation between the total raw score with the PE and the total raw score without the PE. It also shows the correlation between the total raw score (with the PE included) and the PE itself. Table 3 shows that the correlation between the total raw score with the PE and the total raw score without the PE is very strong. It also shows that the total score on the PE is moderately to strongly correlated with the total raw score.

**Table 3. Correlation Between Original and noPE Raw Scores and Between Original Raw Score and PE Score**

	<i>Original, noPE Raw Score</i>	<i>Original, PE Score</i>
G3 CA	0.997	0.541
G4 MA	0.996	0.682
G5 SC	0.988	0.714
G7 CA	0.997	0.594
G8 MA	0.997	0.717
G8 SC	0.986	0.831

Table 4 shows the mean scale score for the original and noPE conditions. Even though there was a slight improvement when the PEs were removed, the difference between mean scale scores is less than one scale score point for all grade content areas except for G4 MA, which has a mean scale score difference of 1.1 points between the original and noPE conditions.

**Table 4. N Counts and Means (Standard Deviations) for the Scale Scores with the Original and noPE Parameters.**

	<i>N Count</i>	<i>Original</i>	<i>noPE</i>
G3 CA	66406	640.5 (36.1)	640.7 (36.6)
G4 MA	67086	647.7 (33.9)	648.8 (34.6)
G5 SC	66264	664.9 (32.4)	665.0 (32.9)
G7 CA	65744	679.0 (36.0)	679.7 (35.8)
G8 MA	65837	708.1 (39.9)	708.9 (39.8)
G8 SC	65777	698.4 (30.9)	698.8 (31.2)

*Test Reliability*

Table 5 shows the test reliability for the original and noPE conditions as well as the number of items and points for each grade/content area. As Table 5 shows, there is very little change in the reliability statistics when the PEs are removed.

**Table 5. Number of Items, Score Points, and Cronbach’s Alpha for the Original and noPE Conditions**

	<i>Original</i>			<i>noPE</i>		
	Number of Items	Number of Score Points	Cronbach’s Alpha	Number of Items	Number of Score Points	Cronbach’s Alpha
G3 CA	56	63	0.91	55	59	0.91
G4 MA	62	69	0.92	61	65	0.92
G5 SC	63	82	0.90	55	68	0.90
G7 CA	63	70	0.91	62	66	0.90
G8 MA	61	68	0.93	60	64	0.92
G8 SC	66	86	0.93	57	71	0.91

*Conditional Standard Error of Measurement (CSEM)*

Table 6 shows the CSEM associated with the LOSS, HOSS, and all cut scores for the original and noPE conditions. As shown in Table 6, the CSEM associated with the cut scores varies by no more than one scale score point. In most cases, the CSEM associated with a particular cut score is the same for both the original and noPE conditions.

The CSEM associated with the LOSS and HOSS tends to be larger in the noPE condition than in the original condition. In many cases, the differences are negligible for the LOSS (G4 MA, G7 CA, G8 MA and G8 SC) and for the HOSS (G3 CA, G4 MA, G7 CA, G8 MA, and G8 SC). The differences are larger than 10 scale score points at the LOSS in G3 CA and G5 SC. The differences are larger than 10 scale score points at the HOSS in G5 SC.

**Table 6. Conditional Standard Error of Measurement for Select Scale Locations**

	<i>G3 CA</i>	<i>G4 MA</i>	<i>G5 SC</i>	<i>G7 CA</i>	<i>G8 MA</i>	<i>G8 SC</i>
LOSS	455	465	470	515	525	540
CSEM Original	56	86	68	40	106	75
CSEM noPE	78	87	93	45	109	84
Basic Cut	592	596	626	634	670	671
CSEM Original	9	9	9	10	11	8
CSEM noPE	9	9	10	10	11	9
Proficient Cut	648	651	669	680	710	703
CSEM Original	10	8	8	9	8	7
CSEM noPE	11	8	9	9	8	8
Advanced Cut	673	688	692	712	741	735
CSEM Original	14	12	9	12	7	8
CSEM noPE	14	13	9	12	8	8
HOSS	790	805	855	865	885	895
CSEM Original	63	85	67	95	79	58
CSEM noPE	68	90	78	98	79	61

*Frequency Distribution of Achievement Levels (ALs)*

Table 7 shows the percentage of students in each achievement level for the original and noPE conditions. In most cases, there are slightly more students classified at or above Proficient in the noPE condition compared to the original condition; however, these differences are all less than one percentage point.

