

Missouri

Assessment Program
Grade-Level Assessments

Technical Report 2010

Submitted to
Missouri Department of Elementary and Secondary Education
December 2010



Developed and published by CTB/McGraw-Hill LLC, a subsidiary of The McGraw-Hill Companies, Inc., 20 Ryan Ranch Road, Monterey, California 93940-5703. Copyright © 2010 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://dese.mo.gov/divimprove/assess/index.html>.

Table of Contents

EXECUTIVE SUMMARY.....	1
E.1 Background.....	1
E.2 Administration.....	1
E.3 Student Performance.....	2
E.4 Validity and Test Scores.....	2
CHAPTER 1: INTRODUCTION.....	4
1.1 Background of the Missouri Assessment Program.....	4
1.2 Purpose of the Missouri Assessment Program.....	5
1.3 Design of the Missouri Assessment Program.....	5
1.4 Overview of this Report.....	5
CHAPTER 2: THE USES OF TEST SCORES.....	10
2.1 Uses of Test Scores.....	10
2.2 Test-Level Scores.....	11
2.2.1 Scale Scores.....	11
2.2.2 Levels of Achievement.....	11
2.2.3 Use of Test-Level Scores.....	12
2.3 Content Standard Subscores.....	12
2.3.1 Use of the Content Standard Subscores.....	12
2.4 Process Standard Subscores.....	13
2.4.1 Use of the Process Standard Subscores.....	13
CHAPTER 3: TEST CONTENT DEVELOPMENT.....	14
3.1 Test Specifications.....	14
3.2 Item Development.....	15
3.2.1 Reading Load.....	16
3.2.2 Item Writing.....	16
3.2.3 Local Pilot Test.....	17
3.2.4 Score, Revise, Rewrite Workshop.....	18
3.2.5 Content and Bias Review Workshop.....	18
3.3 Field Test Selection and Administration.....	19
3.4 Operational Test Selection.....	19
3.5 Universal Design.....	21
3.6 Accommodations.....	22
3.7 Content and Process Standards.....	22
3.8 Summary.....	24
CHAPTER 4: TEST ADMINISTRATION.....	34
4.1 Training of Districts.....	34
4.2 Ancillary Materials.....	35
4.2.1 Return Material Forms and Guidelines.....	38
4.2.2 Security Forms.....	38
4.2.3 Interpretive Guides.....	39
4.3 Test Security Measures.....	39
4.4 Test Administration.....	39
4.4.1 Time.....	39

4.4.2	Accommodations	39
4.5	Summary	40
CHAPTER 5: CONSTRUCTED-RESPONSE SCORING		59
5.1	MAP Scoring Process	59
5.1.1	Selection of Scoring Raters.....	59
5.1.2	Handscoring Training Process	60
5.1.3	Monitoring the Scoring Process.....	62
5.1.4	Security	62
5.2	Inter-Rater Reliability	63
5.3	Summary	64
CHAPTER 6: OPERATIONAL DATA ANALYSES		69
6.1	Calibration Sample.....	69
6.2	Classical Item Statistics	70
6.2.1	Test-Level Statistics.....	70
6.2.2	Item-Level Statistics	70
6.3	Item Response Theory	72
6.3.1	Model Fit.....	72
6.4	Scaling.....	75
6.4.1	Linking Methods.....	75
6.4.2	Anchor Items.....	76
6.4.3	Vertical Scale	76
6.4.4	Lowest and Highest Obtainable Scale Scores.....	77
6.5	Item-Pattern Scoring.....	78
6.6	Summary	78
CHAPTER 7: TEST RESULTS.....		109
7.1	Student Participation.....	109
7.2	Current Administration Data.....	110
7.3	Cross-year, Cross-sectional Comparisons	110
7.4	Reports	111
7.4.1	Description of Each Type of Report	112
7.5	Data Structures.....	116
7.6	Interpreting Test Results	116
7.7	Summary	117
CHAPTER 8: ACHIEVEMENT-LEVEL SETTING		127
8.1	Legislation Affecting MAP Standard Setting.....	127
8.2	Bookmark Standard Setting Procedure.....	128
8.3	Cut Scores	129
8.4	Achievement-Level Descriptors	129
8.5	Summary	129
CHAPTER 9: EVIDENCE OF CONSTRUCT-RELATED VALIDITY		131
9.1	Minimization of Construct-Irrelevant Variance and Construct Under-Representation.....	131
9.2	Reliability.....	131
9.2.1	Test Reliability.....	132
9.2.2	Standard Error of Measurement.....	133
9.2.3	Conditional Standard Error of Measurement.....	134

9.2.4	Classification Accuracy and Consistency	135
9.2.5	Convergent Validity	137
9.3	Principal Components Analysis	138
9.4	Analyses by Content Standard	138
9.4.1	Reliability of Content Standards	139
9.4.2	Correlations among Content Standard Subscores	139
9.4.3	Standard Error of Measurement of Content Standards	140
9.5	Divergent (Discriminant) Validity	140
9.6	Summary	141
CHAPTER 10: FAIRNESS		157
10.1	Minimizing Bias through Careful Test Development	158
10.2	Evaluating Bias through Differential Item Functioning Statistics	159
10.3	Evaluating Bias through Impact Analysis	162
10.3.1	Reliability	162
10.3.2	Effect Size	162
10.4	Summary	164
References		172

Table of Tables

Table E.1: Participation Rates: All Students.....	3
Table E.2: Percentage of Students Classified as <i>Proficient</i> or <i>Advanced</i> in 2006 through 2010 using Census Data: Communication Arts.....	3
Table E.3: Percentage of Students Classified as <i>Proficient</i> or <i>Advanced</i> in 2006 through 2010 using Census Data: Mathematics.....	3
Table E.4: Percentage of Students Classified as <i>Proficient</i> or <i>Advanced</i> in 2006 through 2010 using Census Data: Science.....	3
Table 1. 1: Timeline of Grade-Span MAP.....	8
Table 1. 2: Timeline of Grade-Level MAP.....	8
Table 1. 3: Number of Items that Did Not Map to a Missouri Grade-Level Expectation ..	8
Table 1. 4: Spring 2010 MAP Test Design.....	9
Table 1. 5: Spring 2010 Items Removed from Braille Forms.....	9
Table 3. 1: MAP Test Blueprint: Target Score Points by Content Standard (Communication Arts) or GLE Strand (Mathematics and Science).....	25
Table 3. 2: Elements of Universal Design.....	25
Table 3. 3: Items Omitted from the MAP Spring 2010 Braille Version.....	26
Table 3. 4: MAP 2010 Content Standard Item/Point Distributions, Communication Arts.....	27
Table 3. 5: MAP 2010 GLE Strand Item/Point Distributions, Mathematics.....	28
Table 3. 6: MAP 2010 GLE Strand Item/Point Distributions, Science.....	29
Table 3. 7: MAP 2010 Number of Items/Points Measuring Process Standards, Communication Arts.....	30
Table 3. 8: MAP 2010 Number of Items/Points Measuring Process Standards, Mathematics.....	31
Table 3. 9: MAP 2010 Number of Items/Points Measuring Process Standards, Science.....	33
Table 4. 1: MAP Administration Schedule Timing Guidelines by Session (Time in Minutes).....	42
Table 4. 2: Districts Granted a One-Week Extension of the MAP Testing Window.....	43
Table 4. 3: MAP Accommodations for Students Who Are English Language Learners ..	44
Table 4. 4: MAP Accommodations for Students with Disabilities.....	45
Table 4. 5: Number and Percent of Students Receiving Accommodations by Accommodation Type, MAP 2010 Regular Edition.....	47
Table 4. 6: Number and Percent of Students Receiving Accommodations by Accommodation Type, MAP 2010 Braille Edition.....	50
Table 4. 7: Number and Percent of Students Receiving Accommodations by Accommodation Type, MAP 2010 Large Print Edition.....	53
Table 5. 1: Inter-Rater Reliability, Communication Arts.....	65
Table 5. 2: Inter-Rater Reliability, Mathematics.....	66
Table 5. 3: Inter-Rater Reliability, Science.....	67

Table 6. 1: Large Districts that Were Included in the 80% Calibration Sample	80
Table 6. 2: Summary of Calibration and Census Data: Communication Arts.....	81
Table 6. 3: Summary of Calibration and Census Data: Mathematics.....	83
Table 6. 4: Summary of Calibration and Census Data: Science.....	85
Table 6. 5: MAP Means, Standard Deviations for Raw Scores, <i>p</i> -values, Item-Total Correlation (R_{it}): Communication Arts 2010.....	85
Table 6. 6: MAP Means, Standard Deviations for Raw Scores, <i>p</i> -values, Item-Total Correlation (R_{it}): Mathematics 2010.....	86
Table 6. 7: MAP Means, Standard Deviations for Raw Scores, <i>p</i> -values, Item-Total Correlation (R_{it}): Science 2010.....	86
Table 6. 8: Item Statistics: Grade 3.....	87
Table 6. 9: Item Statistics: Grade 4.....	89
Table 6. 10: Item Statistics: Grade 5.....	91
Table 6. 11: Item Statistics: Grade 6.....	93
Table 6. 12: Item Statistics: Grade 7.....	95
Table 6. 13: Item Statistics: Grade 8.....	97
Table 6. 14: Item Fit Statistics for Misfitting Items.....	99
Table 6. 15: LOSS and HOSS Values by Grade and Content Area	99
Table 7. 1: Participation Rates: All Students.....	118
Table 7. 2: Participation Rates: Males	118
Table 7. 3: Participation Rates: Females.....	118
Table 7. 4: Participation Rates: White	119
Table 7. 5: Participation Rates: Black.....	119
Table 7. 6: Participation Rates: Hispanic.....	119
Table 7. 7: Participation Rates: Asian/Pacific Islander	120
Table 7. 8: Participation Rates: Native American/Alaskan	120
Table 7. 9: Participation Rates: Students Receiving Accommodations.....	120
Table 7. 10: Summary Statistics for Communication Arts.....	121
Table 7. 11: Summary Statistics for Mathematics.....	121
Table 7. 12: Summary Statistics for Science	121
Table 7. 13: Comparison of State-Level Means, 2006 through 2010 Census Data.....	122
Table 7. 14: Comparison of Percentage of Students in each Achievement Level, Communication Arts 2006 through 2010 Census Data	123
Table 7. 15: Comparison of Percentage of Students in each Achievement Level, Mathematics 2006 through 2010 Census Data	124
Table 7. 16: Comparison of Percentage of Students in each Achievement Level, Science 2008 through 2010 Census Data.....	125
Table 7. 17: Summary of Score Reports for Spring 2010.....	126
Table 7. 18: Types of Reports Available to Districts through Crystal Reports	126
Table 8. 1: Communication Arts Cut Scores	130
Table 8. 2: Mathematics Cut Scores	130
Table 8. 3: Science Cut Scores	130
Table 9. 1: Reliability in Communication Arts.....	142

Table 9. 2: Reliability in Mathematics.....	142
Table 9. 3: Reliability in Science.....	142
Table 9. 4: SEM by Subgroup.....	143
Table 9. 5: Conditional Standard Error of Measurement at the Basic, Proficient, & Advanced Cut Scores.....	145
Table 9. 6: Decision Accuracy and Consistency Conditioned on Level of Achievement	146
Table 9. 7: Decision Accuracy and Consistency at Achievement Cut Points	146
Table 9. 8: Principal Component Analysis for Communication Arts.....	147
Table 9. 9: Principal Component Analysis for Mathematics.....	148
Table 9. 10: Principal Component Analysis for Science	148
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	149
Table 9. 12: Reliability (Diagonal) of Each Content Standard, Uncorrected Correlation Coefficient (below Diagonal), and Corrected Correlation Coefficient (above Diagonal) Among Content Standards: Mathematics	150
Table 9. 13: Reliability (Diagonal) of Each Content Standard, Uncorrected Correlation Coefficient (below Diagonal), and Corrected Correlation Coefficient (above Diagonal) Among Content Standards: Science.....	151
Table 9. 14: Mean, Standard Deviation, and Standard Error of Measurement (SEM) of Communication Arts Content Standards	152
Table 9. 15: Mean, Standard Deviation, and Standard Error of Measurement (SEM) of Mathematics Content Standards	153
Table 9. 16: Mean, Standard Deviation, and Standard Error of Measurement (SEM) of Science Content Standards.....	154
Table 9. 17: Inter-Correlation of Communication Arts, Mathematics, and Science Scale Scores.....	154
Table 10. 1: 2010 MAP DIF Statistics: Number of Flagged Items, Communication Arts	165
Table 10. 2: 2010 MAP DIF Statistics: Number of Flagged Items, Mathematics.....	166
Table 10. 3: 2010 MAP DIF Statistics: Number of Flagged Items, Science.....	167
Table 10. 4: Impact Analysis, Grade 3	168
Table 10. 5: Impact Analysis, Grade 4	168
Table 10. 6: Impact Analysis, Grade 5	169
Table 10. 7: Impact Analysis, Grade 6	170
Table 10. 8: Impact Analysis, Grade 7	170
Table 10. 9: Impact Analysis, Grade 8	171

Table of Figures

Figure 4. 1: Sample Script of Examiner’s Manual	56
Figure 4. 2: District Report Form	57
Figure 4. 3: Test Book Accountability Form	58
Figure 6. 1: Item characteristic curve for Grade 3 Communication Arts, Session 3 Item 34	100
Figure 6. 2: Item characteristic curve for Grade 4 Communication Arts, Session 2 Item 30	100
Figure 6. 3: Item characteristic curve for Grade 7 Communication Arts, Session 3 Item 28	100
Figure 6. 4: Item characteristic curve for Grade 8 Communication Arts, Session 2 Item 12	101
Figure 6. 5: Item characteristic curve for Grade 3 Mathematics, Session 3 Item 2.....	101
Figure 6. 6: Item characteristic curve for Grade 6 Mathematics, Session 1 Item 5.....	102
Figure 6. 7: Item characteristic curve for Grade 6 Mathematics, Session 1 Item 17.....	102
Figure 6. 8: Item characteristic curve for Grade 6 Mathematics, Session 3 Item 4.....	103
Figure 6. 9: Item characteristic curve for Grade 7 Mathematics, Session 1 Item 12.....	103
Figure 6. 10: Cross-Grade Articulation of Scale Scores at Selected Percentiles, Communication Arts MAP	104
Figure 6. 11: Cross-Grade Articulation of Scale Scores at Selected Percentiles, Mathematics MAP	105
Figure 6. 12: Cross-Grade Articulation of Scale Scores at Selected Percentiles, Science MAP	106
Figure 6. 13: Communication Arts Test Characteristic Curves by grade, 2010	107
Figure 6. 14: Mathematics Test Characteristic Curves by grade, 2010	107
Figure 6. 15: Science Test Characteristic Curves by grade, 2010	108
Figure 9. 1: SEM Plot Communication Arts, Grades 3 – 8	155
Figure 9. 2: SEM Plot Mathematics, Grades 3 – 8	155
Figure 9. 3: SEM Plot Science, Grade 5 and 8	156

EXECUTIVE SUMMARY

This report is a technical summary of the 2010 operational administration of the Missouri Assessment Program (MAP). The MAP is a grade-level test in Communication Arts and Mathematics administered in Grades 3 through 8. The MAP is a grade-span test in Science administered in Grades 5 and 8. These tests are designed to measure students' knowledge of Communication Arts, Mathematics, and Science. This section provides a summary of the 2010 Technical Report.

E.1 Background

The MAP was originally designed as grade-span tests to measure Missouri's Show-Me Standards. These standards were adopted by the Missouri State Board of Education in 1996. Since their inception, Missouri's Show-Me Standards have been further refined to better delineate Content Standards, Process Standards, and Content Strands/Grade-Level Expectations as Missouri changed its testing program to comply with the requirements of No Child Left Behind. Starting in 2006, grade-level tests were administered in Communication Arts and Mathematics. In 2008, grade-span tests were administered in Science for the first time. In 2010, MAP tests were no longer administered at the high school level. It was replaced by the Missouri End-of-Course Assessments (the technical report for these assessments may be found here: <http://dese.mo.gov/divimprove/assess/tech/>). The MAP tests have therefore undergone multiple alignment analyses to ensure that MAP content reflects these refinements. Further details of the development of the 2010 MAP may be found in Chapter 3 of this report.

E.2 Administration

In the spring of 2010, Missouri administered grade-level MAP tests in Communication Arts and Mathematics to students in Grades 3 through 8 and in Science to students in Grades 5 and 8. The MAP grade-level tests were administered from March 30 to April 24, 2010. A small portion of districts were granted a week-long extension to this testing window because the districts had been adversely affected by winter weather for an extended period of time. For these 70 districts, the test window was March 30 to May 1, 2010. Test administration is discussed in Chapter 4 of this report.

Approximately 550 districts administered Communication Arts and Mathematics MAP tests in Grades 3 through 8. These districts also administered Science MAP tests in Grades 5 and 8. Table E.1 shows participation rates based on the census data.¹ 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. The "accountable" column shows the total number of students who received a test book.

¹ The census data used in this report does not reflect additional cleaning steps that DESE staff implements once CTB releases data to DESE; therefore, the numbers in this report may differ from those in DESE reports using their cleaned data.

The “percent reportable” column shows the percentage of students who received a scale score on MAP. Further analysis of participation rates is provided in Chapter 7 of this report.

E.3 Student Performance

This is the fifth year of the grade-level MAP testing programs in Communication Arts and Mathematics and the third year for the grade-span tests in Science. Tables E.2 and E.3 present the percentage of students classified as *Proficient* or *Advanced* in 2006 through 2010 in Communication Arts and Mathematics, respectively. Table E.4 shows the percentage of students classified as *Proficient* or *Advanced* in 2008 through 2010 in Science.

For all grades and content areas, small to moderate increases in the percentage of students classified as *Proficient* or *Advanced* were observed. More information on student performance may be found in Chapter 7 of this report.

E.4 Validity and Test Scores

Most sections of this Technical Report are designed to provide validity evidence to support the use of MAP test scores. Chapter 2 discusses the uses of MAP scores. Chapter 3 discusses the test development process used to create MAP, which is important to the content-related validity of the MAP scores. Chapter 4 presents information on test administration. Chapter 5 discusses the scoring of constructed-response items, as well as the results of the inter-rater reliability studies. Chapter 6 presents the scaling and linking procedures, as well as the results of other operational data analyses. Chapter 7 reviews the results of the 2010 operational administration and overviews the score reports sent to parents, schools, and districts. Chapter 8 highlights the standard setting procedures used for MAP. Chapter 9 discusses reliability and construct-related validity. In this section, we evaluate the assumption that the content-area MAP tests are unidimensional. For example, the grade-level Mathematics MAP should measure one primary dimension (Mathematics). Chapter 10 overviews the statistical and developmental processes used to assure fairness of the MAP for all examinees. Some analyses in this document are based on the calibration sample while others are based on census data. The sources of data used for particular analyses are indicated throughout the Technical Report.

Table E.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	66947	99.7%	66947	99.8%		
4	67510	99.7%	67510	99.8%		
5	66730	99.7%	66730	99.8%	66730	99.7%
6	67476	99.7%	67476	99.8%		
7	66279	99.6%	66279	99.7%		
8	66463	99.5%	66463	99.6%	66463	99.5%

Table E.2: Percentage of Students Classified as *Proficient* or *Advanced* in 2006 through 2010 using Census Data: Communication Arts

Grade	Communication Arts					
	2006	2007	2008	2009	2010	2010 – 2009
3	42.4	42.6	40.3	40.3	43.1	2.8
4	43.8	45.1	45.1	46.3	50.9	4.6
5	45.0	47.8	48.1	48.8	51.0	2.2
6	42.2	43.6	47.4	47.7	49.6	1.9
7	42.7	44.4	49.0	50.8	51.7	0.9
8	41.5	41.6	48.1	49.7	51.8	2.1

Table E.3: Percentage of Students Classified as *Proficient* or *Advanced* in 2006 through 2010 using Census Data: Mathematics

Grade	Mathematics					
	2006	2007	2008	2009	2010	2010 – 2009
3	43.3	45.0	43.8	44.4	47.1	2.7
4	43.4	44.5	44.2	44.4	48.4	4.0
5	43.3	46.6	45.8	47.2	51.7	4.5
6	43.9	47.8	50.7	50.1	55.4	5.3
7	42.9	44.9	49.5	51.9	54.5	2.6
8	39.8	40.6	43.8	46.4	51.3	4.9

Table E.4: Percentage of Students Classified as *Proficient* or *Advanced* in 2006 through 2010 using Census Data: Science

Grade	Science			
	2008	2009	2010	2010 – 2009
5	44.5	45.1	48.9	3.8
8	43.2	44.8	48.0	3.2

CHAPTER 1: INTRODUCTION

The 2010 Missouri Assessment Program (MAP) marked the fifth administration of grade-level Communication Arts and Mathematics MAP tests in Missouri. It was the third administration of the grade-span Science MAP tests at Grades 5 and 8. The MAP is designed to measure students' knowledge of Communication Arts, Mathematics, and Science. This report provides a technical overview of the Communication Arts, Mathematics, and Science assessments of the 2010 MAP. As such, it presents evidence for the validity of the 2010 MAP scores.

This chapter of the Technical Report serves to describe the background, history, purpose, and design of the MAP, followed by an overview of the major sections for the current report.

1.1 Background of the Missouri Assessment Program

The MAP traces its origin to the 1993 Outstanding Schools Act. This act required that Missouri create a statewide assessment system that measured challenging academic standards. From this Act, grade-span assessments were created that measured Missouri's Show-Me Standards. Originally, MAP was designed to be a grade-span test: Grades 3, 7, and 11 in Communication Arts, Grades 4, 8, and 10 in Mathematics, and Grades 3, 7, and 10 in Science. Table 1.1 provides a brief timeline of the events of the grade-span MAP.

In 2001, the federal No Child Left Behind (NCLB) legislation was enacted, which required states to develop grade-level tests to be administered in Grades 3 through 8 and once in Grades 10 through 12 in both Reading and Mathematics. It also required that states have in place Science assessments to be administered at least once in Grades 3 through 5, Grades 6 through 9, and Grades 10 through 12 by the 2007–2008 school year. Based on the NCLB legislation, student performance, reported in terms of proficiency categories, is used to determine the adequate yearly progress of students at the school, district, and state levels.

In response to NCLB, the Department of Elementary and Secondary Education (DESE) contracted with CTB/McGraw-Hill in 2003 to expand the testing program to grade-level testing for Communication Arts and Mathematics. This contract was renewed in 2007 and extends through 2013. In the spring of 2005, Missouri administered a field test in Communication Arts and Mathematics, which was the basis for the construction of the 2006 and 2007 operational test forms.

The construction of the new Science MAP has been on a different trajectory. In 2005 DESE contracted with CTB/McGraw-Hill to construct a grade-span Science assessment in order to comply with the requirements of NCLB. In the spring of 2006, Missouri administered a field test in Science, which was the basis for the construction of the 2008 and 2010 operational Science forms. The contract to create grade-span Science assessments was renewed in 2007 and extends through 2013.

In 2008, DESE together with Riverside Publishing developed End-of-Course Assessments for use at the high school level. With the development of the new test program, the MAP high school assessments were discontinued. The final administration of the MAP high school assessments was in the spring of 2008.

Table 1.2 shows a timeline of the development history of the NCLB-compliant testing program.

1.2 Purpose of the Missouri Assessment Program

The MAP is designed to measure how well students acquire the skills and knowledge described in Missouri's Grade-Level Expectations (GLEs). The assessments yield information on academic achievement at the student, class, school, district, and state levels. This information is used to diagnose individual student strengths and weaknesses in relation to the instruction of the GLEs and to gauge the overall quality of education throughout Missouri.

1.3 Design of the Missouri Assessment Program

The spring 2010 MAP administration consisted of 14 operational assessments. Each form contained a norm-referenced test form from which norm-referenced scores were derived. The norm-referenced items counted toward the student scale score if they could be mapped to a Missouri GLE. If an item could not be mapped to a Missouri GLE, then it did not count toward the criterion-referenced score. Table 1.3 shows the number of items that could not be mapped to a Missouri GLE. Table 1.4 provides an overview of the 2010 MAP test design.

Braille and large print versions of each operational MAP form were constructed for each grade level/content area to enable visually impaired students to participate in MAP testing. At some grade level(s)/content area(s), it was necessary to drop items from the assessment due to difficulties associated with the Braille translation. Table 1.5 lists the number of items that were omitted from the Braille forms. Note that students taking the Braille forms were given full credit for the omitted items.