**Table 7. Percentage of Students in Each Achievement Level for the Original Achievement Level Data and the No Performance Event Achievement Level Data**

		<i>Below Basic</i>	<i>Basic</i>	<i>Proficient</i>	<i>Advanced</i>	<i>Proficient &amp; Above</i>
G3 CA	Original	8.0	48.6	27.0	16.3	43.3
	noPE	8.0	48.7	26.7	16.5	43.3
	Orig - noPE	0.0	-0.1	0.3	-0.2	0.1
G4 MA	Original	6.0	45.4	39.4	9.1	48.6
	noPE	5.8	45.0	39.1	10.1	49.2
	Orig - noPE	0.2	0.4	0.3	-0.9	-0.6
G5 SC	Original	10.3	40.6	29.7	19.4	49.1
	noPE	10.5	40.4	29.3	19.8	49.2
	Orig - noPE	-0.2	0.2	0.4	-0.5	-0.1
G7 CA	Original	9.7	38.3	35.4	16.6	52
	noPE	9.4	37.9	35.5	17.1	52.6
	Orig - noPE	0.3	0.4	-0.1	-0.5	-0.6
G8 MA	Original	14.9	33.5	32.3	19.4	51.7
	noPE	14.3	33.8	32.2	19.7	51.9
	Orig - noPE	0.6	-0.3	0.1	-0.4	-0.3
G8 SC	Original	16.4	35.3	38.7	9.7	48.3
	noPE	16.4	35.2	37.8	10.6	48.4
	Orig - noPE	0.0	0.1	0.9	-0.9	-0.1

*Percent Perfect Agreement*

Table 8 shows the percent perfect agreement between AL\_original and AL\_noPE using the four achievement levels (“multilevel agreement”) and using dichotomized achievement levels (Proficient and Not Proficient). Grade 8 Science had the lowest levels of perfect agreement. When the assignment of the four achievement levels was compared between the original and noPE conditions, the percent of perfect agreement was 89.8% for Grade 8 Science. When those achievement levels were dichotomized between Proficient/Not Proficient, the percent perfect agreement was 95.1%.

**Table 8. Percent Perfect Agreement Between AL\_original and AL\_noPE**

	<i>Multilevel Agreement</i>	<i>Dichotomous Agreement</i>
G3 CA	95.8%	98.1%
G4 MA	94.7%	97.4%
G5 SC	93.1%	97.2%
G7 CA	95.1%	98.0%
G8 MA	95.3%	98.1%
G8 SC	89.8%	95.1%

*Mean Scores by Change in Achievement Level*

Table 9 shows the number of students, the mean PE score, the mean total raw score without the PE, and the mean total raw score with the PE. This information is provided for students who were classified as Proficient under the original condition but are not Proficient under the noPE condition, students whose achievement level remained the same using both conditions, and students who are now Proficient under the noPE condition but were not Proficient in the original condition. Table 9 shows a pattern where the students who were Proficient had higher mean scores on the PE than did students who are now Proficient. In addition, the students who were Proficient and the students who are now Proficient tend to have similar mean raw scores once the PE is removed.

**Table 9. Mean and Standard Deviation of Raw Score for the PE, noPE, and Original Conditions Disaggregated by Change in Proficiency**

		<i>PE Score</i>			<i>noPE Raw Score</i>		<i>Original Raw Score</i>	
		<b>N</b>	<b>Mean</b>	<b>SD</b>	<b>Mean</b>	<b>SD</b>	<b>Mean</b>	<b>SD</b>
G3 CA	Were Proficient	645	3.9	0.3	47.7	1.1	51.5	1.1
	No Change	65167	2.8	0.8	44.0	9.8	46.8	10.2
	Now Proficient	594	1.9	0.4	47.9	1.0	49.8	1.0
	Total	66406	2.8	0.8	44.0	9.7	46.8	10.1
G4 MA	Were Proficient	652	3.4	0.6	49.7	1.3	53.1	1.4
	No Change	65366	1.8	1.3	47.2	11.0	49.0	11.8
	Now Proficient	1068	0.9	0.9	49.5	2.9	50.4	2.8
	Total	67086	1.8	1.3	47.3	10.8	49.1	11.7
G5 SC	Were Proficient	917	10.9	1.1	42.3	1.2	53.2	1.5
	No Change	64387	8.4	2.5	41.4	11.4	49.8	13.1
	Now Proficient	960	6.4	1.6	43.4	1.2	49.8	1.7
	Total	66264	8.4	2.5	41.4	11.3	49.8	12.9
G7 CA	Were Proficient	458	4.0	0.0	47.8	1.6	51.8	1.6
	No Change	64418	3.0	0.9	46.6	10.6	49.6	11.1
	Now Proficient	868	2.2	0.9	47.1	4.1	49.3	4.1
	Total	65744	3.0	0.9	46.6	10.5	49.6	11.0
G8 MA	Were Proficient	527	3.2	0.9	36.1	1.4	39.3	1.5
	No Change	64610	1.4	1.5	37.4	12.8	38.8	13.8
	Now Proficient	700	0.3	0.6	35.9	4.7	36.2	4.6
	Total	65837	1.4	1.5	37.3	12.7	38.8	13.7
G8 SC	Were Proficient	1593	10.6	1.4	38.4	1.5	49.0	1.8
	No Change	62550	7.3	3.6	38.4	12.0	45.7	14.8
	Now Proficient	1634	4.9	1.6	40.8	1.6	45.7	1.9
	Total	65777	7.3	3.5	38.4	11.7	45.7	14.5