1.4 Overview of this Report

This Technical Report documents in the subsequent chapters the major activities of the testing cycle. This report provides comprehensive detail that confirms that the processes and procedures applied in the MAP adhered to appropriate professional standards and practices of educational assessment. Ultimately, this report serves to document evidence that valid inferences about Missouri student performance can be derived from the MAP. An overview of major activities documented within this report is provided in the following pages.

The Uses of Test Scores (Chapter 2)

Chapter 2 of the Technical Report discusses the concept of validity evidence. This Technical Report is comprised of evidence that supports the use of the MAP scores. In Chapter 2, we discuss some of the uses of the MAP scores.

Item and Test Development (Chapter 3)

Chapter 3 of the Technical Report provides a summary of the major test development activities that occurred to create the spring 2010 operational test forms and the materials developed to inform the public about the testing program. As each major event is presented and discussed, the role of the event in contributing to evidence for validity of the use of test results is discussed.

Test Administration (Chapter 4)

Chapter 4 of the Technical Report serves to describe the processes and activities implemented and information disseminated to help ensure standardized test administration procedures and, thus, uniform test administration conditions for students.

Scoring Constructed-Response Items (Chapter 5)

Chapter 5 of the Technical Report describes the processes and activities for scoring constructed-response items. This chapter discusses how raters are trained and the measures for assuring consistency among raters. Finally, this chapter presents the results of the inter-rater reliability studies.

Operational Data Analyses (Chapter 6)

Chapter 6 of the Technical Report includes a detailed description of the operational analyses of the 2010 MAP, which are comprised of three major parts: the calibration sample; the classical item analysis; and the calibration, scaling, and linking using item response theory (IRT) models. This chapter describes the demographics of the calibration sample and compares it to the state census data. It reports the results of the classical item analysis, as well as the results of the calibration, scaling, and linking.

Test Results and Reporting (Chapter 7)

Chapter 7 of the Technical Report contains information on the results of the spring 2010 MAP administration. Detailed summary statistics based on scale scores and achievement level information are also provided. Finally, this chapter presents information on the score reports sent to parents, schools, and districts.

Standard Setting (Chapter 8)

Chapter 8 of the Technical Report briefly discusses standard setting. It provides an overview of the standard setting activities that occurred for the MAP.

Reliability and Validity Evidence (Chapter 9)

Chapter 9 of the Technical Report provides evidence of reliability and validity of MAP scores. This chapter provides detailed results of the reliability of the tests, as well as information on the decision consistency of the cut scores. It also provides evidence of construct validity for MAP scores.

Fairness (Chapter 10)

Chapter 10 of the Technical Report discusses fairness and how the MAP tests are constructed to be fair to all Missouri students. This chapter summarizes the results of the differential item (DIF) analyses. It also discusses the results of an impact analysis to determine if large differences exist between demographic groups in Missouri.

Table 1. 1: Timeline of Grade-Span MAP

Year	Event
1996	Show-Me Standards Approved
1996	Frameworks for Curriculum Development published
1997	Annotations to the Curriculum Frameworks published
1998	First operational administration of Mathematics MAP (Grades 4, 8, and 10)
1999	First operational administration of Communication Arts MAP (Grades 3, 7, and 11) and Science MAP (Grades 4, 8, and 11)
2000	First operational administration of Social Studies MAP (Grades 4, 8, and 10)
2001	Mathematics Curriculum Supplement published
2005	Last year of grade-span MAP

Table 1. 2: Timeline of Grade-Level MAP

Year	Event
2004	Grade-Level Expectations published
2005	Communication Arts and Mathematics Field Test
2005	Standard Setting for Communication Arts and Mathematics
2006	First Operational Communication Arts and Mathematics MAP
2007	Science Field Test
2008	First Operational Science MAP
2008	Standard Setting for Science
2008	Last Operational Administration of High School MAP
2008	Version 2.0 Grade-Level Expectations (GLEs) published
2009	Last Operational Administration of MAP based on V1.0 GLEs
2010	First Operational Administration of MAP based on V2.0 GLEs

Table 1. 3: Number of Items that Did Not Map to a Missouri Grade-Level Expectation

Content	Grade	Number of Items
Communication Arts	8	1
	3	6
Mathematics	4	6
	5	10
	6	5
	7	1
	8	3
Science	5	3
	8	2

Table 1. 4: Spring 2010 MAP Test Design

Content	Grade	Anchor Items	Operational Items	Total Number of OP Items	Total Raw Score Points
Communication Arts	3	13	43	56	63
	4	13	45	58	62
	5	12	44	56	61
	6	12	44	56	60
	7	14	49	63	70
	8	13	47	60	64
Mathematics	3	12	43	55	59
	4	14	48	62	69
	5	12	46	58	62
	6	12	46	58	62
	7	13	48	61	65
	8	13	48	61	68
Science	5	13	50	63	82
	8	23	43	66	86

Table 1. 5: Spring 2010 Items Removed from Braille Forms

Content Area	Grade	Total Number of Items
Communication Arts	3	2
	8	1
Mathematics	4	2
	5	1
	8	3
Science	5	4
	8	2

CHAPTER 2: THE USES OF TEST SCORES

Validity is the overarching component of the MAP testing program. The following excerpt is from the *Standards for Educational and Psychological Testing* [American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999]:

Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system. This includes evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all examinees (17).

As stated by the *Standards*, the validity of a testing program hinges on the use of the test scores. *Validity evidence* that supports the uses of the MAP test scores is provided in this Technical Report. In this section, we examine some possible uses of the MAP test scores.

The following sections (Chapters 3 through 10) of this Technical Report provide additional evidence for these uses, as well as technical support for some of the interpretations and uses of test scores. The information in Chapters 3 through 10 also provides a firm foundation that the MAP tests measure what they are intended to measure. However, this Technical Report cannot anticipate all possible interpretations and uses of MAP scores. It is recommended that policy and program evaluation studies, in accordance with the *Standards*, be conducted to support some of the uses of the MAP scores. To this end, DESE conducted a study on consequential validity that was implemented by the Assessment Resource Center (see *MAP and Missouri Schools: A Consequential Validity Study*, ARC, 2008).

2.1 Uses of Test Scores

The validity of a test score ultimately rests on how that test score is used. To understand whether a test score is being used properly, we must first understand the purpose of the test. The intended uses of MAP scores include:

- identifying students' strengths and weaknesses on Missouri's Grade-Level Expectations
- communicating expectations for all students
- evaluating school-, district-, and/or state-level programs
- informing stakeholders (teachers, school administrators, district administrators, DESE staff members, parents, and the public) on the status of the progress toward meeting academic achievement standards of the state
- meeting the requirements to measure Adequate Yearly Progress (AYP) by NCLB
- meeting the requirements of the state's accountability program, Missouri School Improvement Program (MSIP)

This Technical Report refers to the use of several kinds of scores: the test-level scores (scale scores and achievement levels), content standard scores, and process standard scores.

2.2 Test-Level Scores

At the test level, an overall scale score that is based on student performance on the entire test is reported. In addition, an associated level of achievement is reported. These scores indicate, in varying ways, a student's achievement in Communication Arts, Mathematics, or Science. Test-level scores are reported at four reporting levels: the state, the school district, the school, and the student.

Custom-written portions of the MAP tests were directly authored by Missouri educators, edited by DESE and CTB staff, and subsequently reviewed and approved for use by Missouri educators. This procedure fosters a close relationship between the items and the Missouri Show-Me Standards from which the MAP was developed. Portions of the MAP tests from CTB's item pool were also aligned to Missouri Content Standards, Process Standards, and GLEs to further solidify the Show-Me Standards as the foundation of the MAP tests. As shown in Table 1.3 in the previous chapter, only one Grade 8 Communication Arts item, three Grade 5 Science items, and two Grade 8 Science items did not map to Missouri standards. In Mathematics, the number of items that did not map to Missouri standards ranged from one item (Grade 7) to ten items (Grade 5). Item development is described in Chapter 3; however, detailed descriptions of processes used to delineate the knowledge, skills, and abilities, including content limits and descriptions for each content area, are beyond the scope of this report.

At the test level, two types of scores are reported to indicate a student's achievement on the MAP: (1) a scale score and (2) its associated level of achievement.

2.2.1 Scale Scores

A scale score indicating a student's total performance is determined for each content area on the MAP. The overall scale score for a content area quantifies the achievement being measured by the Communication Arts, Mathematics, or Science test. In other words, the scale score represents the students' level of achievement, where higher scale scores indicate higher levels of achievement on the test and lower scale scores indicate the lower levels of achievement.

2.2.2 Levels of Achievement

A student's performance on the Communication Arts, Mathematics, or Science MAP tests is reported in one of four levels of achievement: *Below Basic*, *Basic*, *Proficient*, or *Advanced*. The cut scores for the levels of achievement were recommended by Missouri educators and citizens at the Bookmark Standard Setting Workshop in December 2005 for Communication Arts and Mathematics and in July 2008 for Science. The cut scores reflect the expectations of Missouri educators and citizens of what Missouri students should know and be able to do in each grade level(s)/content area(s). The Missouri Show-

Me Standards guided these recommendations, as did Missouri Senate Bill 1080. (See Chapter 8 of this report for a discussion of MAP standard setting.) Thus, MAP achievement levels reflect the achievement standards and abilities intended by the Missouri legislature, Missouri teachers, Missouri citizens, and DESE. Descriptions of each level of achievement in terms of what a student should know and be able to do are provided with the *Guide to Interpreting Results* (see Chapters 4 and 7).

2.2.3 Use of Test-Level Scores

MAP scale scores and achievement levels provide summary evidence of student achievement in Communication Arts, Mathematics, or Science. Classroom teachers may use these scores as evidence of student achievement in these content areas. At the aggregate level, district and school administrators may use this information for activities such as planning curriculum. At the state level, the aggregate test-level scale scores are used for accountability programs associated with NCLB and the MSIP. The results presented in this Technical Report provide evidence that the scale scores are a valid and reliable indicator of student performance in Communication Arts, Mathematics, and Science.

2.3 Content Standard Subscores

The Content Standard subscores indicate student performance in terms of the number- and percent-correct score for each Content Standard in Communication Arts and each GLE strand in Mathematics and Science. Starting in 2008, Content Standard subscores were reported only through DESE's Crystal Reporting system. These scores may be aggregated by the state, districts, or schools to determine the mean Content Standard subscores. These means may be used as indicators of the performance of the school or district in teaching students the knowledge and skills defined for each content area.

2.3.1 Use of the Content Standard Subscores

The purpose of reporting Content Standard subscores on MAP is to show for each student the relationship between the overall achievement being measured and the skills in each of the areas delimited by the Content Standards in Communication Arts and the GLE strands in Mathematics and Science. Teachers may use these subscores for individual students as indicators of strengths and weaknesses, but they are best corroborated by other evidence, such as homework, class participation, diagnostic test scores, or observation. Chapter 3 of this Technical Report provides evidence of content validity that supports the use of the Content Standard subscores. Chapter 9 of this Technical Report provides evidence of construct validity that further supports the use of the Content Standard subscores.

District and school administrators may compare their aggregate results with the state mean to better understand their strengths and weaknesses within a content area. Caution should be exercised when comparing Content Standard subscores between students or across years. The user should be aware that different items will comprise the Content Standards across years and that these items may vary in difficulty.

2.4 Process Standard Subscores

For each MAP content area, Process Standard and Content Standard subscores are determined from the same pool of items. These items were classified by the particular underlying processes used to teach each item's content, and each item's assigned Process Standard was verified by Missouri teachers in a Content Review workshop specifically designed to fulfill that purpose. Content Standard and Process Standard subscores generally show a directly proportional relationship, because the same pool of items is used to measure both sets of standards. Process Standard subscores are only reported through DESE's Crystal Reporting system.

2.4.1 Use of the Process Standard Subscores

The purpose of reporting Process Standard subscores on MAP is to show the achievement of students in each of the areas delimited by the Process Standards in Communication Arts, Mathematics, or Science. When the Process Standard processes are used to teach the content area subject matter, the Process Standard subscores can be said to reflect the strategies Missouri teachers want Missouri students to adopt in the learning and handling of "real world" activities.

Caution should be exercised when making comparisons of Process Standard subscores between students or across years. The user should be aware that different items will comprise the Process Standards across years and that these items may vary in difficulty.

CHAPTER 3: TEST CONTENT DEVELOPMENT

Content-related validity in achievement tests is evidenced by a correspondence between test content and a specification of the content domain. Content-related validity can be demonstrated through consistent adherence to test blueprints, through a high-quality test development process that includes review of items for accessibility to English Language Learners and students with disabilities, and through alignment studies performed by independent groups. In this section, we will provide a detailed discussion of the test development cycle, from aligning items with Missouri’s rigorous Show-Me Standards and GLE strands to selecting items for the final operational test form. In particular, this section will show how MAP follows rigorous procedures to construct tests that reflect the full range of content that MAP is expected to cover.

This chapter is particularly relevant to AERA, APA, & NCME (1999) Standards 3.1, 3.2, and 3.7. It also addresses Standards 3.11, 7.4, and 7.7, which will be discussed in the pertinent sections of this chapter. Standards 3.1, 3.2, and 3.7 are from Chapter 3 of the AERA, APA, & NCME (1999) Standards, which is titled *Test Development and Revision*. Each of these Standards will be presented, as will the way each Standard is addressed in this chapter. AERA, APA, & NCME (1999) Standard 3.1 says,

Tests and testing programs should be developed on a sound scientific basis. Test developers and publishers should compile and document adequate evidence bearing on test development.

The purpose of this chapter is to document the test development process used for MAP. In this chapter, we describe steps taken to create MAP tests from the development of test specifications to the selection of operational forms.

3.1 Test Specifications

AERA, APA, & NCME (1999) Standard 3.2 says,

The purpose(s) of the test, definition of the domain, and the test specifications should be stated clearly so that judgments can be made about the appropriateness of the defined domain for the stated purpose(s) of the test and about the relation of items to the dimensions of the domain they are intended to represent.

The purpose of the test is discussed in Chapter 2. MAP domains are generally defined as the knowledge and skills that are identified within the Missouri Grade Level Expectations (GLEs) and Show-Me Standards. These frameworks are, in turn, based on prior consensus among DESE, Missouri educators, and experienced subject-matter experts that the frameworks represent what is important for teachers to teach and students to learn.

Evidence of validity based on test content includes information about the test specifications, including the test design and test blueprint. Test development involves creating a design framework from the statement of the construct to be measured. The

MAP test specifications evolve from the tension between the constraints of the assessment program and the benefits sought from the examination of students. Many of the benefits sought are not scientific in nature, nor are many of the constraints; rather, they are policy considerations. The 2010 MAP item selection specifications were finalized in August 2009 prior to item selection of the operational forms.

The MAP test specifications consist of a test blueprint and a test design for each grade level/content area. The key structural aspect of the MAP tests is the test blueprint, which specifies the target score points for each Content Standard (Table 3.1). The blueprint represents a compromise between many constraints, including the target weights for each Content Standard recommended by Missouri teachers, availability of items from field testing, and results of multiple reviews by content specialists. Test design elements include such elements as number and type of items/tasks for each of the scores reported (tasks are measured by constructed-response items in MAP). The degree to which the 2010 MAP operational forms matched the test blueprint can be assessed by comparing the targeted score point distributions defined in the test blueprint with the actual point distributions displayed in Tables 3.4–3.6. Actual point distributions on the 2010 MAP operational forms matched blueprint targets within 10%, which was the tolerance for variation approved by DESE.

3.2 Item Development

Item development is discussed in this section in compliance with the AERA, APA, & NCME (1999) Standards. Standard 3.7 states,

The procedures used to develop, review, and try out items, and to select items from the item pool should be documented. If the items were classified into different categories or subtests according to the test specifications, the procedures used for classification and the appropriateness and accuracy of the classification should be documented.

Development of item content for the 2010 MAP Operational Test occurred during the period of 2004–2009. The plan specified two item development and selection cycles. The first cycle included item writing/passage selection workshops; a local pilot study; revision of items based on pilot results; content and bias reviews, item refinements, and form construction; subsequent rounds of formal field testing; selection of operational forms based on statistical data from field testing; and ultimately, operational testing at Grades 3 through 8. The second cycle excluded local pilot testing and item revisions based on pilot results. Each of these steps is described in greater detail in the following sections.

3.2.1 Reading Load

AERA, APA, & NCME (1999) Standard 7.7 is particularly relevant to item development. It says,

In testing applications where the level of linguistic or reading ability is not part of the construct of interest, the linguistic or reading demands of the test should be kept to the minimum necessary for the valid assessment of the intended construct.

MAP item development takes place within well-established content development workflow processes and methodologies. These processes include editing items for both content and style, the latter of which includes multiple reviews of each question to assure proper grammar, punctuation, and compliance to the established style. Clarity and fair access for all examinees also fall within the purview of the style reviews, which occur at scheduled milestones within the overall test development process. A thorough quality assurance review is conducted by a separate entity within the publishing division prior to the actual publication and distribution of the MAP grade-level assessments.

During item writing/content development workshops (described later), content developers are provided with specific training about how to write items that require minimal reading loads for assessing content knowledge outside of the Reading/Communication Arts content domain. For example, Mathematics content developers are trained to recognize and eliminate excessive wordiness in question stems; likewise, Science developers are encouraged to use only strictly relevant information in their items, even for those items which require some kind of background explanation of a scenario or scientific experiment.

Once item writing workshops are complete, content development editors review all item content generated at the workshops and perform a post-workshop analysis. During this process, editors reject items which do not meet specific criteria; items which do not directly assess the intended targets or cannot be modified in such a way as to comply with the established style and quality of the existing MAP items (due to excessive wordiness, linguistic complexity, or overall fair access concerns) are summarily filtered out from the pool. Then, only the remaining material is submitted to a thorough style review.

The established MAP content development workflow calls for style reviews to occur at other milestones which include (but are not limited to) pilot testing, formal content and bias reviews, and form selection. Style reviews also occur after the results of the Score, Revise, Rewrite workshops.

3.2.2 Item Writing

Communication Arts and Mathematics: In February 2005 and January 2007, groups comprised of Missouri educators, Regional Instructional Facilitators (RIFs), DESE staff, and CTB personnel participated in Item Writing Workshops (IWWs) for Communication Arts and Mathematics at Lake of the Ozarks, Missouri. The workshops were conducted

with more than 30 teacher participants per content area. Teacher participants were selected by DESE to represent educational sites throughout Missouri. During the first day of the workshop, Communication Arts participants selected reading passages. During the next three days, Communication Arts participants used selected passages as bases for writing constructed-response (CR) items and writing prompts for the 2010 Operational forms for Grades 3–8. The Mathematics participants wrote CR items and performance-event (PE) items along with scoring guides to create a pool of items for the 2010 Operational forms for Grades 3–8. The content developed at the workshop was based specifically upon the Missouri Show-Me Standards and GLEs. Some selected-response (SR) items were developed by CTB after the workshops to help supplement the item pool and reviewed by DESE. Items were refined after the initial item writing workshop which led to the production of local pilot test forms.

Science: In November 2004, a group comprised of Missouri educators, RIFs, DESE staff, and CTB personnel participated in a four-day Science IWW in Columbia, Missouri. The IWW was conducted with 37 teacher participants selected by DESE on the basis of their prior experience and expertise in item development for MAP Science and to represent educational sites throughout Missouri. The purpose of the IWW was to revise existing items and write new items to ensure a well-balanced item pool for the 2010 MAP Science operational tests. The existing items came from the MAP Science item pool previously developed for operational testing at Grades 3 and 7. During the first two days of the IWW, the existing items were revised to target the new MAP Science GLEs. These new GLEs were the basis for the 2010 assessment to be administered at Grades 5 and 8. During the third and fourth days of the IWW, Science participants wrote new CR items and performance events. A new MAP Science performance event development template was introduced at the IWW. This template specified the types of tasks and numbers of items that comprise a performance event. Science item development was also included in the January 2007 IWW at Lake of the Ozarks and followed the same methods described for the 2004 IWW.

Overall, the IWWs in November 2004, February 2005, and January 2007 provided a basis upon which items written for the Communication Arts, Mathematics, and Science grade-level assessments could be selected for use on small-scale local pilot tests administered throughout Missouri.

3.2.3 Local Pilot Test

In March 2005 (Science) and November 2005 (Communication Arts and Mathematics), small-scale pilot tests were administered in a limited number of classrooms throughout Missouri. These pilot tests consisted of items from the November 2004 (Science) and February 2005 (Communication Arts and Mathematics) IWWs. Teachers who administered the pilot tests were selected by DESE from the pool of IWW participants. The items from the 2007 IWW were not subjected to local pilot testing.

Six Communication Arts forms per grade were piloted, consisting of approximately two SR items and six CR items each for Grades 4, 5, 6, and 8. The six Communication Arts pilot forms for Grades 3 and 7 each contained two SR items, four CR items, and one

writing prompt. Six Mathematics forms per grade were piloted, consisting of approximately twelve SR items and two CR items each for grades 3, 5, 6, and 7. The six Mathematics pilot forms for Grades 4 and 8 each contained twelve SR items, four CR items, and one PE. Ten Science forms per grade, consisting of approximately fifteen CR items, were piloted for each of Grades 5 and 8. In addition to these ten pilot forms, eight PEs were piloted at each grade level.

3.2.4 Score, Revise, Rewrite Workshop

In April 2005 (Science) and February 2006 (Communication Arts and Mathematics), the items included in the 2005 local pilot test underwent further evaluation during Score, Revise, and Rewrite (SRR) Workshops. The items from the 2007 Item Writing Workshop were not subjected to the SRR Workshops.

The purpose of the SRR Workshop was for the participants to score the items piloted in Missouri classrooms and to revise the items and rubrics/scoring guides based on the scoring process, student results, and subsequent discussion. DESE invited approximately 5 to 7 participants per grade level(s)/content area(s), resulting in the direct participation of approximately 100 Missouri educators in this step of the development process. CTB and DESE personnel were present to facilitate the SRR Workshop. The participants individually scored the students' pilot forms, tallied the results, and then reviewed the items as a group. RIFs were also present and participated in the process. Overall, the goal of the workshop was to improve the item quality prior to the next step in the process, Content and Bias Review, and to ensure that quality items were developed for future use in the MAP. Most participants commented that this workshop was successful in this regard.

3.2.5 Content and Bias Review Workshop

Content and Bias Review (CBR) workshops were conducted in May 2005 (Science), May 2006 (Communication Arts and Mathematics), and June 2007 (all three content areas). DESE staff, Missouri educators, RIFs, and CTB staff participated in all meetings. The 2005 and 2006 CBRs were conducted in Columbia, Missouri, and the 2007 CBR was conducted in Jefferson City, Missouri. All three CBRs followed the same procedures. For the Content Review, DESE invited participants from educational sites throughout Missouri to review items, writing prompts, performance events, and scoring guides for content accuracy and grade level appropriateness. In Communication Arts, participants also reviewed passages. In addition, participants in all three content areas verified each item's alignment to the Missouri curriculum by reviewing the Content Standard, Process Standard, and GLE assignment. The Content Review was accomplished over the course of one or two days, and was followed by a one- or two-day Bias Review.

The Bias Review committee was comprised of representatives from various backgrounds whose purpose was to screen the items for any racial, socioeconomic, gender, or other sensitivity issues. This follows AERA, APA, & NCME (1999) Standard 7.4, which states,

Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the domain.

For each content area, over 30 Missouri educators participated in the process to help ensure content validity and screen items for potential bias. Review committees could revise or reject items because of issues related to possible bias. Greater than 90% of reviewed items were accepted by each review committee at each of the three CBRs. The general consensus was that the items as a group were well written and edited, and that the revisions made during and after the SRR Workshop had contributed to a smooth CBR workshop. The accepted items became candidates for the next step in the process, the MAP field test.

3.3 Field Test Selection and Administration

The items approved by CBR committees became the basis for the formation of stand-alone field test forms administered in 2006 and 2007 and embedded field testing in 2008 and 2009. The custom-written material was arranged into test forms using *TerraNova* Survey as a common anchor across forms. (The same anchor was used as the norm-referenced test [NRT] portion of the 2008 and 2009 operational tests; a new *TerraNova* anchor was field tested in 2009 for operational use in 2010. The anchor design is described in more detail in the following section.) Field test items were selected and placed into forms so that the combined coverage of the NRT and custom portions of the test met the established blueprint requirements for content coverage; each field test form was constructed using the same design.