### *Scale Comparability*

The next section of the report addresses scale comparability.

#### *Content Coverage*

Table 10 shows the percent of points that were targeted to each Content Standard with the original MAP blueprint as well as the actual total number of points and total percent of points under the original and noPE conditions. It then shows the difference between the noPE condition and the test Blueprint.

Table 10 shows that the percent of points measuring each Content Standard in Communication Arts and Mathematics is within 10 percentage points of the original target under both the original and noPE conditions. This suggests that the removal of the PE in Communication Arts and Mathematics did not significantly alter the construct being measured.

For Science, the removal of the PE appears to have a greater effect. The Science PE measure the Grade Level Expectation (GLE) “Scientific Inquiry.” The Grade 5 blueprint requires that 25% of the score points measure this GLE while the Grade 8 blueprint requires 28% of the score points measure this GLE. The removal of the PE results in fewer than desirable points measuring Scientific Inquiry. In Grade 5, only 12% of the score points measure this GLE while only 14% of the score points in Grade 8 measure this GLE, both of which represent a more than 10% difference (drop) from the blueprint.

**Table 10. Percentage of Items Measuring Each Standard/Grade-level Expectation with PE and noPE**

Grade/ Content Area	Standard/Grade-level Expectation	Blueprint		Original		noPE		Difference from Blueprint
		Target %	Total Points	% of Total	Total Points	% of Total		
G3 CA	Writing Standard English	22%	12	19%	12	20%	2%	
	Reading (fiction and non-fiction)	68%	45	71%	45	76%	-8%	
	Writing Formally	10%	6	10%	2	3%	7%	
G7 CA	Writing Standard English	22%	16	23%	16	24%	-2%	
	Reading (fiction and non-fiction)	68%	48	69%	48	73%	-5%	
	Writing Formally	10%	6	9%	2	3%	7%	
G4 MA	Number and Operations	35%	24	35%	24	37%	-2%	
	Geometric and Spatial Relationships	15%	10	14%	10	15%	0%	
	Measurement	20%	14	20%	10	15%	4%	
	Data and Probability	10%	7	10%	7	11%	-1%	
	Algebraic Relationships	20%	14	20%	14	22%	-2%	
G8 MA	Number and Operations	10%	13	19%	13	20%	-10%	
	Geometric and Spatial Relationships	25%	16	24%	16	25%	0%	
	Measurement	10%	7	10%	7	11%	-1%	
	Data and Probability	20%	13	19%	9	14%	6%	
	Algebraic Relationships	35%	19	28%	19	30%	5%	
G5 SC	Matter and Energy	13%	11	13%	11	16%	-3%	
	Force and Motion	10%	8	10%	8	12%	-2%	
	Living Organisms	10%	8	10%	8	12%	-2%	
	Organisms with Their Environments	11%	9	11%	9	13%	-2%	
	Earth's Systems	12%	9	11%	9	13%	-1%	
	The Universe	11%	8	10%	8	12%	-1%	
	Scientific Inquiry	25%	22	27%	8	12%	14%	
	Science, Technology, and Human Activity	8%	7	9%	7	10%	-2%	
G8 SC	Matter and Energy	13%	12	14%	12	17%	-4%	
	Force and Motion	8%	7	8%	7	10%	-2%	
	Living Organisms	12%	9	10%	9	13%	0%	
	Organisms with Their Environments	9%	7	8%	7	10%	-1%	
	Earth's Systems	13%	10	12%	10	14%	-1%	
	The Universe	10%	8	9%	8	11%	-1%	
	Scientific Inquiry	28%	25	29%	10	14%	14%	
	Science, Technology, and Human Activity	7%	8	9%	8	11%	-5%	

## Residuals Analyses

Because Science appears to be most impacted by the removal of the PEs, the Grades 5 and 8 Science tests were further studied using residuals analyses and PCA. In the QQ plots of the residuals, the  $x$ -axis represents the residual for the original condition and the  $y$ -axis represents the residual for the noPE condition.

Figure 1 shows the QQ plot of the absolute value of the residuals for the SR items and for the *common* CR items on the Grade 5 Science test. Figure 1 show that the residuals for the SR items and for the CR items are distributed in the same manner for the original and noPE conditions.

Figure 2 shows the QQ plot of the absolute value of the residuals for the SR items and for the *common* CR items on the Grade 8 Science test. Figure 2 show that the residuals for the SR items and for the CR items are distributed in the same manner for the original and noPE conditions.

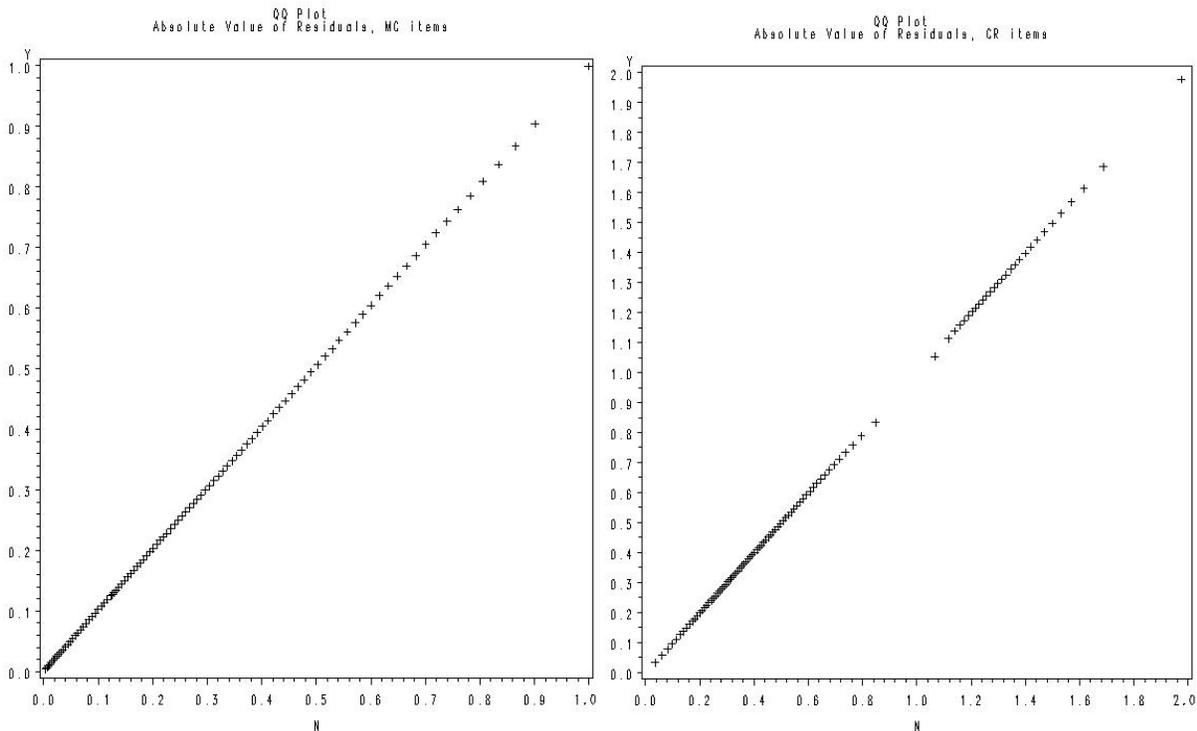
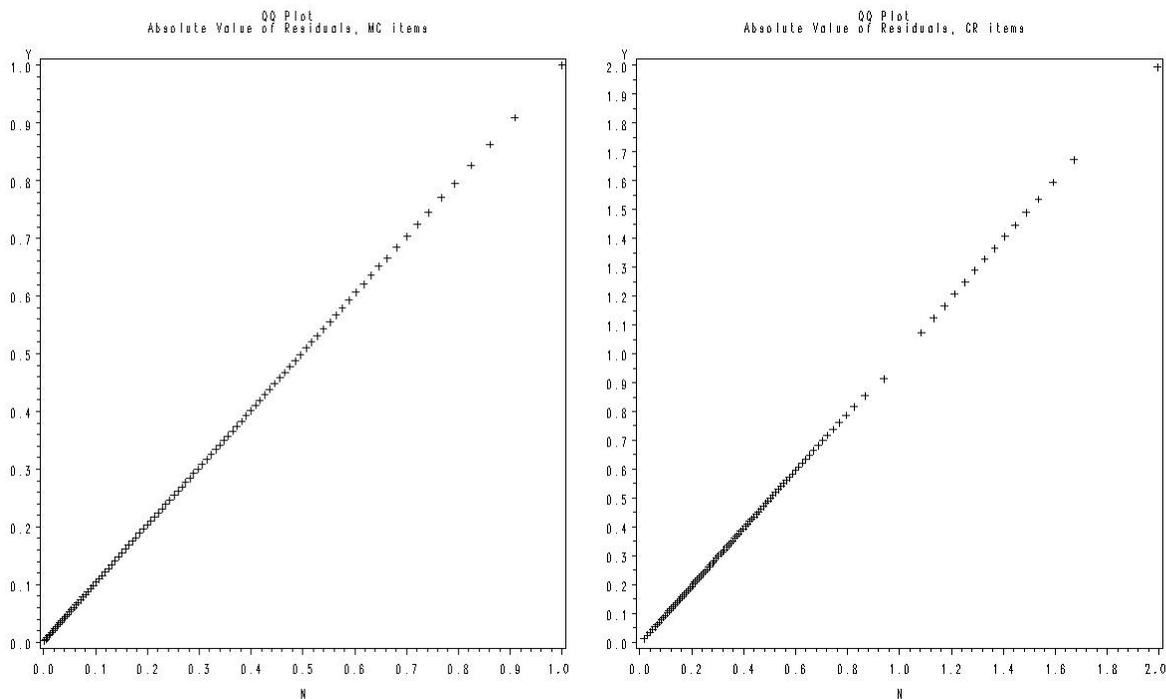