The MAP spring 2006 Science field test consisted of four parallel forms per grade level, which were administered in Grades 5 and 8 in May 2006. The MAP spring 2007 Communication Arts and Mathematics field tests consisted of six parallel forms per grade level/content area which were successfully administered in Grades 3–8 in May 2007. All field test forms were reviewed and approved by DESE prior to administration. The field tests generated item statistics that were used to select two years of parallel operational forms to be administered in 2008 and 2009. Due to budget constraints, not all of the CR items field-tested in 2009 were scored; therefore, some of the items used operationally in 2007 and 2008 were made available for operational selection in 2010.

3.4 Operational Test Selection

The use of an embedded *TerraNova* Survey provides an NRT subtest, which is a requirement of the MAP. For most grade level(s)/content area(s), the intact *TerraNova*

Survey Form E was embedded in the 2010 MAP tests. Due to equating issues with Form E, *TerraNova* Survey Form D was used as the NRT subtest for Grade 8 Science.

The use of the *TerraNova* Survey and its match to the Missouri standards plays an important role in planning for the entire development process leading up to the time of item selection. This is because the test blueprint is applied to the entire test, which includes both the NRT and custom portions. As an NRT product, *TerraNova* items are pre-classified to an existing set of *TerraNova* Reading, Language, Mathematics, or Science standards.² In many cases, the match of *TerraNova* items to the Missouri GLEs could be considered equivalent; nevertheless, the item development process provided for a DESE review of how the items in the *TerraNova* Survey were matched to the Missouri Standards. The match of *TerraNova* Survey Form E items to Missouri standards was initially assessed by DESE in 2008 and then verified by DESE in September 2009 in preparation for the 2010 MAP test.

Operational item selections for 2010 were performed in August and September of 2009 by CTB. The selection process followed strict statistical criteria specified by CTB's Research department and approved by DESE. The selection criteria were based on both content requirements and statistical criteria, including the following:

1. *TerraNova* Survey Form E is the NRT subtest for all grades and content areas, with the exception of Grade 8 Science.
2. Test length and item types match the DESE-approved test design.³
3. Content coverage matches DESE-approved test blueprint.
4. The following items were to be avoided, whenever possible:
 - a. For CR items, 3+ point items where more than 50% were able to attain the top score points.
 - b. p -value ≤ 0.20 or ≥ 0.90
 - c. Omit rates $\geq 5\%$
 - d. Poor Fit statistics (Q1)
 - e. Significant DIF statistics:
 - i. If an item with DIF had to be included for blueprint coverage, examine the item to determine if any content reason exists for the DIF flag (sometimes items will demonstrate statistical bias but no content reason can be determined for the bias).
 - ii. Obtain DESE permission to use the DIF item (meaning someone from DESE should examine the item and agree that no content reason can be determined for the statistical bias).
5. Statistical properties of the test:
 - a. ITEMWIN software must be used to select forms.

² It's important to note that the Communication Arts MAP is comprised of both Reading and Language items that are scaled together. In the *TerraNova* family of tests, Reading and Language are administered in a single booklet but are scaled separately.

³ Due to DESE budget constraints, the 2010 test design contained a higher proportion of SR items than the 2006–2009 MAP tests.

- b. The Standard Error of Measurement (SEM) and Test Characteristic Curve (TCC) of the 2010 operational test must match within 5% of the 2009 MAP.

Production of the 2010 operational test forms and ancillary materials commenced in October 2009. Items were ordered and placed into test books in preparation for operational testing, and the standard process of page reviews between CTB and DESE ensued until final approvals were in place in December 2009. Then, test books and ancillary materials were printed and distributed in support of the spring 2010 operational test.

3.5 Universal Design

Grade-level assessments that are universally designed allow participation of the widest possible range of students, resulting in more valid inferences about students' performance. Universally designed grade-level assessments may reduce the need for accommodations by reducing or eliminating access barriers associated with the tests themselves. Table 3.2 presents the elements of Universal Design (Thompson & Thurlow, 2002). The elements of Universal Design are relevant to both item development and form construction. This section addresses how the elements of Universal Design were addressed in the construction of the spring 2010 test forms.

Universal Design requires that grade-level assessments need to measure the performance of students with a wide range of abilities and skill repertoires, ensuring that students with diverse learning needs receive opportunities to demonstrate competence on the same content. To accommodate the greatest number of students within MAP, the regular print assessment includes simple, clear, and intuitive instructions and procedures, maximum readability and comprehensibility, and maximum legibility. All of these design components are addressed primarily through the physical layout and formatting of the test books. The page specifications and template for test book pages define how directions and test items are placed on the pages, the location and appearance of headers and footers, spacing between an item stem and answer choices, and other page elements to ensure a consistent, legible appearance of printed test books. Written instructions in the test books at the beginning of each test session are clearly and simply stated, and the wording of such instructions is standardized as much as possible across content areas and grade levels to ensure clarity and consistency.

The MAP test books are designed to minimize distractions and to support navigation through the test book. In Grade 3 Communication Arts, the test items are read aloud to the students. In all grade levels and content areas, a "full-page stop" at the end of each testing session indicates that the students cannot turn the page until instructed by the test examiner. Right-facing pages within a session have a "go on" arrow at the bottom right-hand corner to indicate that the test session continues on the next page. Any pages that are intentionally left blank are labeled "Do Not Mark on this Page" to indicate that there are no test materials on that page.

3.6 Accommodations

Students with disabilities or who are English Language Learners may be provided test administration accommodations based on their Individualized Education Plan (IEP). More information on accommodations can be found in Section 4.4.2 of Chapter 4. Accommodation code definitions can be found on the DESE website at:

<http://www.dese.mo.gov/divimprove/assess/special.html>.

Braille and large print versions were constructed for each grade level(s)/content area(s) to enable visually-impaired students to participate in MAP testing. Specific recommendations on how to transcribe items into Braille were provided by an independent Braille expert who collaborated with the Braille publisher to produce the Braille version of the MAP and teacher's notes that accompany the Braille form. DESE conducted a review meeting with a committee of teachers in January 2010 to ensure that both the Braille and large print versions of the 2010 MAP assessment would be accessible to Missouri's visually-challenged students. DESE and the teacher committee made recommendations, as needed, for how to further revise the transcription to best serve the needs of visually-challenged students.

While the goal is to maximize the number of items on the Braille form, it was not possible to transcribe all items into Braille, as some items represent concepts that are simply not appropriate for students who take the Braille form. At some grade level(s)/content area(s), it was necessary to omit items from the Braille version due to bias issues or excessive difficulty associated with the Braille transcription. Table 3.3 lists the items that were omitted from the 2010 Braille versions. The concerns noted by the committee for items that were dropped from the Braille form are brought to the attention of assessment editors and item writers to guide future item development.

3.7 Content and Process Standards

Test content evidence of validity is provided for the MAP with the specification of each of the Content and Process Standards that are influential in acquiring the skills tested in the items/tasks used in each of the MAP tests. If teachers teach using the Content and Process Standards as intended, then student performance should improve on those items that were identified as implicitly tapping these habits of mind and/or explicitly written and clearly intended to measure specific Content Standards.

AERA, APA, & NCME (1999) Standard 3.11 says,

Test developers should document the extent to which the content domain of a test represents the defined domain and test specifications.

The 2010 MAP assessed version 2.0 of the Missouri GLEs for the first time. Prior to selecting the operational test, CTB and DESE performed an in-depth comparison of the version 2.0 GLEs against the former version in place since 2005 (Communication Arts and Mathematics) and 2006 (Science). This comparison was conducted beginning in early 2008 through the approval of the 2010 MAP test specifications. The analysis

included an alignment of the entire MAP item pool to the version 2.0 GLEs, which was reviewed and approved by DESE. The results of the comparison found that the changes to the content domain between the original GLEs and version 2.0 GLEs were limited in scope. A small number of GLEs that were formerly tested were no longer assessable on the statewide test but still present in the curriculum (denoted as “locally assessed”) and a small percentage of the Mathematics GLEs were reclassified to new grade levels. These changes caused the loss of some items from the MAP item pool and resulted in the need to reuse operational items from 2008 for the 2010 MAP. However, the Content Standards/GLE strands used as reporting categories in the 2006–2009 MAP remained intact across grades/content areas in the version 2.0 GLEs. The conclusion from the comparison between the former GLEs and the version 2.0 GLEs was that the same overall content domain would be measured by the 2010 MAP that was measured by the former version (2006–2009).

Between test selection and administration of the 2010 MAP, DESE contracted an independent study to evaluate the alignment of the test forms to the version 2.0 GLEs. The study was conducted in October 2009 by the Human Resources Research Organization (HumRRO) along with Dr. Norman Webb as a subcontractor. The alignment study examined four alignment criteria:

- (1) Categorical concurrence—determines the degree of overall content coverage by the assessment for each content strand.
- (2) Range-of-knowledge representation—indicates the specific content expectations (e.g., standard, GLE) assessed within each strand.
- (3) Balance-of-knowledge representation—provides a statistical index reflecting the distribution of assessed content within each strand (i.e., how evenly the content is assessed).
- (4) Depth-of-knowledge consistency—compares the cognitive complexity ratings of the items with the complexity ratings of each content standard.

The results of the alignment study suggested there were some alignment deficiencies in the 2010 MAP test forms for Communication Arts (Depth-of-Knowledge and Balance-of-knowledge) and Science (Range-of-Knowledge). The Depth-of-Knowledge deficiency in Communication Arts is attributed mainly to the reliance on SR items, which contribute an average of 90% of the total score points on the test. The Balance-of-Knowledge deficiency in Communication Arts is attributed to a historical tendency for item writers to focus on a limited number of GLEs. New items targeting GLEs not traditionally tested in Communication Arts were written during an IWW in 2008, but those items were not field tested due to DESE budget constraints. The Range-of-Knowledge deficiency in Science is mainly attributed to a large number of GLEs at each grade level (149 and 219 assessable GLEs at Grade 5 and Grade 8, respectively). The Science test would need to include many more test items to cover at least 50% of the GLEs, which is the standard to meet the Range-of-Knowledge criterion. With future item and test development, DESE and CTB are committed to implementing the recommendations of the external alignment

study. These recommendations include broadening the scope of item development so that more GLEs can be tested; increasing the cognitive complexity of new test items; and reducing the number of Science GLEs so that a greater proportion can be tested each year.

Table 3.4 provides the distribution of items and points on the 2010 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.

3.8 Summary

In summary, the overall purpose of this chapter is to explicate the procedures used in the development of the MAP grade-level assessments. The efforts by DESE and CTB/McGraw-Hill in developing the MAP address multiple best practices of the test industry but in particular are related to the following AERA, APA, & NCME (1999) Standards:

- Standard 3.1 — Tests and testing programs should be developed on a sound scientific basis. Test developers and publishers should compile and document adequate evidence bearing on test development.
- Standard 3.2 — The purpose(s) of the test, definition of the domain, and the test specifications should be stated clearly so that judgments can be made about the appropriateness of the defined domain for the stated purpose(s) of the test and about the relation of items to the dimensions of the domain they are intended to represent.
- Standard 3.7 — The procedures used to develop, review, and try out items, and to select items from the item pool should be documented. If the items were classified into different categories or subtests according to the test specifications, the procedures used for classification and the appropriateness and accuracy of the classification should be documented.
- Standard 3.11 — Test developers should document the extent to which the content domain of a test represents the defined domain and test specifications.
- Standard 7.4 — Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the domain.
- Standard 7.7 — In testing applications where the level of linguistic or reading ability is not part of the construct of interest, the linguistic or reading demands of the test should be kept to the minimum necessary for the valid assessment of the intended construct.

Table 3. 1: MAP Test Blueprint: Target Score Points by Content Standard (Communication Arts) or GLE Strand (Mathematics and Science)

Content Area Content Standard/ GLE Strand	Grade					
	3	4	5	6	7	8
Communication Arts						
Speaking/Writing Standard English	14	9	11	12	15	14
Reading—Fiction & Nonfiction	44	50	48	48	46	50
Writing Formally & Informally	6	2	2	1	7	1
Mathematics						
Algebraic Relationships	14	14	15	12	20	24
Data and Probability	6	7	11	16	11	14
Geometric and Spatial Relationships	12	11	11	9	12	17
Measurement	11	14	11	9	9	7
Number and Operations	21	25	18	19	15	7
Science						
Matter and Energy			11			12
Force and Motion			8			7
Living Organisms			8			11
Ecology			9			8
Earth Systems			10			12
Universe			9			9
Scientific Inquiry			21			25
Science, Technology, and Human Activity			7			6

Table 3. 2: Elements of Universal Design

Element	Explanation
Inclusive Assessment Population	Tests designed for state, district, or school accountability must include every student except those in the alternate assessment, and this is reflected in assessment design and field testing procedures.
Precisely Defined Constructs	The specific constructs tested must be clearly defined so that all construct irrelevant cognitive, sensory, emotional, and physical barriers can be removed.
Accessible, Non-Biased Items	Accessibility is built into items from the beginning, and bias review procedures ensure that quality is retained in all items.
Amenable to Accommodations	The test design facilitates the use of needed accommodations (e.g., all items can be Brailled).
Simple, Clear, and Intuitive Instructions and Procedures	All instructions and procedures are simple, clear, and presented in understandable language.
Maximum Readability and Comprehensibility	A variety of readability and plain language guidelines are followed (e.g., sentence length and number of difficult words are kept to a minimum) to produce readable and comprehensible text.
Maximum Legibility	Characteristics that ensure easy decipherability are applied to text, to tables, figures, and illustrations, and to response formats.

Table 3. 3: Items Omitted from the MAP Spring 2010 Braille Version

Grade	Content Area	Type	Session	Item
3	Communication Arts	CR	1	4
		SR	3	3
4	Mathematics	PE	1	22
		CR	3	4
5	Mathematics	CR	3	3
	Science	CR	1	1
		SR	2	19
		SR	2	23
CR	3	3		
8	Communication Arts	SR	2	30
	Mathematics	SR	1	17
		SR	2	17
		CR	3	2
	Science	SR	2	4
CR		3	6	

Table 3. 4: MAP 2010 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	0		23	23		23	37%
	Reading Nonfiction	7	7	4	18	14	8	22	35%
	Speaking/Writing Standard English	0	12		12	12		12	19%
	Writing Formally & Informally	0	0	3	3	0	6	6	10%
	Combined Reading from Standards 2 & 3	30	7	4	41	37	8	45	71%
	Total	30	19	7	56	49	14	63	100%
4	Reading Fiction/Poetry/Drama	15	5		20	20		20	32%
	Reading Nonfiction	18	2	4	24	20	8	28	45%
	Speaking/Writing Standard English	0	12		12	12		12	19%
	Writing Formally & Informally	0	0	2	2	0	2	2	3%
	Combined Reading from Standards 2 & 3	33	7	4	44	40	8	48	77%
	Total	33	19	6	58	52	10	62	100%
5	Reading Fiction/Poetry/Drama	16	2	4	22	18	8	26	43%
	Reading Nonfiction	16	5		21	21		21	34%
	Speaking/Writing Standard English	0	12		12	12		12	20%
	Writing Formally & Informally	0	0	1	1	0	2	2	3%
	Combined Reading from Standards 2 & 3	32	7	4	43	39	8	47	77%
	Total	32	19	5	56	51	10	61	100%
6	Reading Fiction/Poetry/Drama	15	2	4	21	17	8	25	42%
	Reading Nonfiction	18	4		22	22		22	37%
	Speaking/Writing Standard English	0	12		12	12		12	20%
	Writing Formally & Informally	0	0	1	1	0	1	1	2%
	Combined Reading from Standards 2 & 3	33	6	4	43	39	8	47	78%
	Total	33	18	5	56	51	9	60	100%
7	Reading Fiction/Poetry/Drama	13	7	4	24	20	8	28	40%
	Reading Nonfiction	20	0		20	20		20	29%
	Speaking/Writing Standard English	0	16		16	16		16	23%
	Writing Formally & Informally	0	0	3	3	0	6	6	9%
	Combined Reading from Standards 2 & 3	33	7	4	44	40	8	48	69%
	Total	33	23	7	63	56	14	70	100%
8	Reading Fiction/Poetry/Drama	15	2	4	21	17	8	25	39%
	Reading Nonfiction	17	4		21	21		21	33%
	Speaking/Writing Standard English	0	16		16	16		16	25%
	Writing Formally & Informally	0	0	2	2	0	2	2	3%
	Combined Reading from Standards 2 & 3	32	6	4	42	38	8	46	72%
	Total	32	22	6	60	54	10	64	100%

Table 3. 5: MAP 2010 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
3	Algebraic Relationships	4	6	1	11	10	2	12	20%
	Data and Probability	3	1	1	5	4	2	6	10%
	Geometric and Spatial Relationships	4	8		12	12		12	20%
	Measurement	1	7	1	9	8	2	10	17%
	Number and Operations	12	5	1	18	17	2	19	32%
	Total	24	27	4	55	51	8	59	100%
4	Algebraic Relationships	5	7	1	13	12	2	14	20%
	Data and Probability	4	1	1	6	5	2	7	10%
	Geometric and Spatial Relationships	2	6	1	9	8	2	10	14%
	Measurement	3	7	1	11	10	4	14	20%
	Number and Operations	12	10	1	23	22	2	24	35%
	Total	26	31	5	62	57	12	69	100%
5	Algebraic Relationships	5	7	1	13	12	2	14	23%
	Data and Probability	2	8		10	10		10	16%
	Geometric and Spatial Relationships	2	7	1	10	9	2	11	18%
	Measurement	3	6	1	10	9	2	11	18%
	Number and Operations	10	4	1	15	14	2	16	26%
	Total	22	32	4	58	54	8	62	100%
6	Algebraic Relationships	5	5	1	11	10	2	12	19%
	Data and Probability	4	10		14	14		14	23%
	Geometric and Spatial Relationships	4	3	1	8	7	2	9	15%
	Measurement	1	6	1	8	7	2	9	15%
	Number and Operations	12	4	1	17	16	2	18	29%
	Total	26	28	4	58	54	8	62	100%
7	Algebraic Relationships	5	13	1	19	18	2	20	31%
	Data and Probability	5	4	1	10	9	2	11	17%
	Geometric and Spatial Relationships	6	4	1	11	10	2	12	18%
	Measurement	1	5	1	7	6	2	8	12%
	Number and Operations	14	0		14	14		14	22%
	Total	31	26	4	61	57	8	65	100%
8	Algebraic Relationships	5	12	1	18	17	2	19	28%
	Data and Probability	4	3	2	9	7	6	13	19%
	Geometric and Spatial Relationships	4	10	1	15	14	2	16	24%
	Measurement	2	3	1	6	5	2	7	10%
	Number and Operations	13	0		13	13		13	19%
	Total	28	28	5	61	56	12	68	100%

Table 3. 6: MAP 2010 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	10%
	Earth's Processes	2	3	2	7	5	4	9	11%
	Force and Motion	0	2	3	5	2	6	8	10%
	Interactions of Organisms	3	2	2	7	5	4	9	11%
	Matter and Energy	6	1	2	9	7	4	11	13%
	Scientific Inquiry	6	2	8	16	8	14	22	27%
	Technology and the Environment	2	3	1	6	5	2	7	9%
	The Universe	1	3	2	6	4	4	8	10%
	Total	22	20	21	63	42	40	82	100%
8	Characteristics of Living Organisms	3	0	3	6	3	6	9	10%
	Earth's Processes	5	1	2	8	6	4	10	12%
	Force and Motion	3	2	1	6	5	2	7	8%
	Interactions of Organisms	2	3	1	6	5	2	7	8%
	Matter and Energy	2	4	3	9	6	6	12	14%
	Scientific Inquiry	7	3	9	19	10	15	25	29%
	Technology and the Environment	1	3	2	6	4	4	8	9%
	The Universe	0	4	2	6	4	4	8	9%
	Total	23	20	23	66	43	43	86	100%

Table 3. 7: MAP 2010 Number of Items/Points Measuring Process Standards, Communication Arts

Grade Level	Process Standard	NRT Items	SR Items	CR Items	Total Items	SR Points	CR Points	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			3	3		6	6
	2.2		12		12	12		12
	2.4	1			1	1		1
	3.5	5	2		7	7		7
4	1.5	4	1		5	5		5
	1.6	20	2	4	26	22	8	30
	2.1			2	2		2	2
	2.2		12		12	12		12
	2.4	1			1	1		1
	3.5	8	4		12	12		12
5	1.4	1			1	1		1
	1.5	8	2		10	10		10
	1.6	14			14	14		14
	2.1			1	1		2	2
	2.2		12		12	12		12
	2.4		1		1	1		1
	3.1		1		1	1		1
	3.4	1			1	1		1
	3.5	8	3	4	15	11	8	19
6	1.5	12	3		15	15		15
	1.6	13	2	1	16	15	2	17
	1.8			1	1		1	1
	2.2		12		12	12		12
	2.4	1			1	1		1
	3.1			1	1		2	2
	3.5	7	1	2	10	8	4	12
7	1.5	6	1		7	7		7
	1.6	21	2	1	24	23	2	25
	2.1		1	3	4	1	6	7
	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
8	1.5	4	4		8	8		8
	1.6	21	1	1	23	22	2	24
	2.1		2	2	4	2	2	4
	2.2		14		14	14		14
	2.4	1			1	1		1
	3.5	6	1	3	10	7	6	13

Table 3. 8: MAP 2010 Number of Items/Points Measuring Process Standards, Mathematics

Grade Level	Process Standard	NRT Items	SR Items	CR Items	Total Items	SR Points	CR Points	Total Points
3	1.10	15	11	1	27	26	2	28
	1.6	2	11		13	13		13
	2.1		1		1	1		1
	3.1			2	2		4	4
	3.2	7	2		9	9		9
	3.3		1	2	3	1	4	5
	3.6		2		2	2		2
4	1.1		1		1	1		1
	1.10	11	5	1	17	16	4	20
	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
	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	
5	1.10	11	2		13	13		13
	1.2		2		2	2		2
	1.5		2		2	2		2
	1.6	5	12	3	20	17	6	23
	1.7		1		1	1		1
	3.1		4		4	4		4
	3.2	5			5	5		5
	3.3		4	1	5	4	2	6
	3.5	1	2		3	3		3
	3.6		3		3	3		3
4.1			1	1		2	2	
6	1.10	8	4		12	12		12
	1.2		1		1	1		1
	1.5		5		5	5		5
	1.6	3	5	2	10	8	4	12
	1.8		2		2	2		2
	3.1	5	2		7	7		7
	3.2	6	3		9	9		9
	3.3	4	2		6	6		6
	3.4		1		1	1		1
	3.5		3		3	3		3
	3.6			1	1		2	2
4.1			1	1		2	2	
7	1.10	9	2		11	11		11
	1.5		2		2	2		2
	1.6	4	14	1	19	18	2	20
	1.8	2			2	2		2
	3.1	6			6	6		6
	3.2	7		1	8	7	2	9
	3.3	3	4	1	8	7	2	9
3.4		3		3	3		3	

Table 3. 8: MAP 2010 Number of Items/Points Measuring Process Standards, Mathematics (Cont'd)

Grade Level	Process Standard	NRT Items	SR Items	CR Items	Total Items	SR Points	CR Points	Total Points
7	3.5		1		1	1		1
	3.6		1		1	1		1
	3.7			1	1		2	2
8	1.10	4	1	1	6	5	2	7
	1.5			1	1		2	2
	1.6	3	8	2	13	11	4	15
	1.8	1	3		4	4		4
	3.1	6	2		8	8		8
	3.2	4	5		9	9		9
	3.3	7	7		14	14		14
	3.5	1	1	1	3	2	4	6
	3.6	1	1		2	2		2
	3.8	1			1	1		1

Table 3. 9: MAP 2010 Number of Items/Points Measuring Process Standards, Science

Grade Level	Process Standard	NRT Items	SR Items	CR Items	Total Items	SR Points	CR Points	Total Points
5	1.1			2	2		2	2
	1.10	12	15	5	32	27	10	37
	1.3	2		2	4	2	4	6
	1.5	5	1	2	8	6	4	10
	1.6	3	4	8	15	7	15	22
	1.8			1	1		4	4
	3.5			1	1		1	1
8	1.1			1	1		1	1
	1.10	16	15	7	38	31	14	45
	1.3	1		5	6	1	7	8
	1.5	3			3	3		3
	1.6	3	4	6	13	7	12	19
	1.7			1	1		1	1
	1.8			1	2	3	1	6
3.8				1	1		2	2

CHAPTER 4: TEST ADMINISTRATION

Chapter 4 of the Technical Report describes the processes and activities implemented and information disseminated to help ensure standardized test administration procedures and, thus, uniform test administration conditions for students. According to the AERA, APA, & NCME *Standards* (1999), the “usefulness and interpretability of test scores require that a test be administered and scored according to the developer’s instructions” (61). Chapter 4 examines how test administration procedures implemented for the MAP strengthen and support the intended score interpretations and reduce construct-irrelevant variance that could threaten the validity of score interpretations.