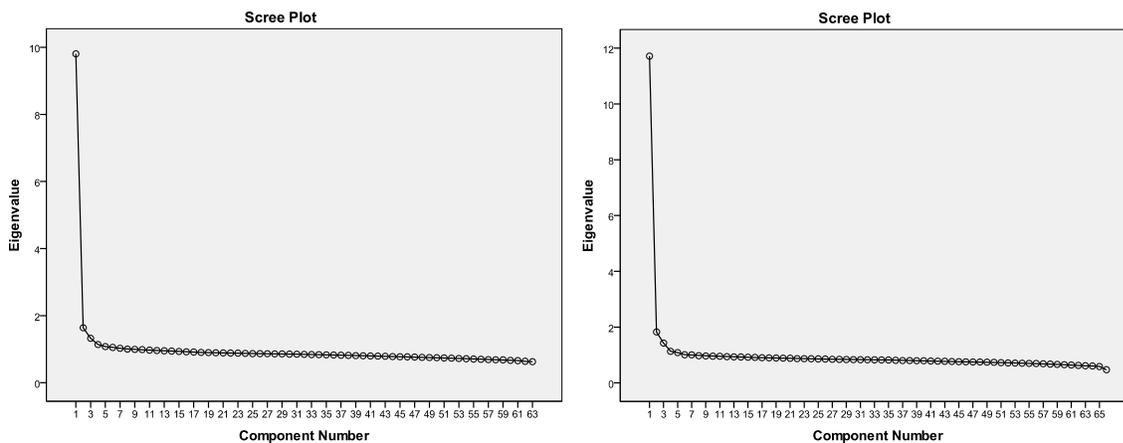
Figure 1. QQ Plots of Absolute Value of Residuals for SR Items (Left) and CR Items (Right), Grade 5 Science



**Figure 2. QQ Plots of Absolute Value of Residuals for SR Items (Left) and CR Items (Right), Grade 8 Science**

### *Principal Components Analyses*

Principal components analyses (PCA) with varimax rotation were conducted for both of the Science tests. Figure 3 shows the scree plot of eigenvalues for Grade 5 Science and for Grade 8 Science. Based on these plots, three factors were extracted by PCA. For both tests, the tables of factor loadings reveal that the items comprising the PE tend to load on the third factor, the CR items load on the first factor, and the SR items are split between the first and second factor.



**Figure 3. Scree Plot of Eigenvalues for Grade 5 Science (Left) and Grade 8 Science (Right)**

## Discussion

There were three underlying questions that guided this research.

1. What is the immediate effect of removing PEs?
2. Will the 2011 MAP scale scores be comparable to the 2010 MAP scale scores?
3. Is a new standard setting warranted given the removal of PEs?

### *Effect of Removing PEs*

Removing the PEs has the immediate effect of decreasing the test length and testing time. The results of this study show that there is very little change in the mean scale score and in the distribution of students in each achievement level. In general, the study shows a negligible amount of improvement across the grade/content areas.

There is good agreement in classification of students into achievement levels between the original and noPE conditions in all grades/content areas. Generally most students will be classified into the same achievement level on both tests. There is a small percentage of students who will change from Proficient to not Proficient and vice versa once the PE is removed. The students who lose their Proficient status tended to perform well on the PE while the students who achieved Proficient status tended to perform poorly on the PE.

The removal of the PEs does not unduly impact the test reliability statistics. Overall, the test reliability for each grade/content area is above 0.90. This is not surprising since CR items affect test reliability less than SR items do. In addition to the classical measurement of test reliability, the CSEM shows that there was little change in measurement error at key locations along the test scale.

There were a few grades where there were larger differences in the CSEM at the LOSS and/or HOSS between the original and noPE conditions. This suggests that the removal of the PE in these grades/content areas may result in less precise measurement for students at the very low and/or upper end of the test scale. In general, though, there are few students at the LOSS/HOSS and the test is designed to more precisely measure the area where the majority of the students lie (in the middle of the test scale).

Generally speaking, the psychometric criteria examined in this research suggest that the removal of the PE has little impact overall for the vast majority of Missouri students. It is important to remember that this research used the 2010 operational data. Students taking the 2010 test were administered the PE. Even though the PE was able to be removed for psychometric analysis, we could not model the way students will behave on the test when the PE is not administered. It is quite possible that student performance will improve with the removal of the PE. The PE item itself probably increases student fatigue on the test and may decrease motivation.

### *Comparability*

It is not enough that the test statistics are similar across years; it is also important that the test construct remain comparable across years. In order to make valid cross-year comparisons, the same construct must be measured from year to year. From the analyses of the percentage of items measuring each GLE/Content Standard, it appears that the same construct is generally being measured in all grades/content areas.

In Science, the Scientific Inquiry strand has less emphasis when the PE is not included in Grades 5 and 8; therefore, for these two grades/content areas, we undertook further study of the test construct through residuals analyses. In both cases, the results of the residuals analyses suggest that the measurement model is functioning the same way for the two conditions.

This analysis, however, only examined the items in common between the two conditions. It is possible that the PE measures the construct in a fundamentally different way than the remaining CR and SR items. To study this question, PCA were conducted and three factors were extracted for each Science test. Based on these analyses, the Science PE items seem to measure a different underlying construct than do CR and SR items. At the same time, the eigenvalue plot for both Grade 5 and Grade 8 Science suggest that there is one main factor for the test with two small factors.

This question has been addressed in the literature as practitioners have frequently struggled with the balance of SR to CR items. The SR items are known to be reliable, efficient, and cost-effective, and the literature suggests that, when written appropriately, they may address the same complexity of skills as do CR items (Kennedy & Walstad, 1997). Nevertheless, educators continue to perceive that CR items are a better indicator of student ability than are SR items. The analyses undertaken here suggest that even though the PE may measure the Science construct differently than do the SR and CR items, the effect is not so strong that the unidimensionality of the test is undermined. In other words, it still appears that there is one main construct underlying the test.

### *Standard Setting*

In part, DESE requested this research to anticipate the possibility of a new standard setting or cut score review. It is difficult to say with certainty if a standard setting will be warranted. The results of this research suggest that a standard setting will not be needed; however, it is not possible to anticipate the myriad ways that student scores will increase or decrease because of the removal of the PEs. It is also not possible to anticipate all the ways in which teachers will react when PEs are removed. If DESE decides that a standard setting is warranted, then it is advised that they hold this only after the test has been administered operationally and operational test data is available.

### *Conclusions*

Overall, the results of this research suggest that the MAP test results are not substantively affected by the removal of the PE items. While there are some differences, these differences do not appear to be critical. The study suggests that DESE can continue to report MAP results without the need to rescale the test or to reset standards.