Chapter 4 demonstrates adherence to AERA, APA, & NCME (1999) Standards 3.19, 3.20, 5.1, 5.2, 5.3, 5.4, 5.6, and 5.7 in the MAP. Each Standard will be explicated within the relevant section of this chapter.

4.1 Training of Districts

To ensure that the MAP’s grade-level assessments are administered and scored in accordance with the department’s mandates, the Department takes a primary role in communicating with and training district personnel. The development of the grade-level assessments is a collaborative effort between the Department and CTB/McGraw-Hill. The Department conveys to districts the purpose of the grade-level assessments and the importance of test administration being consistent with test industry standards. The tests and the standard administration practice must also meet the State Board of Education policies and the mandates of both state and federal legislation.

To accomplish these goals, the Department provides train-the-trainer opportunities for the RIFs who, in turn, convey test administration training to districts. The RIFs also conduct Quality Assurance visits during testing to ensure district adherence to the standardized administration of the tests.

The RIFs are responsible to the districts within their region. They disseminate information to each district, offer assistance with test administration, and serve as the liaisons between the Department and the districts. The Department also communicates directly with districts, answering questions particular to the Grade-Level Assessment, as well as general assessment questions. The Department also provides assistance with and interpretation of Grade-Level Assessment data and test results.

The Assistant Director of Assessment trained the RIFs in the following components of Grade-Level Assessment administration: the *Test Coordinator’s Manual*; the *Examiner’s Manual*; the dates for testing; appropriate protocols for test administration and security; guidance on the timing and administration of tests; and changes made to the test since the last administration in spring 2009.

Appendix A of this report contains the Department's presentations on the *Test Coordinator's Manual* and the *Examiner's Manual*. During these presentations, the Assistant Director of Assessment walked the Regional Instructional Facilitators and other Department staff through an annotated version of the *Test Coordinator's Manual* and the *Examiner's Manual*. The Regional Instructional Facilitators, in turn, used this information to train district-level staff.

4.2 Ancillary Materials

Test administration ancillary materials for the MAP contribute to the body of evidence of the validity of score interpretation. This section examines how the test materials address the AERA, APA, & NCME (1999) Standards related to test administration procedures.

For the spring 2010 test administration, CTB/McGraw-Hill produced two types of administration manuals: the *Test Coordinator's Manual* and the *Examiner's Manual*. DESE Curriculum and Assessment staff reviewed, provided feedback, and gave final approval for each manual.

The *Test Coordinator's Manual* is common to all grades and content areas. It provides an overview of MAP and any changes made to MAP for 2010. It gives guidelines for testing, such as the inclusion of special populations, the use of translators, and the invalidation procedures. It also details the Test Coordinator's role in the testing process by outlining nine steps the Test Coordinator should follow. These steps are:

- Step 1: Review Testing Materials
- Step 2: Distribute Testing Materials
- Step 3: Collect Testing Materials
- Step 4: Check the Organization of Materials Collected
- Step 5: Check the Student Information Sheet (SIS)
- Step 6: Check the Group Information Sheet (GIS)
- Step 7: Complete the School/Group List
- Step 8: Organize Materials for the District Test Coordinator
- Step 9: Package and Ship Testing Materials

The *Examiner's Manuals* are specific to each grade. The MAP *Examiner's Manuals* also outline steps that should be followed when administering MAP. These steps include:

- Step 1: Preparing for Testing
- Step 2: Organize Your Classroom
- Step 3: Check your Testing Materials
- Step 4: Before Testing
- Step 5: Administer the Assessments
- Step 6: Invalidations and Make-ups
- Step 7: After Testing: Student Status Coding
- Step 8: Assemble Materials for Return

These steps provide instructions on pre-test and post-test procedures such as:

- Test security
- Standardized testing protocols for norm-referenced information
- Using student barcode labels
- Completing the student information sheet, including recording test accommodations

This section presents the AERA, APA, & NCME (1999) Standards relevant to test administration and how information in the MAP *Examiner's Manuals* and *Test Coordinator Manual* address these Standards.

Standard 3.19 *The directions for test administration should be presented with sufficient clarity and emphasis so that it is possible for others to replicate adequately the administration conditions under which the data on reliability and validity, and, where appropriate, norms were obtained.*

The MAP *Examiner's Manuals* provide instructions for before-, during-, and after-testing activities with sufficient detail and clarity to support reliable test administrations by qualified test administrators. To ensure uniform administration conditions throughout the state, instructions in the *Examiner's Manuals* describe the following: the materials that the examiner and students need for testing; how to verify that pre-coded student information on student barcode labels is correct; how to fill out the Student Information Sheet if the student barcode label is incorrect; how to prepare the testing environment; the test schedule, including testing times; and how to administer the tests.

Standard 3.20 *The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample material, practice or sample questions, criteria for scoring, and a representative item identified with each major area in the test's classification or domain should be provided to the test takers prior to the administration of the test or included in the testing material as part of the standard administration instructions.*

To ensure clarity of instructions to students, the manuals include scripts that the examiner is instructed to read verbatim to students. Examiners are instructed to follow the script and to repeat any part of the directions as many times as needed, but to not modify the words used. Examiners may use professional judgment to respond to student questions, but they may not reword test items, suggest answers, or evaluate student work during the testing session. A sample of a script is presented in Figure 4.1.

Sample test items are provided in each content area to familiarize students with how to fill in answers. Sample items are also provided in the *Examiner's Manuals*.

Standard 5.1 *Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer, unless the situation or a test taker's disability dictates that an exception should be made.*

To ensure the usefulness and interpretability of test scores and to minimize sources of construct-irrelevant variance, it is essential that the MAP is administered according to the prescribed test schedule. The *Test Coordinator's Manual* includes instructions for scheduling the test within the state testing window of March 29 through April 23, 2010, with a one-week extension until April 30, 2010, for 70 districts. The *Examiner's Manuals* contain the schedule for timing each test session and whether timing is to be strictly enforced. The test timing schedule is presented in Table 4.1.

Standard 5.2 *Modifications or disruptions of standardized test administration procedures or scoring should be documented.*

Department staff administer reports on testing concerns which include a wide range of improper activities that may occur during testing, including the following: copying and reviewing grade-level assessment questions with students; cueing students during testing either verbally or with written materials on the classroom walls; cueing students nonverbally, such as tapping or nodding the head; using a calculator on parts of the test where it is not allowed; allowing too much time on *TerraNova* sections of the test; allowing students to correct or complete answers after tests have been returned to the teacher; splitting sessions into two parts; ignoring the standardized directions in the test books; reading the Communication Arts Assessment to students; paraphrasing parts of the test to students; changing or completing (or allowing other school personnel to change or complete) student answers; allowing accommodations that are not written in the Individualized Education Program (IEP); allowing accommodations for students who do not have an IEP; allowing students to use dictionaries on parts of the grade-level assessment other than the writing prompt; or defining terms on the test.

Testing concerns are gathered from school officials, students, parents, and other interested parties who call the Department to state their allegation. A narrative of the conversation is written and read back to them. The superintendent of the district in which the allegation is made is then contacted and read the narrative. A letter is sent to confirm the conversation and to ask the superintendent to investigate the claim. A MAP Quality Assurance Concern District Response Report is sent for the superintendent to use for replying to the allegation. This report is shown in Figure 4.2.

Standard 5.4 *The testing environment should furnish reasonable comfort with minimal distractions.*

Step 2 in the *Examiner's Manual* overviews the steps that teachers should take to prepare their classroom for administering the MAP test. These include:

- Plan for the distribution and collection of materials.
- Plan seating arrangements. Allow enough space between students to prevent the sharing of answers.

- Eliminate distractions such as bells or telephones.
- Use a Do Not Disturb sign on the door of the testing room.
- Make sure classroom maps, charts, and any other materials that relate to the **content and processes** of the test are covered, removed, or out of the students' view.
- When administering the timed portion of the test, write on the board the starting and stopping times for the test.

Standard 5.6 *Reasonable efforts should be made to assure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent means.*

The *Examiner's Manuals* and *Test Coordinator's Manual* present instructions for post-test activities to ensure that test materials are handled properly and to ensure the integrity of student information and test scores. Detailed instructions guide test examiners in completing required information on students' scannable test books. For students who were administered a large print or Braille version of the MAP, examiners are instructed to transcribe students' responses from the large print or Braille test book to a regular-edition test book exactly as they responded in the large print or Braille test book.

Standard 5.7 *Test users have the responsibility of protecting the security of test materials at all times.*

Throughout the manuals, test coordinators and examiners are reminded of test security requirements and procedures to maintain test security. Specific actions that are direct violations of test security are so noted. Detailed information about test security procedures are presented in Section 4.3.

4.2.1 Return Material Forms and Guidelines

The *Test Coordinator's Manual* instructs test coordinators in procedures for organizing and packing materials and returning them to CTB/McGraw-Hill for scanning and scoring. DESE curriculum and assessment staff have opportunities to review, provide feedback, and give final approval. The purpose of the instructions is to ensure that used and unused test materials are properly accounted for and student answer documents are organized properly for return shipment. Proper organization of materials and accurate completion of the school/group list document contributes to accurate score reports and helps in delivery of such reports in a timely manner.

4.2.2 Security Forms

As soon as test books are received by a district, the district test coordinator assures that the first and last security barcode on the tests match the packing list they received. The district test coordinator then packages the tests to be sent to schools. Upon returning tests to CTB/McGraw-Hill, school and district test coordinators are required to complete and submit a *Test Book Accountability Form* that details the number of scorable and nonscorable books returned. This form also requires that districts/schools document nonstandard situations, including lost, damaged, destroyed, extra, or missing test books. The *Test Book Accountability Form* is shown in Figure 4.3.

4.2.3 Interpretive Guides

Essential to making valid interpretations of test scores is an understanding of what the test scores mean and how to interpret score reports. The *Guide to Interpreting Results* is written for Missouri teachers and administrators who receive MAP score reports from the 2010 administration. More detail about the guide can be found in Chapter 7.

4.3 Test Security Measures

Maintaining the security of all test materials is crucial to preventing the possibility of random or systematic errors, such as unauthorized exposure of test items that would affect the valid interpretation of test scores. Several test security measures are implemented for the MAP. Test security procedures are discussed throughout the *Test Examiner Manuals* and *Test Coordinator's Manual*.

Test coordinators and examiners are instructed to keep all test materials in locked storage, except during actual test administration, and access to secure materials must be restricted to authorized individuals only (e.g., test examiners and the school test coordinator). During the testing sessions, test examiners are directly responsible for the security of the MAP and must account for all test materials at all times. The test examiners must supervise the test administrations at all times.

4.4 Test Administration

The 2010 MAP test was administered to students within the state testing window of March 29 through April 23, 2010, with a one-week extension until April 30, 2010, for 70 districts adversely affected by winter weather. Table 4.2 shows those districts that were given a one-week extension of the testing window. Systems chose when and how to administer the MAP within this window. Each session within each content area of the MAP was required to be administered in one block of time.

4.4.1 Time

Each section of each content area test was timed to provide sufficient time for students to attempt all items. The *Examiner's Manuals* provided examiners with timing guidelines for the custom portions of MAP. Strict timing guidelines were given for the norm-referenced portions of the test. For MAP's custom sessions, examiners were instructed to allow students to complete the assessment if s/he was making adequate progress. For the norm-referenced portion of the test, students received an accommodation for additional time if so needed and documented on their IEP. The timing schedule of the MAP is presented in Table 4.1.

4.4.2 Accommodations

Accommodations are allowed on MAP. Test accommodations may be used with students who qualify under IDEA and have an IEP or Section 504 of the Americans' with Disabilities Act and have a 504 plan, or who are identified as English Language Learners. Accommodations must be specified in the qualifying student's IEP and must be consistent with accommodations used during daily classroom instruction and testing. The

use of any accommodation must be indicated on the student information sheet at the time of test administration. AERA, APA, and NCME (1999) Standard 5.3 states:

When formal procedures have been established for requesting and receiving accommodation, test takers should be informed of these procedures in advance of testing.

In compliance with this, each grade-specific MAP *Examiner's Manual* contains the list of accommodations permissible for the MAP assessments. The tables of accommodations presented in each *Examiner's Manual* are shown in Tables 4.3 and 4.4. If a specific accommodation is not on the list of accommodations in the *Examiner's Manual*, the accommodation may still be permitted. However, for accountability purposes, there are some accommodations that will invalidate a student's test results, such as an oral administration of the Communication Arts test or paraphrasing any of the tests. Detailed information regarding testing accommodations can be found at the DESE website:

<http://dese.mo.gov/divimprove/assess/ancillaries.html>

Braille and large print forms are provided to students with vision disabilities.

Tables 4.5 through 4.7 summarize the numbers of reportable students receiving accommodations by accommodation type for the 2010 MAP, the Braille edition of the 2010 MAP, and the large print edition of the 2010 MAP. The analyses in Tables 4.5 through 4.7 are based on census data and include only those students who received accommodations and received a scale score on the Communication Arts, Mathematics, or Science MAP.

In 2010, the setting and timing accommodations appear to be the most frequently used for the Communication Arts, Mathematics, and Science MAP. For the Science and Mathematics MAP, having the test read aloud was among the more frequently used accommodations. For the Mathematics MAP, using calculators was also among the more frequently used accommodations.

On the Braille and large print editions of the MAP, the setting and timing accommodations are again among the most frequently used accommodations. Common accommodations for both the Braille and large print editions included using a scribe for the Communication Arts, Mathematics, and Science MAP tests, having the test read aloud for the Mathematics and Science MAP tests, and using a calculator for the Mathematics MAP tests.

4.5 Summary

In summary, the overall purpose of each of the test administration workshops and the ancillary materials is to keep districts informed about policies and procedures related to testing in general and the MAP in particular. The information imparted is clearly related to standardizing the administration of the MAP, maintaining the security of the

assessment, allowing access to the assessments for special populations by clearly delineating appropriate accommodations, and by providing guidance on appropriate interpretations of the test results. These communication and training efforts by DESE and the ancillary information developed by CTB/McGraw-Hill address multiple best practices of the testing industry but in particular are related to the following Standards (AERA, APA, & NCME, 1999):

- Standard 3.19— The directions for test administration should be presented with sufficient clarity and emphasis so that it is possible for others to replicate adequately the administration conditions under which the data on reliability and validity, and, where appropriate, norms were obtained.
- Standard 3.20— The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample material, practice or sample questions, criteria for scoring, and a representative item identified with each major area in the test’s classification or domain should be provided to the test takers prior to the administration of the test or included in the testing material as part of the standard administration instructions.
- Standard 5.1—Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer, unless the situation or a test taker’s disability dictates that an exception should be made.
- Standard 5.2— Modifications or disruptions of standardized test administration procedures or scoring should be documented.
- Standard 5.3—When formal procedures have been established for requesting and receiving accommodation, test takers should be informed of these procedures in advance of testing.
- Standard 5.4—The testing environment should furnish reasonable comfort with minimal distractions.
- Standard 5.6—Reasonable efforts should be made to assure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent means.
- Standard 5.7—Test users have the responsibility of protecting the security of test materials at all times.

Table 4. 1: MAP Administration Schedule Timing Guidelines by Session (Time in Minutes)

Grade	Session	Communication Arts	Mathematics	Science
3	1	45 - 55*	25 - 35*	
	2	60 - 90	50**	
	3	51**	25 - 35	
	4	15 - 20		
4	1	45 - 55*	40 - 55*	
	2	51**	50**	
	3	15 - 20	25 - 35	
5	1	45 - 55*	25 - 35*	50 - 70*
	2	50**	50**	45**
	3	15 - 20	25 - 35	55 - 70
6	1	45 - 55*	25 - 35*	
	2	51**	50**	
	3	15 - 20	25 - 35	
7	1	45 - 55*	25 - 35*	
	2	60 - 90	50**	
	3	51**	25 - 35	
	4	15 - 20		
8	1	45 - 55*	40 - 55*	50 - 70*
	2	51**	50**	45**
	3	15 - 20	25 - 35	55 - 70

*Session 1 allows an additional 10 minutes, if needed (not included in these test times)

**Strictly timed TerraNova sessions (all other sessions include time ranges as guidelines only)

***Includes 30 minutes of strictly timed TerraNova items plus 40-55 minutes for custom items

Table 4. 2: Districts Granted a One-Week Extension of the MAP Testing Window

District	
Alton R-IV	Kingston K-14
Arcadia Valley R-II	Kirbyville R-VI
Ava R-I	Knox County R-I
Bakersfield R-IV	Laclede Co. R-I
Belton 124	Lakeland R-III
Bismarck R-V	Lonedell R-XIV
Blue Eye R-V	Mansfield R-IV
Bradleyville R-I	Mark Twain R-VIII
Branson R-IV	Maryville R-II
Cassville R-IV	Niangua R-V
Central R-III	Norwood R-I
Chadwick R-I	Oregon-Howell
Clark County R-I	Osceola School
Clinton Co. R-III	Potosi R-III
Couch R-I	Reeds Spring R-IV
Dora R-III	Richards R-V
East Carter R-II	Richmond R-XVI
Excelsior Springs	Richwoods R-VII
Fair View R-XI	Roscoe C-1
Fairfax R-III	Savannah R-III
Fayette R-III	Seneca R-VII
Forsyth R-III	Seymour R-II
Gainesville R-V	Shell Knob 78
Glenwood R-VIII	Sherwood Cass R-VIII
Greenville R-II	Smithville R-II
Hartville	Sparta R-III
Holden R-III	Ste. Genevieve R-II
Hollister R-V	Sullivan
Houston R-I	Taneyville R-II
Howell Valley R-I	Thayer R-II
Imagine Schools	Thornfield R-I
Independence 30	W. St. Francois Co. R-III
Junction Hill C-12	Washington
Kansas City 33	Weaubleau R-III
Kearney R-I	West Plains R-VII

Table 4. 3: MAP Accommodations for Students Who Are English Language Learners

Accommodations List for Students Who Are English Language Learners (ELL)			
The following are the only accommodations allowed for ELL students:			
Code	Invalidates	Administration Accommodations	Description
04	√	Oral reading of assessment (<i>Not permissible for Communication Arts Assessment</i>) (See Note 1.)	The Test Examiner reads items verbatim to the student in an isolated setting so that other students will not benefit or be disturbed.
11	√	Oral reading in native language (<i>Not permissible for Communication Arts Assessment</i>) (See Note 1.)	
		Timing Accommodations	Description
20		Extended time to complete strictly timed sessions (See Note 2.)	ELL students may need to complete the assessments over more than one test period.
21		Administer test using more than allotted periods	Dates for taking the Grade-Level Assessments must occur within the testing window.
22		Other: Specify	Other timing accommodations.
		Response Accommodations	Description
35		Use of scribe to record student response in test booklet	The student conveys verbal responses to a scribe in an isolated, individual setting so that other students cannot benefit or be disturbed. The scribe cannot suggest ideas, words, or concepts. The scribe records the student's answers verbatim. The student should indicate capitalization and punctuation if language mechanics are being assessed.
		Oral response	The student provides an oral response to the Test Examiner.
43	√	Use of bilingual dictionary (<i>Not permissible for Communication Arts Assessment</i>) (See Note 1.)	
		Setting Accommodations	Description
50		Testing individually	The room should be free of noises, conversation, and distractions from adjoining rooms. Individual testing is appropriate when, for example, responses are given orally or questions are paraphrased.
51		Testing with small groups	The location should be free of noises, conversation, and distractions from adjoining rooms. Students may not interact with one another about questions or answers. The Test Examiner must be present at all times. Testing in small groups is not appropriate for students who give responses orally or require paraphrasing of questions.
53		Other: Specify	Other setting accommodations.

NOTES

Note 1: *Oral reading, oral reading in native language, or signing during any Communication Arts Assessment will result in the LOSS (Lowest Obtainable Scale Score). The use of a bilingual dictionary during the Communication Arts Assessment will result in the LOSS (Lowest Obtainable Scale Score).*

Note 2: *If used, the student score cannot be compared with scores generated under standard conditions.*

Table 4. 4: MAP Accommodations for Students with Disabilities

Accommodations List for Students with Disabilities			
Code	Invalidates	Administration Accommodations	Description
01		Braille edition of assessment	Braille editions of the assessment require special processing. Consult the Braille edition test materials for specific instructions.
02		Large Print edition of assessment	Large Print editions of the assessment require special processing. Consult the Large Print test materials for specific instructions.
04	√	Oral reading of assessment (See Note 1.)	The Test Examiner reads items verbatim to the student in an isolated setting so that other students will not benefit or be disturbed.
04		Oral reading of assessment to Blind/Partial Sight students (See Note 1.)	The Test Examiner reads items verbatim to the student who cannot read Braille in an isolated setting so that other students will not benefit or be disturbed.
05		Signing (See Note 1.)	A certified sign language interpreter or deaf education instructor may sign directions for the Communication Arts Assessments. The Mathematics and Science Assessments may have both directions and the test items signed for students.
06	√	Paraphrasing (See Note 2.)	The Test Examiner paraphrases questions to help student understanding in an isolated setting. Terms may be defined as long as they: 1) are not the actual concept or content being assessed, 2) would not give clues, or 3) would not disclose the answer.
10		Other administration accommodations	
		Use of assistive devices	An assistive device that permits a student to read and/or respond to the assessment is used. Examples of assistive devices include computers that assist students with fine-motor problems, text enlargers that enable students to independently read and answer test questions, or augmentative communication devices.
		Use of visual aids: Specify	Visual aids include any type of optical or non-optical devices used to enhance visual capability. Examples of visual aids include bold-line felt-tip markers, lamps, filters, bold-lined paper, writing guides, or other adaptations that alter the visual environment by adjusting the space, illumination, color, contrast, or other physical features of the environment.
		Timing Accommodations	Description
20		Extend time to complete strictly timed sessions (See Note 3.)	Extended time to complete strictly timed sessions is allowed for a student whose disability may cause him/her to be unable to meet time constraints.
21		Administer assessment using more than allotted periods	Students with disabilities may need to complete the assessments over more than one test period as a result of fatigue and/or loss of concentration. Some students may require additional breaks. Dates for taking the Grade-Level Assessment must occur within the testing window.
22		Other: Specify	Other timing accommodations
		Response Accommodations	Description
35		Use of scribe to record student response in test booklet	The student conveys verbally or signs responses to a scribe in an isolated, individual setting so that other students cannot benefit or be disturbed. The scribe cannot suggest ideas, words, or concepts. The scribe records the student's answers verbatim. The student should indicate capitalization and punctuation if language mechanics are being assessed.
		Student taped response	The student speaks responses into a tape recorder in an isolated setting so that other students cannot benefit or be disturbed. The Test Examiner must be present at all times.
		Signed response	The student uses sign language to convey responses. A certified sign language interpreter or deaf education instructor records responses.
		Pointing to respond	The student points to correct responses and the administrator records responses in the Grade-Level Assessment test book.
		Oral response	The student provides oral responses to the Test Examiner.

NOTES

Note 1: Oral reading, oral reading in native language, or signing during the Communication Arts Assessment will result in the LOSS (Lowest Obtainable Scale Score). The use of a bilingual dictionary during the Communication Arts Assessment will result in the LOSS (Lowest Obtainable Scale Score). Students identified as blind/visually impaired (who do not read Braille) may use the oral reading accommodation if it is their primary instructional method.

Note 2: Paraphrasing test questions invalidates all Grade-Level Assessment student scores for accountability purposes.

Note 3: If used, the student score cannot be compared with scores generated under standard conditions.

Note 4: Use of magnifying equipment, amplification equipment, graph paper, and testing with the teacher facing the student are not listed as accommodations because these are no longer required to be reported as accommodations for the Grade-Level Assessments.

Table 4. 4: MAP Accommodations for Students with Disabilities (cont'd)

Accommodations List for Students with Disabilities			
Code	Invalidates	Administration Accommodations	Description
		Use of a Braille	A student records responses using a Braille. Examples of a Braille include a Braillewriter, a slate and stylus, or an electronic Braille note taker.
		Use of a communication device	The student uses a communication device to provide responses to the Test Examiner.
		Use of a computer/word processor/typewriter to respond	The student uses a computer/word processor to write the responses. (Provide a non-networked computer to avoid inappropriate use of the computer to access answers.) The student uses a typewriter to write the responses.
39		Use of a calculator/math table/ abacus	In sessions of the Grade-Level Assessment where calculators are allowed, the accommodation code should not be marked. The use of a calculator represents an accommodation when it is used on a section of the assessment for which calculator use is not allowed. Students may use talking calculators, but only in an isolated setting. Students may use tables to assist in simple addition, subtraction, multiplication, and division facts using whole numbers. Students may use an abacus to perform mathematical computations by sliding beads along rods.
44		Other: Specify (See Note 4.)	Other response accommodations
		Setting Accommodations	Description
50		Testing individually	The location should be free of noises, conversation, and distractions from adjoining rooms. Individual testing is appropriate when, for example, responses are given orally or questions are paraphrased.
51		Testing in small groups	The location should be free of noises, conversation, and distractions from adjoining rooms. Students may not interact with one another about questions or answers. The Test Examiner must be present at all times. Testing in small groups is not appropriate for students who give responses orally or require paraphrasing of questions.
53		Other: Specify	Other setting accommodations

Table 4. 5: Number and Percent of Students Receiving Accommodations by Accommodation Type, MAP 2010 Regular Edition

Grade	Accommodation	Communication Arts		Mathematics		Science	
		Frequency	Percent	Frequency	Percent	Frequency	Percent
3	Regular Edition	66702	100.00%	66765	100.00%		
	Oral reading	37	0.06%	4558	6.83%		
	Oral reading blind	5	0.01%				
	Signing of assessment	6	0.01%	16	0.02%		
	Paraphrasing	1	0.00%	5	0.01%		
	Other administration	95	0.14%	58	0.09%		
	Oral reading in native language	10	0.01%	173	0.26%		
	Extend time— <i>TerraNova</i> session	2958	4.43%	2913	4.36%		
	Administer using > allotted periods	2832	4.25%	2730	4.09%		
	Other timing	488	0.73%	472	0.71%		
	Use of scribe	1928	2.89%	1529	2.29%		
	Use of calculator, math table, etc.	61	0.09%	1598	2.39%		
	Use of bilingual dictionary	2	0.00%	19	0.03%		
	Other response	84	0.13%	75	0.11%		
	Testing individually	2236	3.35%	2020	3.03%		
	Testing in small group	4315	6.47%	4629	6.93%		
	Other setting	286	0.43%	281	0.42%		
4	Regular Edition	67261	100.00%	67351	100.00%		
	Oral reading	44	0.07%	5041	7.48%		
	Oral reading blind	5	0.01%				
	Signing of assessment	3	0.00%	16	0.02%		
	Paraphrasing	1	0.00%	3	0.00%		
	Other administration	119	0.18%	65	0.10%		
	Oral reading in native language	16	0.02%	229	0.34%		
	Extend time— <i>TerraNova</i> session	3177	4.72%	3155	4.68%		
	Administer using > allotted periods	3169	4.71%	3116	4.63%		
	Other timing	629	0.94%	602	0.89%		
	Use of scribe	1994	2.96%	1723	2.56%		
	Use of calculator, math table, etc.	88	0.13%	2202	3.27%		
	Use of bilingual dictionary	0	0.00%	25	0.04%		
	Other response	108	0.16%	85	0.13%		
	Testing individually	2449	3.64%	2237	3.32%		
	Testing in small group	4953	7.36%	5323	7.90%		
	Other setting	310	0.46%	312	0.46%		
5	Regular Edition	66461	100.00%	66541	100.00%	66519	100.00%
	Oral reading	38	0.06%	5036	7.57%	4878	7.33%
	Oral reading blind	8	0.01%				
	Signing of assessment	7	0.01%	24	0.04%	24	0.04%
	Paraphrasing	3	0.00%	3	0.00%	2	0.00%
	Other administration	151	0.23%	56	0.08%	50	0.08%

Table 4. 5: Number and Percent of Students Receiving Accommodations by Accommodation Type, MAP 2010 Regular Edition (Cont'd)

Grade	Accommodation	Communication Arts		Mathematics		Science	
		Frequency	Percent	Frequency	Percent	Frequency	Percent
5	Oral reading in native language	6	0.01%	140	0.21%	109	0.16%
	Extend time— <i>TerraNova</i> session	3259	4.90%	3233	4.86%	2995	4.50%
	Administer using > allotted periods	3323	5.00%	3307	4.97%	3157	4.75%
	Other timing	585	0.88%	565	0.85%	548	0.82%
	Use of scribe	1768	2.66%	1560	2.34%	1686	2.53%
	Use of calculator, math table, etc.	125	0.19%	2611	3.92%	994	1.49%
	Use of bilingual dictionary	5	0.01%	31	0.05%	32	0.05%
	Other response	109	0.16%	90	0.14%	86	0.13%
	Testing individually	2104	3.17%	1945	2.92%	1928	2.90%
	Testing in small group	5126	7.71%	5478	8.23%	5173	7.78%
	Other setting	296	0.45%	294	0.44%	288	0.43%
6	Regular Edition	67215	100.00%	67272	100.00%		
	Oral reading	39	0.06%	4390	6.53%		
	Oral reading blind	8	0.01%				
	Signing of assessment	7	0.01%	20	0.03%		
	Paraphrasing	4	0.01%	6	0.01%		
	Other administration	67	0.10%	38	0.06%		
	Oral reading in native language	6	0.01%	171	0.25%		
	Extend time— <i>TerraNova</i> session	2853	4.24%	2810	4.18%		
	Administer using > allotted periods	2904	4.32%	2881	4.28%		
	Other timing	579	0.86%	554	0.82%		
	Use of scribe	1333	1.98%	1083	1.61%		
	Use of calculator, math table, etc.	185	0.28%	3379	5.02%		
	Use of bilingual dictionary	0	0.00%	61	0.09%		
	Other response	77	0.11%	60	0.09%		
	Testing individually	1684	2.51%	1459	2.17%		
Testing in small group	5618	8.36%	5975	8.88%			
Other setting	159	0.24%	171	0.25%			
7	Regular Edition	65987	100.00%	66005	100.00%		
	Oral reading	18	0.03%	3676	5.57%		
	Oral reading blind	5	0.01%				
	Signing of assessment	3	0.00%	9	0.01%		
	Paraphrasing	0	0.00%	1	0.00%		
	Other administration	41	0.06%	30	0.05%		
	Oral reading in native language	4	0.01%	170	0.26%		
	Extend time— <i>TerraNova</i> session	2226	3.37%	2264	3.43%		
	Administer using > allotted periods	2275	3.45%	2231	3.38%		
	Other timing	524	0.79%	501	0.76%		
	Use of scribe	968	1.47%	685	1.04%		
	Use of calculator, math table, etc.	212	0.32%	3400	5.15%		
	Use of bilingual dictionary	0	0.00%	82	0.12%		

Table 4. 5: Number and Percent of Students Receiving Accommodations by Accommodation Type, MAP 2010 Regular Edition (Cont'd)

Grade	Accommodation	Communication Arts		Mathematics		Science	
		Frequency	Percent	Frequency	Percent	Frequency	Percent
7	Other response	69	0.10%	49	0.07%		
	Testing individually	1150	1.74%	917	1.39%		
	Testing in small group	5542	8.40%	5832	8.84%		
	Other setting	112	0.17%	107	0.16%		
8	Regular Edition	66097	100.00%	66125	100.00%	66061	100.00%
	Oral reading	45	0.07%	3566	5.39%	3504	5.30%
	Oral reading blind	4	0.01%				
	Signing of assessment	10	0.02%	24	0.04%	23	0.03%
	Paraphrasing	5	0.01%	4	0.01%	5	0.01%
	Other administration	29	0.04%	16	0.02%	14	0.02%
	Oral reading in native language	3	0.00%	132	0.20%	121	0.18%
	Extend time— <i>TerraNova</i> session	2222	3.36%	2240	3.39%	2120	3.21%
	Administer using > allotted periods	2304	3.49%	2345	3.55%	2276	3.45%
	Other timing	560	0.85%	547	0.83%	532	0.81%
	Use of scribe	835	1.26%	657	0.99%	764	1.16%
	Use of calculator, math table, etc.	262	0.40%	3682	5.57%	2303	3.49%
	Use of bilingual dictionary	0	0.00%	85	0.13%	90	0.14%
	Other response	66	0.10%	55	0.08%	50	0.08%
	Testing individually	1050	1.59%	901	1.36%	939	1.42%
	Testing in small group	5218	7.89%	5521	8.35%	5212	7.89%
Other setting	118	0.18%	117	0.18%	111	0.17%	

Table 4. 6: Number and Percent of Students Receiving Accommodations by Accommodation Type, MAP 2010 Braille Edition

Grade	Accommodation	Communication Arts		Mathematics		Science	
		Frequency	Percent	Frequency	Percent	Frequency	Percent
3	Braille Edition	6	100.00%	6	100.00%		
	Oral reading	0	0.00%	2	33.33%		
	Oral reading blind	1	16.67%				
	Signing of assessment	0	0.00%	0	0.00%		
	Paraphrasing	0	0.00%	0	0.00%		
	Other administration	1	16.67%	1	16.67%		
	Oral reading in native language	0	0.00%	0	0.00%		
	Extend time— <i>TerraNova</i> session	2	33.33%	2	33.33%		
	Administer using > allotted periods	3	50.00%	3	50.00%		
	Other timing	0	0.00%	0	0.00%		
	Use of scribe	4	66.67%	4	66.67%		
	Use of calculator, math table, etc.	0	0.00%	2	33.33%		
	Use of bilingual dictionary	0	0.00%	0	0.00%		
	Other response	0	0.00%	0	0.00%		
	Testing individually	4	66.67%	4	66.67%		
	Testing in small group	1	16.67%	1	16.67%		
	Other setting	0	0.00%	0	0.00%		
4	Braille Edition	5	100.00%	6	100.00%		
	Oral reading	0	0.00%	2	33.33%		
	Oral reading blind	2	40.00%				
	Signing of assessment	0	0.00%	0	0.00%		
	Paraphrasing	0	0.00%	0	0.00%		
	Other administration	0	0.00%	0	0.00%		
	Oral reading in native language	0	0.00%	0	0.00%		
	Extend time— <i>TerraNova</i> session	1	20.00%	1	16.67%		
	Administer using > allotted periods	2	40.00%	2	33.33%		
	Other timing	0	0.00%	1	16.67%		
	Use of scribe	2	40.00%	2	33.33%		
	Use of calculator, math table, etc.	2	40.00%	3	50.00%		
	Use of bilingual dictionary	0	0.00%	0	0.00%		
	Other response	0	0.00%	0	0.00%		
	Testing individually	3	60.00%	3	50.00%		
	Testing in small group	1	20.00%	2	33.33%		
	Other setting	1	20.00%	1	16.67%		
5	Braille Edition	5	100.00%	5	100.00%	5	100.00%
	Oral reading	0	0.00%	3	60.00%	3	60.00%
	Oral reading blind	1	20.00%				
	Signing of assessment	0	0.00%	0	0.00%	0	0.00%
	Paraphrasing	0	0.00%	0	0.00%	0	0.00%
	Other administration	1	20.00%	1	20.00%	1	20.00%

Table 4. 6: Number and Percent of Students Receiving Accommodations by Accommodation Type, MAP 2010 Braille Edition (Cont'd)

Grade	Accommodation	Communication Arts		Mathematics		Science	
		Frequency	Percent	Frequency	Percent	Frequency	Percent
5	Oral reading in native language	0	0.00%	0	0.00%	0	0.00%
	Extend time— <i>TerraNova</i> session	3	60.00%	3	60.00%	3	60.00%
	Administer using > allotted periods	3	60.00%	3	60.00%	3	60.00%
	Other timing	0	0.00%	0	0.00%	0	0.00%
	Use of scribe	5	100.00%	5	100.00%	5	100.00%
	Use of calculator, math table, etc.	1	20.00%	4	80.00%	2	40.00%
	Use of bilingual dictionary	0	0.00%	0	0.00%	0	0.00%
	Other response	0	0.00%	0	0.00%	0	0.00%
	Testing individually	5	100.00%	5	100.00%	5	100.00%
	Testing in small group	0	0.00%	0	0.00%	0	0.00%
	Other setting	1	20.00%	1	20.00%	1	20.00%
6	Braille Edition	7	100.00%	7	100.00%		
	Oral reading	0	0.00%	2	28.57%		
	Oral reading blind	3	42.86%				
	Signing of assessment	0	0.00%	0	0.00%		
	Paraphrasing	0	0.00%	0	0.00%		
	Other administration	0	0.00%	0	0.00%		
	Oral reading in native language	0	0.00%	0	0.00%		
	Extend time— <i>TerraNova</i> session	2	28.57%	2	28.57%		
	Administer using > allotted periods	2	28.57%	2	28.57%		
	Other timing	0	0.00%	0	0.00%		
	Use of scribe	4	57.14%	4	57.14%		
	Use of calculator, math table, etc.	0	0.00%	1	14.29%		
	Use of bilingual dictionary	0	0.00%	0	0.00%		
	Other response	0	0.00%	0	0.00%		
	Testing individually	6	85.71%	6	85.71%		
Testing in small group	0	0.00%	0	0.00%			
Other setting	1	14.29%	1	14.29%			
7	Braille Edition	5	100.00%	NR			
	Oral reading	0	0.00%				
	Oral reading blind	1	20.00%				
	Signing of assessment	0	0.00%				
	Paraphrasing	0	0.00%				
	Other administration	0	0.00%				
	Oral reading in native language	0	0.00%				
	Extend time— <i>TerraNova</i> session	1	20.00%				
	Administer using > allotted periods	2	40.00%				
	Other timing	0	0.00%				
	Use of scribe	3	60.00%				
	Use of calculator, math table, etc.	1	20.00%				
	Use of bilingual dictionary	0	0.00%				

Table 4. 6: Number and Percent of Students Receiving Accommodations by Accommodation Type, MAP 2010 Braille Edition (Cont'd)

Grade	Accommodation	Communication Arts		Mathematics		Science	
		Frequency	Percent	Frequency	Percent	Frequency	Percent
7	Other response	0	0.00%				
	Testing individually	5	100.00%				
	Testing in small group	0	0.00%				
	Other setting	0	0.00%				
8	Braille Edition	8	100.00%	7	100.00%	7	100.00%
	Oral reading	0	0.00%	1	14.29%	1	14.29%
	Oral reading blind	0	0.00%				
	Signing of assessment	0	0.00%	0	0.00%	0	0.00%
	Paraphrasing	0	0.00%	0	0.00%	0	0.00%
	Other administration	1	12.50%	2	28.57%	2	28.57%
	Oral reading in native language	0	0.00%	0	0.00%	0	0.00%
	Extend time— <i>TerraNova</i> session	4	50.00%	4	57.14%	4	57.14%
	Administer using > allotted periods	4	50.00%	4	57.14%	4	57.14%
	Other timing	0	0.00%	0	0.00%	0	0.00%
	Use of scribe	3	37.50%	3	42.86%	3	42.86%
	Use of calculator, math table, etc.	0	0.00%	6	85.71%	2	28.57%
	Use of bilingual dictionary	0	0.00%	0	0.00%	0	0.00%
	Other response	0	0.00%	1	14.29%	0	0.00%
	Testing individually	3	37.50%	4	57.14%	4	57.14%
	Testing in small group	4	50.00%	3	42.86%	3	42.86%
Other setting	0	0.00%	0	0.00%	0	0.00%	

NR=Not reported due to sample size less than 5 students

Table 4. 7: Number and Percent of Students Receiving Accommodations by Accommodation Type, MAP 2010 Large Print Edition

Grade	Accommodation	Communication Arts		Mathematics		Science	
		Frequency	Percent	Frequency	Percent	Frequency	Percent
3	Large Print Edition	43	100%	43	100%		
	Oral reading	0	0.00%	19	44.19%		
	Oral reading blind	0	0.00%				
	Signing of assessment	0	0.00%	0	0.00%		
	Paraphrasing	0	0.00%	0	0.00%		
	Other administration	1	2.33%	1	2.33%		
	Oral reading in native language	0	0.00%	0	0.00%		
	Extend time— <i>TerraNova</i> session	16	37.21%	17	39.53%		
	Administer using > allotted periods	16	37.21%	16	37.21%		
	Other timing	3	6.98%	3	6.98%		
	Use of scribe	17	39.53%	16	37.21%		
	Use of calculator, math table, etc.	0	0.00%	7	16.28%		
	Use of bilingual dictionary	0	0.00%	0	0.00%		
	Other response	0	0.00%	0	0.00%		
	Testing individually	18	41.86%	20	46.51%		
	Testing in small group	19	44.19%	17	39.53%		
	Other setting	3	6.98%	3	6.98%		
4	Large Print Edition	34	100%	36	100%		
	Oral reading	0	0.00%	18	50.00%		
	Oral reading blind	1	2.94%				
	Signing of assessment	0	0.00%	0	0.00%		
	Paraphrasing	0	0.00%	0	0.00%		
	Other administration	0	0.00%	0	0.00%		
	Oral reading in native language	0	0.00%	1	2.78%		
	Extend time— <i>TerraNova</i> session	15	44.12%	15	41.67%		
	Administer using > allotted periods	14	41.18%	14	38.89%		
	Other timing	1	2.94%	1	2.78%		
	Use of scribe	13	38.24%	13	36.11%		
	Use of calculator, math table, etc.	0	0.00%	13	36.11%		
	Use of bilingual dictionary	0	0.00%	0	0.00%		
	Other response	0	0.00%	0	0.00%		
	Testing individually	18	52.94%	18	50.00%		
	Testing in small group	14	41.18%	15	41.67%		
	Other setting	1	2.94%	1	2.78%		
5	Large Print Edition	34	100%	34	100%	34	100%
	Oral reading	0	0.00%	14	41.18%	14	41.18%
	Oral reading blind	0	0.00%				
	Signing of assessment	0	0.00%	0	0.00%	0	0.00%
	Paraphrasing	0	0.00%	0	0.00%	0	0.00%
	Other administration	2	5.88%	2	5.88%	2	5.88%

Table 4. 7: Number and Percent of Students Receiving Accommodations by Accommodation Type, MAP 2010 Large Print Edition (Cont'd)

Grade	Accommodation	Communication Arts		Mathematics		Science	
		Frequency	Percent	Frequency	Percent	Frequency	Percent
5	Oral reading in native language	0	0.00%	0	0.00%	0	0.00%
	Extend time— <i>TerraNova</i> session	15	44.12%	15	44.12%	13	38.24%
	Administer using > allotted periods	15	44.12%	14	41.18%	12	35.29%
	Other timing	3	8.82%	3	8.82%	3	8.82%
	Use of scribe	13	38.24%	13	38.24%	13	38.24%
	Use of calculator, math table, etc.	1	2.94%	10	29.41%	7	20.59%
	Use of bilingual dictionary	0	0.00%	0	0.00%	0	0.00%
	Other response	1	2.94%	1	2.94%	1	2.94%
	Testing individually	15	44.12%	15	44.12%	15	44.12%
	Testing in small group	15	44.12%	16	47.06%	16	47.06%
	Other setting	0	0.00%	0	0.00%	0	0.00%
6	Large Print Edition	38	100%	36	100%		
	Oral reading	1	2.63%	12	33.33%		
	Oral reading blind	1	2.63%				
	Signing of assessment	0	0.00%	0	0.00%		
	Paraphrasing	0	0.00%	0	0.00%		
	Other administration	2	5.26%	2	5.56%		
	Oral reading in native language	0	0.00%	0	0.00%		
	Extend time— <i>TerraNova</i> session	21	55.26%	18	50.00%		
	Administer using > allotted periods	12	31.58%	10	27.78%		
	Other timing	3	7.89%	4	11.11%		
	Use of scribe	17	44.74%	14	38.89%		
	Use of calculator, math table, etc.	2	5.26%	10	27.78%		
	Use of bilingual dictionary	0	0.00%	0	0.00%		
	Other response	1	2.63%	0	0.00%		
Testing individually	20	52.63%	15	41.67%			
Testing in small group	11	28.95%	13	36.11%			
Other setting	2	5.26%	2	5.56%			
7	Large Print Edition	41	100%	43	100%		
	Oral reading	0	0.00%	20	46.51%		
	Oral reading blind	1	2.44%				
	Signing of assessment	0	0.00%	0	0.00%		
	Paraphrasing	0	0.00%	0	0.00%		
	Other administration	2	4.88%	1	2.33%		
	Oral reading in native language	0	0.00%	1	2.33%		
	Extend time— <i>TerraNova</i> session	10	24.39%	10	23.26%		
	Administer using > allotted periods	11	26.83%	12	27.91%		
	Other timing	5	12.20%	4	9.30%		
	Use of scribe	20	48.78%	17	39.53%		
	Use of calculator, math table, etc.	2	4.88%	20	46.51%		
	Use of bilingual dictionary	0	0.00%	0	0.00%		

Table 4. 7: Number and Percent of Students Receiving Accommodations by Accommodation Type, MAP 2010 Large Print Edition (Cont'd)

Grade	Accommodation	Communication Arts		Mathematics		Science	
		Frequency	Percent	Frequency	Percent	Frequency	Percent
7	Other response	0	0.00%	0	0.00%		
	Testing individually	19	46.34%	18	41.86%		
	Testing in small group	13	31.71%	16	37.21%		
	Other setting	1	2.44%	1	2.33%		
8	Large Print Edition	34	100%	34	100%	33	100%
	Oral reading	0	0.00%	13	38.24%	13	39.39%
	Oral reading blind	2	5.88%				
	Signing of assessment	0	0.00%	0	0.00%	0	0.00%
	Paraphrasing	0	0.00%	0	0.00%	0	0.00%
	Other administration	1	2.94%	1	2.94%	1	3.03%
	Oral reading in native language	0	0.00%	0	0.00%	0	0.00%
	Extend time— <i>TerraNova</i> session	13	38.24%	12	35.29%	13	39.39%
	Administer using > allotted periods	10	29.41%	10	29.41%	10	30.30%
	Other timing	2	5.88%	2	5.88%	1	3.03%
	Use of scribe	11	32.35%	11	32.35%	11	33.33%
	Use of calculator, math table, etc.	3	8.82%	18	52.94%	12	36.36%
	Use of bilingual dictionary	0	0.00%	0	0.00%	0	0.00%
	Other response	0	0.00%	0	0.00%	0	0.00%
	Testing individually	13	38.24%	12	35.29%	12	36.36%
	Testing in small group	12	35.29%	13	38.24%	12	36.36%
Other setting	2	5.88%	2	5.88%	2	6.06%	

Figure 4. 1: Sample Script of Examiner’s Manual
Directions for Administering the Mathematics Assessment

SESSION 1

Punch out all the manipulatives prior to testing.

If this is the first day of testing:

- *Distribute the test books. Ensure that students write their names and district/school on their test book covers. (If this is not the first day of testing, make sure each student has his or her own test book.)*
- *Ensure that all students use nonmechanical No. 2 pencils.*
- *Take a moment to have the students look through the test book.*
- *Hold up a student’s test book and point out the STOP pages. Tell the students that whenever they see one of the STOP pages, they should not continue.*
- *Distribute manipulatives and scratch paper. Scratch paper may include graph or grid paper. Teachers may keep the manipulatives after the test is administered.*
- *Be sure that students understand what each picture means.*



This picture means that you will use your ruler.



This picture means that you will use your pattern blocks.



In Session 1 of the Mathematics Assessment, you may use these manipulatives. You may not need to use both of them.

For the questions in this session, you will select from a list of given answer choices. Use scratch, grid, or graph paper to work the problems. Remember to fill in the circle in the test book that goes with the answer you choose.

You should read each question very carefully and do your best to answer clearly and completely. Your score on these questions will depend on how well you follow directions and show your understanding of mathematics.

CHAPTER 5: CONSTRUCTED-RESPONSE SCORING

In this section, we first describe the scoring process used for MAP. In particular, we focus on the MAP handscoring process. At the end of this section, we describe and report the results of the inter-rater reliability study conducted on the handscoring of MAP CR items.

Chapter 5 adheres to AERA, APA, & NCME (1999) Standards 3.22, 3.23, and 5.9. Each of these Standards will be presented in the pertinent section of this chapter. Standard 3.22 provides some general guidance for Chapter 5:

Procedures for scoring and, if relevant, scoring criteria should be presented by the test developer in sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scores or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical if test can be scored locally.

Chapter 5 explains the procedures used for scoring the MAP CR items. The scoring criteria used for each item is not presented in this chapter to preserve the integrity of the items for future use.

5.1 MAP Scoring Process

Selected-response items were scored by CTB using electronic scanning equipment. Constructed-response items were scored by human raters who were trained by CTB.

5.1.1 Selection of Scoring Raters

AERA, APA, & NCME (1999) Standard 3.23 specifies:

The process for selecting, training, and qualifying scorers should be documented by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the score scale, and the procedures for training scorers should result in a degree of agreement among scorers that allows for the scores to be interpreted as originally intended by the test developer. Scorer reliability and potential drift over time in raters' scoring standards should be evaluated and reported by the person(s) responsible for conducting the training session.

Sections 5.1.1 and 5.1.2 explain how raters (scorers) are selected and trained for the MAP handscoring process. Section 5.1.3 describes how the raters are monitored throughout the MAP handscoring process.

CTB/McGraw-Hill and Kelly Services strive to develop a highly qualified, experienced core of raters so that the integrity of all projects is appropriately maintained.

Recruitment

The MAP 2010 project was staffed with a large number of returning raters and team leaders who had previous experience with MAP and other handscoring projects. Kelly Services also recruited new team leaders and raters for employment. Recruitment sources included advertisements in newspapers in Indianapolis, Indiana; Centennial, Colorado, and nearby areas; and Internet sources.

CTB requires that all raters and team leaders possess a bachelor's degree or higher. Kelly Services carefully screened all new applicants and required them to produce either a transcript or a copy of the degree. Kelly Services also required a one- to two-hour interview/screening process. Individuals who did not present proper documentation or had less than desirable work records were eliminated during this process. Kelly Services verified that 100% of all potential raters met the degree requirement. All experienced raters and team leaders had already successfully completed the screening process.

The Interview Process

All potential raters completed a pre-interview activity. For some parts of the pre-interview activity, applicants were shown examples of test responses and were supplied with a scoring guide. In a brief introduction, they became acquainted with the application of a rubric. After the introduction, applicants applied the scoring guide to score the sample responses. The applicant's scores were used for discussion during the interview process to determine the applicant's trainability as well as his/her ability to understand and implement the standards set forth in the sample scoring guide.

Kelly Services interviewed each applicant and determined the applicant's suitability for a specific content area and grade level. Applicants with strong leadership skills were questioned further to determine whether they were qualified to be team leaders.

When Kelly Services determined applicants were qualified, the applicants were recommended for employment. All assignments were made according to availability and suitability. Before being hired, all applicants were required to read, agree to, and sign a nondisclosure agreement outlining the CTB/McGraw-Hill business ethics and security procedures.

5.1.2 Handscoring Training Process

Training Material Development

All materials necessary for scoring were developed by CTB. These materials include the scoring guides and training papers used to complete the handscoring of CR and extended-response (writing essays and performance events) items.

Missouri operational items have been previously field tested. Prior to actual scoring, handscoring supervisors assembled materials based on the rubrics. Student answer documents were randomly sampled to ensure that a representative sample of possible responses was used. Supervisors selected anchor papers and training papers and recommended clarifications to rubrics. All materials were presented during the Training

Material Review Meeting (TMRM) and scores and annotations were approved by DESE participants.

From that point, training and qualifying materials were developed based on the rubric and scoring philosophies discussed during the TMRM.

Training Material Review Meeting

CTB prepared all anchors, scoring guides, and student response samples for DESE and Missouri participant review. Each response was scored and annotated by DESE participants.

Training and Qualifying Procedures

Handscoring involves training and qualifying team leaders and raters, monitoring scoring accuracy and production, and ensuring security of both the test materials and the scoring facilities. An explanation of the training and qualification procedures follows.

All readers were trained and qualified in a specific Rater Item Block (RIB) consisting of one item to be scored, except in Grades 5 and 8 Science where there was one multi-item RIB. Raters were trained using the following steps:

- Reviewing CR items
- Reviewing rubrics
- Reviewing anchor papers
- Explaining scoring strategies, followed by a question-and-answer period
- Scoring a training set, followed by sharing established scores
- Scoring additional training sets
- Qualifying Round 1
- Qualifying Round 2 (if necessary)
- Explaining condition codes and sensitive paper procedures
- Explaining nonstandard response (nsr) or computer-generated response (cgr) procedures
- Explaining unscannable image procedures

All raters were trained and qualified using the same procedures and criteria. Qualification standards for every item were predetermined by DESE. In order to score an item, readers must have met the specific standards for that item. The qualification standards were:

- 4-point item: 80% qualification
- 3-point item: 80% qualification
- 2-point item: 90% qualification
- 1-point item: 100% qualification

5.1.3 Monitoring the Scoring Process

AERA, APA, & NCME (1999) Standard 5.9 says:

When test scoring involves human judgment, scoring rubrics should specify criteria for scoring. Adherence to established scoring criteria should be monitored and checked regularly. Monitoring procedures should be documented.

Section 5.1.3 explains the monitoring procedures that CTB uses to ensure that handscoring raters follow established scoring criteria while items are being scored. Detailed scoring rubrics are available for all CR items, which specify the criteria for scoring those CR items. These rubrics will not be presented here in order to preserve the integrity of the items for use in future MAP forms.

Daily Accuracy Checks

Throughout the course of handscoring, calibration sets of pre-scored papers (checksets/validity sets) were administered daily to each rater to monitor scoring accuracy and to maintain a consistent focus on the established rubrics and guidelines. Checksets were executed via imaging software that provided images in such a way that the reader did not know when a checkset was administered.

In addition to the checkset process, CTB's handscoring protocol included the use of read-behinds. The read-behind was another valuable rater-reliability monitoring technique that allowed a team leader to review a reader's scored documents and provide feedback and counseling as appropriate.

Approximately 5% of Communication Arts, Mathematics, and Science papers were scored by a second reader to establish inter-rater reliability statistics for all CR items. This procedure is called a "double-blind read" because the second reader does not know the first reader's score.

Recalibration of Raters

Recalibration in handscoring refers to the process in which raters who begin to drift away from scoring accuracy are realigned to correct scoring. After a thorough review of the rubric, anchors, and training papers, a recalibration round is administered to a reader who has drifted; accuracy on this round must meet or exceed the qualification rate. A rater who continues to exhibit drift is released.

5.1.4 Security

Security guards were on site whenever employees were present in the building. All employees were issued photo identification badges and required to wear them in plain view at all times. Visitors and employees who forgot their badges were issued visitors' badges and were required to wear them in plain view. All employees and visitors were subject to inspection of their personal effects.

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 through 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 84% 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 7 item (Session 1, Item 3) had weighted Kappa values that indicate only moderate agreement between the raters.

As shown in Table 5.2, raters demonstrated at or above 98% 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 96% perfect and adjacent agreement for all Science items. The weighted Kappa statistic indicates good or very good inter-rater agreement for all Science items.

5.3 Summary

The information presented in this chapter summarizes the steps taken by CTB to ensure accuracy in the handscoring process. The inter-rater reliability statistics presented in Section 5.2 demonstrate that the items are scored reliably. These efforts by CTB address multiple best practices of the testing industry, but are particularly related to AERA, APA, & NCME (1999) Standards 3.22, 3.23, and 5.9.:

- Standard 3.22—Procedures for scoring and, if relevant, scoring criteria should be presented by the test developer in sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scores or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical if test can be scored locally.
- Standard 3.23—The process for selecting, training, and qualifying scorers should be documented by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the score scale, and the procedures for training scorers should result in a degree of agreement among scorers that allows for the scores to be interpreted as originally intended by the test developer. Scorer reliability and potential drift over time in raters' scoring standards should be evaluated and reported by the person(s) responsible for conducting the training session.
- Standard 5.9—When test scoring involves human judgment, scoring rubrics should specify criteria for scoring. Adherence to established scoring criteria should be monitored and checked regularly. Monitoring procedures should be documented.

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

Grade	Session	Item #	# Points	% Perfect	% Adjacent	% Perfect & Adjacent*	Weighted Kappa
3	1	3	2	86	14	100	0.83
	1	4	2	72	26	98	0.71
	1	5	2	78	20	98	0.78
	1	6A	2	86	13	99	0.87
	1	6B	1	96	3	99	0.88
	1	6C	1	98	2	100	0.80
	2	1	4	73	26	99	0.79
4	1	3	2	88	12	100	0.90
	1	4	2	89	11	100	0.90
	1	5	2	94	5	99	0.93
	1	6A	2	83	17	100	0.82
	1	6B	1	99	1	100	0.94
	1	6C	1	100	1	101	0.91
5	1	3	2	69	29	98	0.66
	1	4A	2	77	21	98	0.77
	1	4B	2	97	3	100	0.84
	1	5	2	84	12	96	0.82
	1	6	2	81	18	99	0.74
6	1	3	2	91	9	100	0.74
	1	4	2	78	20	98	0.71
	1	5	2	56	34	90	0.43
	1	6A	2	80	19	99	0.84
	1	6B	1	96	4	100	0.86
7	1	3	2	74	10	84	0.58
	1	4	2	85	15	100	0.71
	1	5	2	76	23	99	0.78
	1	6A	2	88	11	99	0.83
	1	6B	1	97	3	100	0.92
	1	6C	1	98	2	100	0.64
	2	1	4	70	29	99	0.78
8	1	3	2	83	16	99	0.85
	1	4	2	71	24	95	0.71
	1	5	2	72	25	97	0.70
	1	6A	2	69	30	99	0.67
	1	6B	1	97	3	100	0.87
	1	6C	1	97	3	100	0.82

* 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

Grade	Session	Item #	# Points	% Perfect	% Adjacent	% Perfect & Adjacent*	Weighted Kappa
3	3	1	2	95	5	100	0.91
	3	2	2	93	5	98	0.94
	3	3	2	95	6	101	0.96
	3	4	2	97	3	100	0.96
4	1	22	4	80	18	98	0.92
	3	1	2	96	4	100	0.95
	3	2	2	93	7	100	0.95
	3	3	2	94	6	100	0.92
	3	4	2	92	8	100	0.93
5	3	1	2	96	3	99	0.97
	3	2	2	92	8	100	0.94
	3	3	2	97	3	100	0.98
	3	4	2	95	5	100	0.95
6	3	1	2	90	10	100	0.94
	3	2	2	95	5	100	0.96
	3	3	2	90	10	100	0.88
	3	4	2	95	5	100	0.95
7	3	1	2	96	4	100	0.96
	3	2	2	98	2	100	0.98
	3	3	2	97	2	99	0.97
	3	4	2	95	5	100	0.94
8	1	20	4	83	15	98	0.94
	3	1	2	91	9	100	0.94
	3	2	2	97	3	100	0.98
	3	3	2	98	2	100	0.98
	3	4	2	81	19	100	0.82

* 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

Grade	Session	Item #	# Points	% Perfect	% Adjacent	% Perfect & Adjacent*	Weighted Kappa
5	1	1	2	98	2	100	0.98
	1	2	2	83	18	101	0.85
	1	3	2	91	8	99	0.92
	1	4	2	84	14	98	0.86
	1	5	2	86	12	98	0.87
	1	6	2	88	12	100	0.90
	1	7	2	83	16	99	0.84
	1	8	2	80	18	98	0.82
	1	9	2	93	7	100	0.94
	1	10	2	86	14	100	0.87
	1	11	2	95	4	99	0.96
	1	12	2	88	11	99	0.85
	1	13	2	95	4	99	0.96
	3	1	2	98	2	100	0.98
	3	2	1	85	14	99	0.71
	3	3	4	87	9	96	0.90
	3	4	2	100	1	101	0.96
	3	5	1	87	13	100	0.74
	3	6	1	94	6	100	0.80
	3	7	2	87	12	99	0.68
3	8	1	88	12	100	0.73	
8	1	1	2	95	5	100	0.96
	1	2	2	84	15	99	0.85
	1	3	2	93	7	100	0.94
	1	4	2	81	17	98	0.85
	1	5	2	96	4	100	0.96
	1	6	2	91	9	100	0.91
	1	7	2	78	21	99	0.72
	1	8	2	76	22	98	0.68
	1	9	2	90	10	100	0.85
	1	10	2	90	9	99	0.89
	1	11	2	83	15	98	0.80
	1	12	2	93	7	100	0.90
	1	13	2	93	7	100	0.85
	1	14	2	93	7	100	0.77
	3	1	1	90	10	100	0.78
	3	2	1	88	13	101	0.73
	3	3	2	79	20	99	0.72
	3	4	1	96	1	97	0.98
	3	5	1	89	11	100	0.75
	3	6	4	73	24	97	0.87

Table 5. 3: Inter-Rater Reliability, Science (Cont'd)

Grade	Session	Item #	# Points	% Perfect	% Adjacent	% Perfect & Adjacent*	Weighted Kappa
8	7	3	1	87	12	99	0.71
	8	3	2	91	8	99	0.93
	9	3	2	88	11	99	0.88

* 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

This chapter of the MAP Technical Report describes the analyses that occurred on the operational data. These analyses include a classical item analysis and an examination of the raw scores and an IRT analysis involving calibrating, scaling, and linking. All of these analyses were conducted using the calibration sample.

In the following section, we first discuss the calibration sample. Next, we present the classical item statistics, including aggregate raw score statistics and individual item-level statistics. Then, we discuss the IRT models used for calibrating the data and address how well these models fit the Missouri data. If the IRT models fit the empirical item response distributions for the population for which we want to make generalizations (i.e., Missouri students), then the claim is strengthened that the scores are valid indicators of an underlying ability. The lowest obtainable scale score (LOSS) and highest obtainable scale score (HOSS) for MAP are presented. Finally, we provide a general overview of scaling and discuss the methods used to link the MAP results to the *TerraNova* scale.

Chapter 6 demonstrates adherence in the MAP to AERA, APA, & NCME (1999) Standards 1.5, 2.8, 3.18, 4.2, 4.11, 4.13, and 6.4. Each Standard will be explicated within the appropriate section of this chapter. Standard 6.4 provides general guidance that is relevant to this chapter. It states:

The population for whom the test is intended and the test specifications should be documented. If applicable, the item pool and scale development procedures should be described in the relevant test manuals. If normative data are provided, the norming population should be described in terms of relevant demographic variables, and the year(s) in which the data were collected should be reported.

In section 6.1, we will discuss the calibration sample and compare it to the general population. The test specifications and item pool are discussed in Chapter 3. The scale development procedures are discussed in section 6.4 of this chapter. Information regarding reported data are discussed in detail in Chapter 7. Information on the normative data may be found in the *TerraNova, Third Edition: Technical Addendum Forms E and F* (2009).

6.1 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 2010, the grade-level calibration samples were comprised of at least 80% of the total student population for that grade. Several large school districts were identified for

inclusion in the 80% sample. These districts are listed in Table 6.1. Data from these districts had to be included in the calibration sample before data analysis procedures could begin. These large districts were identified because past data processing has demonstrated that large districts often return data at the end of the data-return window while small districts often return data early in the data-return window. Since the calibration sample was going to be based on the first 80% of data to be returned, it was important to identify large districts to ensure the calibration data were representative of the state.

Tables 6.2 through 6.4 examine the representativeness of the calibration sample compared to the census data. These tables demonstrate that the calibration sample was representative of the state.

6.2 Classical Item Statistics

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

6.2.1 Test-Level Statistics

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

6.2.2 Item-Level Statistics

Tables 6.8 through 6.13 present the item statistics for each item by grade level/content area. The tables include test session, item number and part (if applicable), p -value, item-total correlation (R_{it}), and omit rate for each item by grade level(s)/content area(s).

p-value: The p -value is a measure of item difficulty. For a SR 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 CR 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 0.70 (Grade 8) to 0.78 (Grade 4). Table 6.6 shows that the average p -values associated with the Mathematics MAP range from 0.59 (Grade 8) to 0.80 (Grade

3). Table 6.7 demonstrates that the average p -values associated with the Science MAP range from 0.60 (Grade 8) to 0.64 (Grade 5).

It is important that one examines the range of p -values, 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 Communication Arts MAP has items with p -values ranging from the low 0.20s to the 0.90s (see Tables 6.8 through 6.13). The p -values on the Mathematics MAP tend to range from the 0.10s and 0.20s to the 0.90s (see Tables 6.8 through 6.13). The Science MAP has items with p -values ranging from 0.10s to the 0.90s (see Tables 6.10 and 6.13). Such a broad range of p -values indicates that the items measure well throughout the range of skills at a given grade, and hence supports the accuracy of the MAP test scores.

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 0.20 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 0.20 threshold was further analyzed to assure that the item was correctly keyed.

Omit Rates: The omit rate for each item indicate 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.

This examination of omit rates complies with Standards 2.8 and 3.18 of the AERA, APA, & NCME (1999) Standards. Both Standards are concerned with speededness of a test. Standard 2.8 states:

Test users should be informed about the degree to which rate of work may affect examinee performance.

The results in this section will show that, overall, student test scores are not adversely affected by the rate at which they complete the test. In general, students have ample time to complete all sections of the test. Related to this, Standard 3.18 states:

For tests that have time limits, test development research should examine the degree to which scores include a speed component and evaluate the appropriateness of that component, given the domain the test is designed to measure.

Again, the results presented in Tables 6.8 through 6.13 show that the majority of tests did not have a speed component. These results are particularly relevant to the *TerraNova* component of the test, which is a strictly timed administration. The results of our analyses suggest that the majority of students were able to complete the test in the prescribed amount of time.

6.3 Item Response Theory

A marginal maximum-likelihood procedure was used to simultaneously estimate the item parameters using the three-parameter/two-parameter partial credit (3PL/2PPC) IRT models (Bock & Aitkin, 1981; Thissen, 1982). Under the 3PL model, the probability that a student with trait or scale score θ will respond correctly to SR item j is

$$P_j(\theta) = c_j + (1 - c_j) / [1 + \exp(-1.7a_j(\theta - b_j))].$$

In the equation, a_j is the item discrimination, b_j is the item difficulty, and c_j is the probability of a correct response by a very low-ability student. Under the 2PPC model, the probability that a student with trait or scale score θ will respond in category k to partial-credit item j is

$$P_{jk}(\theta) = \exp(z_{jk}) / \sum_{i=1}^{m_j} \exp(z_{ji}),$$

$$\text{where } z_{jk} = (k - 1)f_j - \sum_{i=0}^{k-1} g_{ji} \text{ and } g_{j0} = 0 \text{ for all } j.$$

The summary output of the 3PL and 2PPC models is in two different metrics. The location and discrimination parameters for the SR items are in the traditional 3PL metric and are labeled b and a , respectively. In the 2PPC model, f (alpha) and g (gamma) are analogous to b and a , where alpha is the discrimination parameter and gamma over alpha (g/f) is the location where adjacent trace lines cross on the ability scale. Because of the different metrics used, the 3PL parameters b and a are not directly comparable to the 2PPC parameters f and g ; however, they can be converted to a common metric. The two metrics are related by $b = g/f$ and $a = f / 1.7$ (Burket, 1995). As a result of this procedure, the SR and CR items are placed on the same scale. Note that for the 2PPC model, there are $m_j - 1$ (where m_j is a score level j) independent g 's and one f , for a total of m_j independent parameters estimated for each item, while there is one a and one b per item in the 3PL model.

6.3.1 Model Fit

A procedure developed by Yen (1981) was used to assess model-to-data fit for all test items. In this procedure, students are rank ordered on the basis of their $\hat{\theta}$ values and sorted into ten cells with ten percent of the sample in each cell. Each item j in each decile

i has a response from N_{ij} examinees. The fitted IRT models are used to calculate an expected proportion E_{ijk} of examinees who respond to item j in category k . The observed proportion O_{ijk} is also tabulated for each decile, and the approximate chi-square statistic

$$Q_{1j} = \sum_{i=1}^{10} \sum_{k=1}^{m_j} \frac{N_{ij} (O_{ijk} - E_{ijk})^2}{E_{ijk}}.$$

Q_{1j} should be approximately chi-square distributed with degrees of freedom (DF) equal to the number of “independent” cells, $10(m_j-1)$, minus the number of estimated parameters. For the 3PL model, $m_j=2$, so $DF = 10(2-1) - 3 = 7$. For the 2PPC model, $DF = 10(m_j - 1) - m_j = 9m_j - 10$. Since DF differs between SR and CR items and between CR items with different score levels m_j , Q_{1j} is transformed, yielding the test statistic

$$Z_j = \frac{Q_{1j} - DF}{\sqrt{2DF}}.$$

This statistic is useful for flagging items that fit relatively poorly. Z_j is sensitive to sample size, and cutoff values for flagging an item based on Z_j have been developed and were used to identify items for the item review. The cutoff value is $(N/1500 \times 4)$ for a given test, where N is the sample size.

Nine MAP operational items were flagged for poor fit. In Communication Arts, one item was flagged for poor fit in each of Grades 3, 4, 7, and 8. In Mathematics, one item was flagged for poor fit in Grade 3, three items were flagged for poor fit in Grade 6, and one item in Grade 7. Table 6.14 shows the chi-square statistic and the Z -statistic for each 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.

Each flagged item was examined more closely by studying its item characteristic curve (ICC) at each non-zero score point. 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.

Figures 6.1 through 6.9 show the ICCs for each of the misfitting MAP items. The smooth line in each of these figures represents the predicted relationship between examinee performance on the item and examinee ability, and the jagged line represents the

observed relationship.³ Large differences between the two lines indicate poor fit. Each 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.

With large numbers of observations such as there are for the Missouri calibration samples, items may be flagged for statistically significant differences; however, these differences may not be of practical importance. In the case of the nine MAP items flagged for misfit, the differences do not seem to be of practical importance. Misfitting items that have content validity are often retained for use in one assessment and monitored over a period of usage. A large number of misfitting items in an assessment would indicate that caution should be exercised in the interpretation of the overall score. No MAP test had more than three items flagged for misfit.

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.

Figure 6.2 presents the ICC for Session 2, Item 30 (SR item) on the Grade 4 Communication Arts test. There is poor fit throughout the ability range.

Figure 6.3 presents the ICC for Session 3, Item 28 (SR item) on the Grade 7 Communication Arts test. There is poor fit throughout the ability range.

Figure 6.4 presents the ICC for Session 2, Item 12 (SR item) on the Grade 8 Communication Arts test. There is poor fit throughout the ability range.

Figure 6.5 presents the ICC for Session 3, Item 2 (2-point CR item) on the Grade 3 Mathematics test. There is poor fit throughout the ability range of level 2. There is poor fit at the upper end of the ability range of level 3.

Figure 6.6 presents the ICC for Session 1, Item 5 (SR item) on the Grade 6 Mathematics test. There is poor fit throughout the ability range.

Figure 6.7 presents the ICC for Session 1, Item 17 (SR item) on the Grade 6 Mathematics test. There is poor fit throughout the ability range.

Figure 6.8 presents the ICC for Session 3, Item 4 (2-point CR item) on the Grade 6 Mathematics test. As shown, there is poor fit at the low end of the ability range for level 1, throughout the distribution for level 2, and at the upper end of the range for level 3.

Figure 6.9 presents the ICC for Session 1, Item 12 (SR item) on the Grade 7 Mathematics test. There is poor fit throughout the ability range.

³ For CR items, there will be one graph for each score level. For example, a 2-point item will have three graphs for 0, 1, and 2 score points.

6.4 Scaling

The purpose of scaling a test is to enhance its validity by increasing the comparability of test takers' scores. In this section, we explicate the way in which the MAP scales are produced to comply with Standard 4.2 of the AERA, APA, & NCME (1999) Standards, which states:

The construction of scales used for reporting scores should be described clearly in the test documents.

The MAP scores are produced using the three-parameter logistic, two-parameter partial credit (3PL/2PPC) IRT model (explained previously) that assumes that each of the items and tasks is an independent indicator of the underlying ability governing the propensity for students to answer an item correctly (or with greater correctness in the case of the multilevel CR items).

Scaling and linking of complex assessment data were performed using PARDUX (Burket, 1995), which is proprietary software developed by CTB/McGraw-Hill. PARDUX is designed to produce a single scale by jointly analyzing data resulting from students' responses to both SR items and CR items. In PARDUX, items are calibrated based on IRT using the 3PL model (Lord & Novick, 1968) for SR items and the 2PPC model (Yen, 1993) for CR items. PARDUX is also used to link the scales developed by two calibrations through the common-item procedure developed by Stocking & Lord (1983).

6.4.1 Linking Methods

CTB uses a common-item, non-equivalent groups design to link the current year's assessment to the established MAP scale. The embedded *TerraNova* form serves as the anchor set, and the non-equivalent groups are comprised of at least 80% of the census data in each grade. After the initial IRT item calibration, item parameters were linked to the MAP scale using the Stocking & Lord (1983) equating procedure.

Standard 4.11 of the AERA, APA, & NCME (1999) Standards states:

When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions or other linkages were established and on the accuracy of equating functions.

The Stocking & Lord (1983) procedure minimizes the mean squared difference between the two TCCs, one based on estimates from the previous calibration and the other on transformed estimates from the current calibration. Let $\hat{\psi}_j$ be the TCC based on estimates from a previous calibration and $\hat{\psi}_j^*$ be the TCC based on transformed estimates from the current calibration.

$$\hat{\psi}_j = \hat{\psi}(\theta_j) = \sum_{i=1}^n P_i(\theta_j; a_i, b_i, c_i),$$

and

$$\hat{\psi}_j^* = \hat{\psi}(\theta_j) = \sum_{i=1}^n P_i(\theta_j; \frac{a_i}{M_1}, M_1 b_i + M_2, c_i).$$

The TCC method determines the scaling constants (M_1 and M_2) by minimizing the following quadratic loss function (F):

$$F = \frac{1}{N} \sum_{a=1}^N (\hat{\psi}_j - \hat{\psi}_j^*)^2.$$

The standard error of the equating (SEE) is difficult and cumbersome to estimate for IRT equating procedures, like Stocking and Lord (Kolen & Brennan, 1995; Michaelides & Haertel, 2004). The estimation of the SEE is beyond the scope of this report. It is anticipated that the SEE would be small because 80% of the census data is used for the purposes of linking each year. The large sample size (55,000+) should ensure that the equating estimates are fairly stable.

6.4.2 Anchor Items

AERA, APA, & NCME (1999) Standard 4.13 requires information about the anchors, stating:

In equating studies that employ an anchor test design, the characteristics of the anchor test and its similarity to the forms being equated should be presented, including both content specifications and empirically determined relationships among test scores. If anchor items are used, as in some IRT-based and classical equating studies, the representativeness and psychometric characteristics of anchor items should be presented.

The content representation of the anchor items is shown in Tables 3.4, 3.5, and 3.6 of Chapter 3. Appendix B provides further details on psychometric characteristics of the anchor items.

6.4.3 Vertical Scale

The scale on which the MAP scale scores are reported is based in part on the *Terra Nova* standardized achievement test, which makes it possible to report national percentile scores in addition to the criterion-referenced scale scores of MAP. Although the MAP

scale is unique to Missouri, the characteristic growth seen on the scale from grade to grade for the standardized test has been utilized and built upon to give MAP its vertical scale characteristics. The vertical scale is sometimes referred to as a growth scale.

Evidence of the validity of the MAP growth scale is provided by the increase of the scale score at selected percentiles as grade level increases. Figures 6.10, 6.11, and 6.12 display the scale scores for several points on the score distributions for each grade of the Communication Arts, Mathematics, and Science MAP tests, respectively. These scale scores indicate the growth, or change, in score by grade at the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th percentiles. Ideally, the scale score associated with each percentile will increase from grade to grade. Figure 6.10 shows the selected percentiles for the Communication Arts MAP. Considering all but the 1st and 99th percentiles, the scale scores progress upward from Grades 3 through 5 and then flatten from Grades 5 to 6 before continuing to progress upward again from Grade 7 to 8. At the 1st, 5th, and 10th percentiles, there is a decrease in scale score from Grade 6 to 7. At the 99th percentile, there is a decrease in scale score from Grade 4 to 6.

Figure 6.11 shows the selected percentiles for the Mathematics MAP. Except for the 1st and 99th percentiles, there is an upward progression of scale scores across all grades. At the 1st percentile, there is a decrease in scale score between Grades 6 and 7. At the 99th percentile, there is a decrease in scale scores between Grades 3 and 4.

Figure 6.12 shows the selected percentiles for the Science MAP. There is an upward progression of scale scores across the two Science grades.

Figures 6.13 through 6.15 show the TCCs by grade for the MAP Communication Arts, Mathematics, and Science, respectively. Because these tests were linked to the *TerraNova* scale, they have an underlying vertical scale. By plotting the TCCs together, we can demonstrate that the tests increase in difficulty as the grade levels increase. Figure 6.13 shows that the TCCs for Communication Arts for Grades 5, 6, and 7 overlap. Grades 5 and 6 TCCs are very close to each other, separating only in the middle of the TCCs. Grade 7 TCC crosses Grades 5 and 6 TCCs at the lower end. During the selection of the forms, the pre-equated TCCs were examined and efforts were made to further separate Grades 5 through 7 TCCs while, at the same time, protecting against scale drift. The available item pool was insufficient to create tests that resulted in the optimal increases in test difficulty. For Grade 7, the mean scale score is higher than Grades 5 and 6. The Grades 5 and 6 mean scale scores were nearly identical. DESE continues to work on differentiating skills in these grades, which may help pull apart the Grades 5 and 6 TCCs.

For both Mathematics (Figure 6.14) and Science (Figure 6.15), the TCCs indicate that test difficulty increases with grade level.

6.4.4 Lowest and Highest Obtainable Scale Scores

A maximum likelihood procedure cannot produce scale score estimates for students with perfect scores or scores below the level expected by guessing. Also, although maximum

likelihood estimates are available for students with extreme scores other than zero or perfect, occasionally these estimates have standard errors of measurement that are very large, and differences between these extreme values have little meaning. Therefore, scores are established for these students based on a rational but necessarily non-maximum likelihood procedure. These values, which are set separately by grade, are called the LOSS and the HOSS. Table 6.15 shows the LOSS and HOSS values used for each grade of the Communication Arts, Mathematics, and Science MAP tests.

6.5 Item-Pattern Scoring

MAP scale scores are derived using item-pattern scoring; thus, these scale scores are based on the student's responses to all items on a given test, and scale scores account for the characteristics of the items that are in the test (such as item difficulty). A scale score can be interpreted as a highly probable estimate of a student's ability in a given content area.

Using item-pattern scoring, a student's scale score is based on the student's responses to each item (his/her item-response vector). Each item uses optimal item weights in terms of item information, meaning that items do not contribute equally to the overall scale score. Students with the same raw score may be assigned to different scale scores depending on which items they answered correctly.

The procedures applied here are similar to those followed in the development of the *TerraNova* and *TerraNova*, The Second Edition tests. For additional information on the technical details of the item-pattern scoring, readers can also refer to Yen & Candell (1991) and to the technical report for *TerraNova* The Second Edition (CTB/McGraw-Hill, 2003).

6.6 Summary

In summary, the overall purpose of the operational data analysis is to ensure that the test items, as well as the overall test, are functioning appropriately. It also helps maintain the test scale across the years so that test results may be appropriately compared across years. The data analyses undertaken by CTB/McGraw-Hill address multiple best practices of the testing industry but in particular are related to the following Standards (AERA, APA, & NCME, 1999):

- Standard 1.5—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.
- Standard 2.8—Test users should be informed about the degree to which rate of work may affect examinee performance.
- Standard 3.18—For tests that have time limits, test development research should examine the degree to which scores include a speed component and evaluate the appropriateness of that component, given the domain the test is designed to measure.

- Standard 4.2—The construction of scales used for reporting scores should be described clearly in the test documents.
- Standard 4.11—When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions or other linkages were established and on the accuracy of equating functions.
- Standard 4.13—In equating studies that employ an anchor test design, the characteristics of the anchor test and its similarity to the forms being equated should be presented, including both content specifications and empirically determined relationships among test scores. If anchor items are used, as in some IRT-based and classical equating studies, the representativeness and psychometric characteristics of anchor items should be presented.
- Standard 6.4—The population for who the test is intended and the test specifications should be documented. If applicable, the item pool and scale development procedures should be described in the relevant test manuals. If normative data are provided, the norming population should be described in terms of relevant demographic variables, and the year(s) in which the data were collected should be reported.

Table 6. 1: Large Districts that Were Included in the 80% Calibration Sample

District Name
Columbia
St. Joseph
North Kansas
Springfield
Blue Springs
Lee's Summit
Kansas City
Fort Zumwalt
Francis Howell
Hazelwood
Ferguson Florissant
Rockwood
Mehlville
Parkway
St. Louis City

Table 6. 2: Summary of Calibration and Census Data: Communication Arts

	Communication Arts, Grade 3				
	Calibration Sample		Census Data		Diff (Calib % - Census %)
	N	%	N	%	
All Students	66678		66751		
Gender					
Male	33761	50.63	33809	50.65	-0.02
Female	32841	49.25	32866	49.24	0.01
Unknown	76	0.11	76	0.11	0.00
Race/Ethnicity					
White	49983	74.96	50029	74.95	0.01
Black	11819	17.73	11841	17.74	-0.01
Hispanic	3093	4.64	3096	4.64	0.00
Asian/Pacific Islander	1411	2.12	1412	2.12	0.00
Native American/ Alaskan	308	0.46	309	0.46	0.00
Unknown	64	0.10	64	0.10	0.00
	Communication Arts, Grade 4				
All Students	67225		67301		
Gender					
Male	34492	51.31	34542	51.32	-0.01
Female	32648	48.57	32674	48.55	0.02
Unknown	85	0.13	85	0.13	0.00
Race/Ethnicity					
White	50434	75.02	50485	75.01	0.01
Black	12103	18.00	12119	18.01	-0.01
Hispanic	2963	4.41	2972	4.42	-0.01
Asian/Pacific Islander	1361	2.02	1361	2.02	0.00
Native American/ Alaskan	281	0.42	281	0.42	0.00
Unknown	83	0.12	83	0.12	0.00
	Communication Arts, Grade 5				
All Students	66440		66500		
Gender					
Male	34025	51.21	34059	51.22	-0.01
Female	32334	48.67	32358	48.66	0.01
Unknown	81	0.12	83	0.12	0.00
Race/Ethnicity					
White	50107	75.42	50144	75.40	0.02
Black	11884	17.89	11899	17.89	0.00
Hispanic	2772	4.17	2773	4.17	0.00
Asian/Pacific Islander	1279	1.93	1283	1.93	0.00
Native American/ Alaskan	313	0.47	314	0.47	0.00
Unknown	85	0.13	87	0.13	0.00

Table 6. 2: Summary of Calibration and Census Data: Communication Arts (Cont'd)

	Communication Arts, Grade 6				
	Calibration Sample		Census Data		Diff (Calib % - Census %)
	N	%	N	%	
All Students	67200		67260		
Gender					
Male	34455	51.27	34490	51.28	-0.01
Female	32674	48.62	32698	48.61	0.01
Unknown	71	0.11	72	0.11	0.00
Race/Ethnicity					
White	50725	75.48	50768	75.48	0.00
Black	12056	17.94	12070	17.95	-0.01
Hispanic	2744	4.08	2746	4.08	0.00
Asian/Pacific Islander	1259	1.87	1259	1.87	0.00
Native American/ Alaskan	337	0.50	338	0.50	0.00
Unknown	79	0.12	79	0.12	0.00
	Communication Arts, Grade 7				
All Students	65990		66034		
Gender					
Male	33607	50.93	33641	50.94	-0.01
Female	32318	48.97	32328	48.96	0.01
Unknown	65	0.10	65	0.10	0.00
Race/Ethnicity					
White	50300	76.22	50325	76.21	0.01
Black	11553	17.51	11571	17.52	-0.01
Hispanic	2515	3.81	2515	3.81	0.00
Asian/Pacific Islander	1227	1.86	1228	1.86	0.00
Native American/ Alaskan	326	0.49	326	0.49	0.00
Unknown	69	0.10	69	0.10	0.00
	Communication Arts, Grade 8				
All Students	66079		66139		
Gender					
Male	33594	50.84	33637	50.86	-0.02
Female	32388	49.01	32405	49.00	0.01
Unknown	97	0.15	97	0.15	0.00
Race/Ethnicity					
White	50864	76.97	50901	76.96	0.01
Black	11273	17.06	11291	17.07	-0.01
Hispanic	2274	3.44	2276	3.44	0.00
Asian/Pacific Islander	1220	1.85	1223	1.85	0.00
Native American/ Alaskan	344	0.52	344	0.52	0.00
Unknown	104	0.16	104	0.16	0.00

Table 6. 3: Summary of Calibration and Census Data: Mathematics

	Mathematics, Grade 3				
	Calibration Sample		Census Data		Diff (Calib % - Census %)
	N	%	N	%	
All Students	66803		66814		
Gender					
Male	33829	50.64	33837	50.64	0.00
Female	32897	49.24	32900	49.24	0.00
Unknown	77	0.12	77	0.12	0.00
Race/Ethnicity					
White	50011	74.86	50021	74.87	-0.01
Black	11841	17.73	11841	17.72	0.01
Hispanic	3118	4.67	3119	4.67	0.00
Asian/Pacific Islander	1458	2.18	1458	2.18	0.00
Native American/ Alaskan	310	0.46	310	0.46	0.00
Unknown	65	0.10	65	0.10	0.00
	Mathematics, Grade 4				
All Students	67384		67394		
Gender					
Male	34587	51.33	34596	51.33	0.00
Female	32707	48.54	32708	48.53	0.01
Unknown	90	0.13	90	0.13	0.00
Race/Ethnicity					
White	50492	74.93	50498	74.93	0.00
Black	12123	17.99	12127	17.99	0.00
Hispanic	3010	4.47	3010	4.47	0.00
Asian/Pacific Islander	1393	2.07	1393	2.07	0.00
Native American/ Alaskan	282	0.42	282	0.42	0.00
Unknown	84	0.12	84	0.12	0.00
	Mathematics, Grade 5				
All Students	66562		66580		
Gender					
Male	34077	51.20	34089	51.20	0.00
Female	32401	48.68	32407	48.67	0.01
Unknown	84	0.13	84	0.13	0.00
Race/Ethnicity					
White	50138	75.33	50151	75.32	0.01
Black	11909	17.89	11914	17.89	0.00
Hispanic	2790	4.19	2790	4.19	0.00
Asian/Pacific Islander	1322	1.99	1322	1.99	0.00
Native American/ Alaskan	315	0.47	315	0.47	0.00
Unknown	88	0.13	88	0.13	0.00

Table 6. 3: Summary of Calibration and Census Data: Mathematics (Cont'd)

	Mathematics, Grade 6				
	Calibration Sample		Census Data		Diff (Calib % - Census %)
	N	%	N	%	
All Students	67307		67315		
Gender					
Male	34523	51.29	34528	51.29	0.00
Female	32711	48.60	32714	48.60	0.00
Unknown	73	0.11	73	0.11	0.00
Race/Ethnicity					
White	50764	75.42	50768	75.42	0.00
Black	12070	17.93	12074	17.94	-0.01
Hispanic	2765	4.11	2765	4.11	0.00
Asian/Pacific Islander	1290	1.92	1290	1.92	0.00
Native American/ Alaskan	338	0.50	338	0.50	0.00
Unknown	80	0.12	80	0.12	0.00
	Mathematics, Grade 7				
All Students	66043		66052		
Gender					
Male	33648	50.95	33655	50.95	0.00
Female	32330	48.95	32332	48.95	0.00
Unknown	65	0.10	65	0.10	0.00
Race/Ethnicity					
White	50294	76.15	50300	76.15	0.00
Black	11565	17.51	11566	17.51	0.00
Hispanic	2530	3.83	2531	3.83	0.00
Asian/Pacific Islander	1263	1.91	1263	1.91	0.00
Native American/ Alaskan	325	0.49	325	0.49	0.00
Unknown	66	0.10	67	0.10	0.00
	Mathematics, Grade 8				
All Students	66151		66166		
Gender					
Male	33649	50.87	33660	50.87	0.00
Female	32403	48.98	32407	48.98	0.00
Unknown	99	0.15	99	0.15	0.00
Race/Ethnicity					
White	50874	76.91	50886	76.91	0.00
Black	11272	17.04	11274	17.04	0.00
Hispanic	2300	3.48	2301	3.48	0.00
Asian/Pacific Islander	1255	1.90	1255	1.90	0.00
Native American/ Alaskan	345	0.52	345	0.52	0.00
Unknown	105	0.16	105	0.16	0.00

Table 6. 4: Summary of Calibration and Census Data: Science

	Science, Grade 5				
	Calibration Sample		Census Data		Diff (Calib % - Census %)
	N	%	N	%	
All Students	66555		66558		
Gender					
Male	34077	51.20	34081	51.20	0.00
Female	32395	48.67	32394	48.67	0.00
Unknown	83	0.12	83	0.12	0.00
Race/Ethnicity					
White	50137	75.33	50143	75.34	-0.01
Black	11904	17.89	11902	17.88	0.01
Hispanic	2791	4.19	2790	4.19	0.00
Asian/Pacific Islander	1321	1.98	1321	1.98	0.00
Native American/ Alaskan	314	0.47	314	0.47	0.00
Unknown	88	0.13	88	0.13	0.00
	Science, Grade 8				
All Students	66087		66101		
Gender					
Male	33603	50.85	33611	50.85	0.00
Female	32389	49.01	32395	49.01	0.00
Unknown	95	0.14	95	0.14	0.00
Race/Ethnicity					
White	50865	76.97	50875	76.97	0.00
Black	11226	16.99	11229	16.99	0.00
Hispanic	2298	3.48	2298	3.48	0.00
Asian/Pacific Islander	1253	1.90	1254	1.90	0.00
Native American/ Alaskan	345	0.52	345	0.52	0.00
Unknown	100	0.15	100	0.15	0.00

Table 6. 5: MAP Means, Standard Deviations for Raw Scores, *p*-values, Item-Total Correlation (R_{it}): Communication Arts 2010

Grade	Total Items	Total Points	Mean Raw Score (SD)	Mean <i>p</i> -value (SD)	Mean R_{it} (SD)
3	56	63	46.82 (10.15)	0.76 (0.14)	0.38 (0.09)
4	58	62	47.34 (10.59)	0.78 (0.15)	0.41 (0.10)
5	56	61	44.03 (10.04)	0.74 (0.16)	0.38 (0.09)
6	56	60	42.86 (10.27)	0.72 (0.14)	0.37 (0.09)
7	63	70	49.55 (11.00)	0.71 (0.17)	0.35 (0.11)
8	60	64	43.73 (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 2010

Grade	Total Items	Total Points	Mean Raw Score (SD)	Mean p -value (SD)	Mean R_{it} (SD)
3	55	59	45.81 (9.71)	0.80 (0.14)	0.40 (0.08)
4	62	69	49.55 (11.94)	0.74 (0.14)	0.39 (0.09)
5	58	62	44.67 (11.07)	0.73 (0.15)	0.38 (0.11)
6	58	62	43.32 (11.54)	0.71 (0.15)	0.40 (0.08)
7	61	65	41.40 (12.06)	0.66 (0.17)	0.39 (0.10)
8	61	68	38.71 (13.76)	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 2010

Grade	Total Items	Total Points	Mean Raw Score (SD)	Mean p -value (SD)	Mean R_{it} (SD)
5	63	82	49.79 (12.95)	0.64 (0.21)	0.34 (0.10)
8	66	86	46.22 (14.48)	0.60 (0.23)	0.37 (0.11)

Table 6. 8: Item Statistics: Grade 3

Communication Arts					Mathematics				
Session	Item	<i>p</i> -value	R _{it}	Omit Rate	Session	Item	<i>p</i> -value	R _{it}	Omit Rate
1	1	0.73	0.26	0.09	1	1	0.56	0.41	0.08
1	2	0.71	0.35	0.10	1	2	0.51	0.37	0.17
1	3	0.75	0.47	0.32	1	3	0.80	0.35	0.09
1	4	0.64	0.46	0.61	1	4	0.89	0.47	0.32
1	5	0.64	0.48	0.41	1	5	0.66	0.32	0.20
1	6A	0.67	0.31	0.55	1	6	0.87	0.34	0.23
1	6B	0.84	0.41	0.54	1	7	0.71	0.44	1.23
1	6C	0.97	0.30	0.54	1	8	0.77	0.50	0.21
1	10	0.69	0.25	0.35	1	9	0.77	0.44	0.85
1	11	0.47	0.20	0.40	1	10	0.69	0.13	0.23
1	12	0.42	0.20	0.41	1	11	0.94	0.37	0.34
2	1	0.71	0.46	0.24	1	12	0.93	0.45	1.24
3	1	0.98	0.32	0.10	1	13	0.91	0.35	0.24
3	2	0.94	0.45	0.17	1	14	0.76	0.37	0.75
3	3	0.89	0.35	0.46	1	15	0.92	0.35	0.20
3	4	0.81	0.44	0.42	1	16	0.91	0.27	0.24
3	5	0.92	0.40	0.68	1	17	0.83	0.49	3.16
3	6	0.90	0.18	0.18	1	18	0.91	0.42	0.19
3	7	0.68	0.21	0.22	1	19	0.85	0.40	0.31
3	8	0.89	0.44	0.40	2	1	0.89	0.29	0.12
3	9	0.67	0.40	0.84	2	4	0.79	0.38	0.67
3	10	0.90	0.47	0.71	2	5	0.95	0.34	0.79
3	11	0.85	0.40	2.10	2	6	0.67	0.40	0.85
3	12	0.71	0.31	2.43	2	7	0.55	0.47	1.26
3	13	0.83	0.45	0.20	2	8	0.89	0.35	1.07
3	14	0.92	0.44	0.27	2	9	0.78	0.48	1.37
3	15	0.81	0.50	0.46	2	10	0.92	0.37	0.10
3	16	0.60	0.36	0.73	2	11	0.95	0.38	0.14
3	17	0.77	0.32	0.42	2	12	0.92	0.38	0.15
3	18	0.70	0.41	0.58	2	13	0.91	0.43	0.19
3	19	0.85	0.42	0.84	2	14	0.88	0.42	0.24
3	20	0.79	0.32	0.24	2	15	0.66	0.39	0.39
3	21	0.78	0.35	0.33	2	16	0.83	0.38	1.86
3	22	0.71	0.45	0.39	2	17	0.88	0.43	0.54
3	23	0.63	0.39	0.65	2	19	0.86	0.33	0.45

Table 6. 8: Item Statistics: Grade 3 (Cont'd)

Communication Arts					Mathematics				
Session	Item	<i>p</i> -value	R _{it}	Omit Rate	Session	Item	<i>p</i> -value	R _{it}	Omit Rate
3	24	0.40	0.20	0.32	2	20	0.80	0.48	0.24
3	25	0.87	0.45	0.45	2	21	0.92	0.41	0.34
3	26	0.85	0.48	0.63	2	22	0.76	0.35	0.80
3	27	0.62	0.47	0.72	2	23	0.84	0.51	0.60
3	28	0.74	0.31	0.91	2	24	0.69	0.54	0.34
3	29	0.97	0.36	0.26	2	25	0.51	0.32	0.37
3	30	0.75	0.46	0.32	2	26	0.90	0.41	0.36
3	31	0.86	0.46	0.41	2	30	0.79	0.36	0.33
3	32	0.61	0.31	0.79	3	1	0.82	0.53	0.11
3	33	0.88	0.41	0.40	3	2	0.57	0.50	0.22
3	34	0.80	0.43	0.56	3	3	0.48	0.58	0.25
3	35	0.45	0.23	0.78	3	4	0.31	0.39	0.47
3	36	0.61	0.34	1.11	3	5	0.90	0.45	0.42
3	37	0.92	0.37	0.68	3	6	0.95	0.30	0.32
3	38	0.88	0.51	0.88	3	7	0.73	0.50	0.38
3	39	0.77	0.39	1.00	3	8	0.92	0.43	0.68
4	1	0.84	0.43	0.16	3	9	0.80	0.40	0.28
4	2	0.76	0.41	0.25	3	10	0.87	0.42	0.41
4	3	0.59	0.40	1.02	3	11	0.88	0.37	0.36
4	4	0.77	0.46	0.28	3	12	0.92	0.29	0.18
4	5	0.56	0.43	0.45					

Table 6. 9: Item Statistics: Grade 4

Communication Arts					Mathematics				
Session	Item	<i>p</i> -value	R _{it}	Omit Rate	Session	Item	<i>p</i> -value	R _{it}	Omit Rate
1	1	0.86	0.37	0.05	1	1	0.95	0.31	0.10
1	2	0.88	0.50	0.06	1	2	0.92	0.35	0.09
1	3	0.42	0.47	0.45	1	3	0.70	0.34	0.14
1	4	0.45	0.45	0.41	1	4	0.76	0.43	0.22
1	5	0.78	0.53	0.54	1	5	0.86	0.43	0.15
1	6A	0.73	0.49	0.42	1	6	0.75	0.49	0.20
1	6B	0.93	0.40	0.42	1	7	0.82	0.25	0.40
1	6C	0.97	0.32	0.42	1	8	0.80	0.28	0.13
1	7	0.91	0.38	0.30	1	9	0.82	0.44	0.13
1	8	0.53	0.17	0.35	1	10	0.71	0.31	0.40
1	9	0.56	0.33	1.13	1	11	0.77	0.42	0.18
1	10	0.75	0.39	0.37	1	12	0.79	0.55	0.52
1	11	0.78	0.38	0.48	1	13	0.94	0.29	0.13
1	12	0.21	0.05	0.38	1	14	0.55	0.46	0.25
2	1	0.93	0.40	0.12	1	15	0.57	0.42	0.42
2	2	0.84	0.48	0.14	1	16	0.74	0.26	0.66
2	3	0.80	0.45	0.13	1	17	0.50	0.37	0.17
2	4	0.90	0.45	0.17	1	18	0.67	0.37	0.30
2	5	0.72	0.45	0.25	1	19	0.69	0.43	0.22
2	6	0.82	0.50	0.72	1	20	0.80	0.36	0.48
2	7	0.93	0.40	0.16	1	21	0.80	0.34	0.20
2	8	0.77	0.48	0.43	1	22	0.45	0.61	0.31
2	9	0.66	0.37	0.24	2	1	0.86	0.23	0.11
2	10	0.86	0.51	0.43	2	2	0.70	0.37	0.28
2	11	0.95	0.29	0.30	2	3	0.67	0.34	0.90
2	12	0.92	0.36	0.33	2	4	0.53	0.41	1.46
2	13	0.80	0.28	0.42	2	5	0.68	0.49	0.57
2	14	0.86	0.49	0.73	2	6	0.80	0.36	0.70
2	15	0.76	0.40	0.87	2	8	0.85	0.32	1.37
2	16	0.76	0.44	1.07	2	9	0.73	0.55	1.81
2	17	0.71	0.29	1.34	2	10	0.78	0.44	1.98
2	18	0.67	0.30	1.52	2	12	0.86	0.36	0.11
2	19	0.72	0.44	1.62	2	13	0.75	0.37	0.30
2	20	0.85	0.50	0.13	2	14	0.93	0.26	0.10
2	21	0.95	0.44	0.17	2	15	0.90	0.27	0.19

Table 6. 9: Item Statistics: Grade 4 (Cont'd)

Communication Arts					Mathematics				
Session	Item	<i>p</i> -value	R _{it}	Omit Rate	Session	Item	<i>p</i> -value	R _{it}	Omit Rate
2	22	0.89	0.52	0.24	2	16	0.89	0.34	0.19
2	23	0.54	0.28	0.37	2	18	0.79	0.47	0.27
2	24	0.82	0.40	1.13	2	19	0.86	0.35	0.23
2	25	0.95	0.44	0.23	2	20	0.83	0.28	0.15
2	26	0.96	0.46	0.17	2	21	0.89	0.37	0.33
2	27	0.91	0.50	0.27	2	22	0.71	0.42	0.41
2	28	0.59	0.41	0.28	2	23	0.94	0.41	0.29
2	29	0.83	0.52	3.82	2	24	0.83	0.51	0.85
2	30	0.83	0.37	0.23	2	25	0.91	0.39	0.21
2	31	0.67	0.37	0.79	2	26	0.90	0.43	0.29
2	32	0.72	0.47	0.49	2	27	0.64	0.31	0.27
2	33	0.76	0.58	0.42	2	30	0.82	0.39	1.07
2	34	0.78	0.37	0.51	2	31	0.84	0.44	0.17
2	35	0.75	0.48	1.26	3	1	0.63	0.52	0.29
2	36	0.92	0.32	0.81	3	2	0.46	0.49	0.34
2	37	0.88	0.49	0.69	3	3	0.75	0.47	0.31
2	38	0.87	0.49	0.77	3	4	0.43	0.49	0.32
2	39	0.86	0.50	0.92	3	5	0.69	0.37	0.64
3	1	0.73	0.52	0.16	3	6	0.87	0.21	0.27
3	2	0.97	0.18	0.18	3	7	0.40	0.36	0.39
3	3	0.59	0.50	0.19	3	8	0.89	0.34	0.55
3	4	0.86	0.51	0.33	3	9	0.49	0.25	0.22
3	5	0.61	0.36	0.19	3	10	0.65	0.51	0.52
					3	11	0.54	0.23	0.36
					3	12	0.66	0.55	0.45
					3	13	0.61	0.37	0.37
					3	14	0.78	0.44	0.75

Table 6. 10: Item Statistics: Grade 5

Communication Arts					Mathematics					Science				
Session	Item	<i>p</i> -value	R _{it}	Omit Rate	Session	Item	<i>p</i> -value	R _{it}	Omit Rate	Session	Item	<i>p</i> -value	R _{it}	Omit Rate
1	1	0.93	0.26	0.06	1	1	0.83	0.34	0.07	1	1	0.77	0.32	0.21
1	2	0.87	0.37	0.07	1	2	0.68	0.42	0.11	1	2	0.56	0.45	0.26
1	3	0.66	0.38	0.42	1	3	0.55	0.42	0.13	1	3	0.40	0.46	0.47
1	4A	0.45	0.33	0.30	1	4	0.73	0.40	0.18	1	4	0.53	0.55	0.37
1	4B	0.97	0.34	0.30	1	5	0.82	0.41	0.15	1	5	0.32	0.41	1.96
1	5	0.41	0.30	0.33	1	6	0.69	0.32	0.12	1	6	0.47	0.42	0.43
1	6	0.60	0.30	0.33	1	7	0.72	0.38	0.18	1	7	0.51	0.42	0.46
1	7	0.86	0.36	0.18	1	8	0.88	0.37	0.17	1	8	0.46	0.49	0.80
1	8	0.55	0.32	0.26	1	9	0.76	0.15	0.20	1	9	0.44	0.53	0.48
1	9	0.58	0.28	1.87	1	10	0.40	0.40	0.27	1	10	0.39	0.45	0.45
1	10	0.62	0.34	0.22	1	11	0.70	0.62	0.10	1	11	0.46	0.49	0.79
1	11	0.41	0.26	0.25	1	12	0.77	0.53	0.22	1	12	0.25	0.48	0.95
1	12	0.46	0.12	0.22	1	13	0.89	0.31	0.19	1	13	0.23	0.44	0.88
2	1	0.66	0.28	0.15	1	14	0.74	0.17	0.18	2	2	0.97	0.21	0.12
2	2	0.51	0.34	0.22	1	15	0.45	0.39	0.21	2	3	0.90	0.34	0.13
2	3	0.91	0.42	0.22	1	16	0.94	0.30	0.18	2	4	0.93	0.30	0.12
2	4	0.91	0.35	0.19	1	17	0.86	0.29	0.09	2	5	0.89	0.17	0.19
2	5	0.75	0.40	0.35	1	18	0.70	0.51	0.12	2	6	0.78	0.36	0.45
2	6	0.89	0.41	0.48	1	19	0.57	0.31	0.13	2	7	0.74	0.32	0.32
2	7	0.76	0.43	0.58	1	20	0.70	0.60	0.26	2	8	0.72	0.29	0.36
2	8	0.70	0.45	0.73	1	21	0.77	0.35	0.40	2	9	0.71	0.20	0.30
2	9	0.85	0.53	1.04	2	1	0.74	0.37	0.16	2	10	0.77	0.50	0.69
2	10	0.75	0.50	4.10	2	2	0.49	0.31	0.27	2	11	0.72	0.45	1.30
2	11	0.55	0.38	1.49	2	3	0.83	0.38	2.10	2	12	0.82	0.31	0.26
2	12	0.87	0.51	1.88	2	4	0.51	0.33	0.26	2	14	0.69	0.36	0.84
2	13	0.91	0.39	2.98	2	5	0.64	0.32	0.51	2	15	0.67	0.28	0.64
2	14	0.74	0.40	3.43	2	6	0.61	0.44	0.53	2	16	0.65	0.31	0.81
2	15	0.80	0.35	4.27	2	7	0.64	0.48	0.84	2	17	0.65	0.19	0.27
2	16	0.96	0.36	4.53	2	9	0.66	0.49	1.54	2	18	0.77	0.33	0.37
2	17	0.72	0.47	0.14	2	10	0.60	0.14	0.24	2	19	0.65	0.31	0.26
2	18	0.81	0.34	0.17	2	11	0.92	0.23	0.19	2	20	0.74	0.31	0.52
2	19	0.94	0.45	0.22	2	12	0.84	0.39	0.26	2	21	0.63	0.31	0.46
2	20	0.85	0.51	0.34	2	13	0.44	0.33	0.51	2	22	0.71	0.28	0.30
2	21	0.78	0.50	0.17	2	14	0.62	0.61	0.29	2	24	0.51	0.31	0.41
2	22	0.56	0.38	0.21	2	15	0.89	0.46	0.64	2	25	0.42	0.32	0.44

Table 6. 10: Item Statistics: Grade 5 (Cont'd)

Communication Arts					Mathematics					Science				
Session	Item	<i>p</i> -value	R _{it}	Omit Rate	Session	Item	<i>p</i> -value	R _{it}	Omit Rate	Session	Item	<i>p</i> -value	R _{it}	Omit Rate
2	23	0.78	0.45	0.32	2	19	0.82	0.37	0.24	2	26	0.91	0.23	0.25
2	24	0.98	0.27	0.13	2	21	0.82	0.52	0.25	2	27	0.86	0.35	0.23
2	25	0.93	0.40	0.18	2	22	0.99	0.21	0.23	2	28	0.89	0.40	0.21
2	26	0.65	0.32	0.34	2	23	0.66	0.44	0.19	2	29	0.81	0.15	2.70
2	27	0.32	0.06	0.30	2	25	0.74	0.47	0.17	2	30	0.92	0.38	0.26
2	28	0.92	0.41	0.18	2	26	0.98	0.23	0.26	2	31	0.88	0.35	0.40
2	29	0.87	0.51	0.33	2	29	0.93	0.29	0.33	2	32	0.74	0.20	0.82
2	30	0.78	0.43	0.18	2	30	0.91	0.29	0.29	2	33	0.87	0.38	0.46
2	31	0.91	0.42	0.19	3	1	0.52	0.59	0.32	2	34	0.87	0.43	1.82
2	32	0.71	0.44	0.31	3	2	0.51	0.46	0.33	2	35	0.69	0.30	0.52
2	33	0.73	0.39	0.24	3	3	0.73	0.45	0.21	2	36	0.64	0.33	0.30
2	34	0.71	0.43	0.30	3	4	0.78	0.57	0.30	2	37	0.71	0.36	0.94
2	35	0.75	0.47	0.42	3	5	0.85	0.41	0.25	2	38	0.44	0.29	1.94
2	36	0.79	0.46	0.75	3	6	0.69	0.34	0.28	2	39	0.66	0.44	2.64
2	37	0.72	0.47	0.50	3	7	0.69	0.34	0.72	2	40	0.59	0.32	0.19
2	38	0.72	0.39	0.75	3	8	0.91	0.24	0.29	2	41	0.77	0.41	0.39
3	1	0.71	0.45	0.14	3	9	0.59	0.49	0.33	2	42	0.60	0.43	0.73
3	2	0.74	0.48	0.22	3	10	0.61	0.41	0.41	2	43	0.56	0.29	0.37
3	3	0.74	0.32	0.15	3	11	0.89	0.19	0.21	2	44	0.47	0.42	0.62
3	4	0.58	0.29	0.18	3	12	0.68	0.34	0.34	2	45	0.14	0.01	0.77
3	5	0.70	0.44	0.24	3	13	0.62	0.31	0.45	3	1	0.78	0.30	0.20
					3	14	0.90	0.37	0.43	3	2	0.51	0.28	0.42
					3	15	0.90	0.24	0.16	3	3	0.75	0.38	0.30
										3	4	0.96	0.27	0.23
										3	5	0.40	0.30	0.60
										3	6	0.20	0.27	1.53
										3	7	0.11	0.28	0.82
										3	8	0.66	0.37	0.46

Table 6. 11: Item Statistics: Grade 6

Communication Arts					Mathematics				
Session	Item	<i>p</i> -value	R _{it}	Omit Rate	Session	Item	<i>p</i> -value	R _{it}	Omit Rate
1	1	0.64	0.21	0.07	1	1	0.57	0.51	0.12
1	2	0.81	0.40	0.06	1	2	0.53	0.43	0.10
1	3	0.91	0.42	0.26	1	3	0.82	0.32	0.11
1	4	0.78	0.46	0.33	1	4	0.60	0.46	0.09
1	5	0.59	0.39	0.50	1	5	0.85	0.27	0.11
1	6A	0.57	0.57	0.50	1	6	0.82	0.43	0.15
1	6B	0.83	0.23	0.48	1	7	0.66	0.39	0.19
1	7	0.63	0.30	0.14	1	8	0.84	0.36	0.15
1	8	0.60	0.26	0.17	1	9	0.89	0.34	0.10
1	9	0.49	0.28	0.35	1	10	0.88	0.43	0.14
1	10	0.83	0.38	0.18	1	11	0.63	0.31	0.35
1	11	0.52	0.15	0.24	1	12	0.58	0.42	0.16
1	12	0.62	0.39	0.19	1	13	0.89	0.34	0.17
2	1	0.64	0.41	0.13	1	14	0.83	0.49	0.20
2	2	0.89	0.30	0.34	1	15	0.53	0.46	0.19
2	3	0.91	0.34	0.13	1	16	0.83	0.27	0.16
2	4	0.64	0.45	0.20	1	17	0.69	0.25	0.13
2	5	0.49	0.33	0.20	1	18	0.66	0.27	0.19
2	6	0.43	0.30	0.18	1	19	0.77	0.41	0.16
2	7	0.80	0.48	0.30	2	1	0.81	0.22	0.10
2	8	0.75	0.36	0.47	2	2	0.82	0.30	0.21
2	9	0.87	0.34	0.18	2	3	0.75	0.46	1.44
2	10	0.63	0.27	0.19	2	4	0.68	0.40	0.25
2	11	0.87	0.35	0.19	2	5	0.69	0.41	0.53
2	12	0.85	0.25	0.17	2	6	0.72	0.47	0.56
2	13	0.78	0.39	0.28	2	7	0.73	0.39	1.00
2	14	0.84	0.34	0.31	2	8	0.40	0.39	1.02
2	15	0.77	0.48	0.44	2	9	0.60	0.27	1.59
2	16	0.86	0.47	0.57	2	10	0.88	0.32	0.12
2	17	0.89	0.46	0.72	2	11	0.90	0.43	0.15
2	18	0.79	0.52	0.56	2	12	0.47	0.47	0.17
2	19	0.55	0.48	0.83	2	13	0.54	0.49	0.27
2	20	0.86	0.45	0.83	2	14	0.82	0.49	0.56
2	21	0.69	0.42	1.03	2	15	0.77	0.50	0.67
2	22	0.53	0.25	1.24	2	16	0.89	0.39	0.23

Table 6. 11: Item Statistics: Grade 6 (Cont'd)

Communication Arts					Mathematics				
Session	Item	<i>p</i> -value	R _{it}	Omit Rate	Session	Item	<i>p</i> -value	R _{it}	Omit Rate
2	23	0.78	0.34	1.59	2	17	0.93	0.20	0.20
2	24	0.86	0.52	0.22	2	19	0.55	0.47	0.45
2	25	0.78	0.24	0.25	2	20	0.44	0.41	0.32
2	26	0.88	0.48	0.42	2	22	0.58	0.35	0.24
2	27	0.51	0.32	2.55	2	23	0.86	0.40	0.29
2	28	0.82	0.39	3.95	2	25	0.92	0.37	0.47
2	29	0.66	0.40	0.29	2	26	0.49	0.38	0.32
2	30	0.88	0.48	0.58	2	27	0.91	0.40	0.28
2	31	0.50	0.28	0.33	2	28	0.69	0.45	0.27
2	32	0.52	0.38	0.56	2	31	0.66	0.50	0.36
2	33	0.79	0.42	0.60	3	1	0.51	0.61	0.54
2	34	0.63	0.45	0.68	3	2	0.53	0.52	1.22
2	35	0.79	0.50	0.77	3	3	0.34	0.52	0.50
2	36	0.53	0.28	0.82	3	4	0.72	0.44	0.63
2	37	0.87	0.40	1.03	3	5	0.79	0.50	0.26
2	38	0.58	0.36	1.27	3	6	0.70	0.38	0.39
2	39	0.66	0.36	1.50	3	7	0.61	0.40	0.26
3	1	0.76	0.41	0.37	3	8	0.65	0.45	0.37
3	2	0.88	0.36	0.16	3	9	0.90	0.40	0.71
3	3	0.86	0.29	0.18	3	10	0.71	0.38	0.33
3	4	0.54	0.26	0.20	3	11	0.82	0.41	0.26
					3	12	0.82	0.37	0.25
					3	13	0.93	0.38	0.38

Table 6. 12: Item Statistics: Grade 7

Communication Arts					Mathematics				
Session	Item	<i>p</i> -value	R _{it}	Omit Rate	Session	Item	<i>p</i> -value	R _{it}	Omit Rate
1	1	0.78	0.28	0.07	1	1	0.85	0.48	0.08
1	2	0.61	0.29	0.12	1	2	0.67	0.47	0.16
1	3	0.60	0.44	0.57	1	3	0.75	0.36	0.16
1	4	0.61	0.34	0.46	1	4	0.49	0.37	0.17
1	5	0.66	0.51	0.52	1	5	0.75	0.31	0.25
1	6A	0.87	0.49	1.05	1	6	0.77	0.52	0.23
1	6B	0.82	0.34	1.05	1	7	0.50	0.42	0.37
1	6C	0.98	0.19	1.05	1	8	0.53	0.41	0.22
1	7	0.84	0.34	0.10	1	9	0.74	0.26	0.21
1	8	0.83	0.26	0.29	1	10	0.56	0.30	0.28
1	9	0.80	0.33	0.15	1	11	0.80	0.42	0.27
1	10	0.55	0.19	1.54	1	12	0.79	0.27	0.18
1	11	0.59	0.37	0.22	1	13	0.23	0.27	0.46
1	12	0.72	0.42	0.47	1	14	0.60	0.44	0.23
1	13	0.49	0.24	0.72	1	15	0.67	0.50	0.18
1	14	0.20	0.11	0.73	1	16	0.57	0.38	0.22
1	15	0.35	0.04	0.58	1	17	0.43	0.24	0.40
1	16	0.33	0.20	0.71	1	18	0.67	0.43	0.18
2	1	0.76	0.52	0.34	2	1	0.80	0.28	0.20
3	1	0.95	0.36	0.17	2	2	0.66	0.44	0.36
3	2	0.97	0.31	0.16	2	3	0.56	0.45	0.95
3	3	0.52	0.16	0.18	2	4	0.54	0.52	0.95
3	4	0.93	0.31	0.22	2	5	0.69	0.46	0.22
3	5	0.92	0.28	0.16	2	6	0.74	0.49	0.42
3	6	0.79	0.30	0.20	2	7	0.69	0.41	1.15
3	7	0.94	0.40	0.26	2	8	0.73	0.34	0.69
3	8	0.88	0.31	0.16	2	9	0.74	0.51	1.22
3	9	0.64	0.31	0.26	2	10	0.84	0.36	0.31
3	10	0.71	0.25	0.47	2	11	0.74	0.45	0.25
3	11	0.86	0.27	0.23	2	12	0.73	0.51	0.32
3	12	0.51	0.28	0.26	2	13	0.81	0.38	0.41
3	13	0.88	0.39	0.31	2	14	0.71	0.44	0.50
3	14	0.73	0.38	0.48	2	15	0.71	0.19	0.58
3	15	0.85	0.39	0.49	2	16	0.97	0.13	0.20
3	16	0.65	0.36	0.67	2	17	0.60	0.47	1.49

Table 6. 12: Item Statistics: Grade 7 (Cont'd)

Communication Arts					Mathematics				
Session	Item	<i>p</i> -value	R _{it}	Omit Rate	Session	Item	<i>p</i> -value	R _{it}	Omit Rate
3	17	0.52	0.28	0.73	2	18	0.85	0.27	0.26
3	18	0.82	0.50	0.67	2	19	0.79	0.42	0.40
3	19	0.80	0.46	0.75	2	20	0.84	0.45	0.52
3	20	0.59	0.44	0.86	2	21	0.75	0.40	1.19
3	21	0.78	0.52	0.88	2	22	0.60	0.35	0.42
3	22	0.62	0.38	0.96	2	23	0.84	0.25	0.39
3	23	0.69	0.37	1.13	2	25	0.66	0.33	0.33
3	24	0.78	0.36	1.00	2	26	0.44	0.48	0.36
3	25	0.85	0.23	0.56	2	27	0.85	0.32	0.29
3	26	0.52	0.21	0.37	2	28	0.69	0.41	0.32
3	27	0.86	0.48	0.33	2	29	0.51	0.18	0.58
3	28	0.80	0.33	0.31	2	30	0.95	0.29	0.32
3	29	0.67	0.50	0.91	2	31	0.57	0.36	0.25
3	30	0.81	0.39	0.26	2	32	0.53	0.49	0.31
3	31	0.78	0.38	0.39	3	1	0.80	0.41	0.60
3	32	0.80	0.40	0.46	3	2	0.17	0.42	1.03
3	33	0.66	0.41	0.38	3	3	0.27	0.54	3.41
3	34	0.64	0.44	0.56	3	4	0.30	0.53	1.13
3	35	0.65	0.39	1.20	3	5	0.73	0.45	0.28
3	36	0.82	0.48	0.57	3	6	0.83	0.35	0.30
3	37	0.74	0.42	0.58	3	7	0.72	0.34	0.45
3	38	0.73	0.46	0.59	3	8	0.53	0.36	0.46
3	39	0.68	0.48	1.04	3	9	0.50	0.38	0.28
4	1	0.82	0.35	0.35	3	10	0.64	0.48	0.34
4	2	0.43	0.26	0.37	3	11	0.42	0.22	0.27
4	3	0.79	0.44	0.73	3	12	0.63	0.38	0.46
4	4	0.72	0.38	1.23					
4	5	0.35	0.04	0.31					

Table 6. 13: Item Statistics: Grade 8

Communication Arts					Mathematics					Science				
Session	Item	<i>p</i> -value	R _{it}	Omit Rate	Session	Item	<i>p</i> -value	R _{it}	Omit Rate	Session	Item	<i>p</i> -value	R _{it}	Omit Rate
1	1	0.77	0.27	0.28	1	1	0.67	0.36	0.39	1	1	0.30	0.52	0.97
1	2	0.59	0.21	0.19	1	2	0.67	0.34	0.28	1	2	0.43	0.54	0.97
1	3	0.45	0.46	3.61	1	3	0.41	0.33	0.17	1	3	0.47	0.56	0.54
1	4	0.57	0.42	1.13	1	4	0.62	0.24	0.22	1	4	0.51	0.51	0.79
1	5	0.41	0.37	0.87	1	5	0.49	0.34	0.23	1	5	0.43	0.47	2.43
1	6A	0.69	0.44	1.22	1	6	0.64	0.39	0.20	1	6	0.32	0.45	0.76
1	6B	0.87	0.35	1.22	1	7	0.87	0.38	0.25	1	7	0.40	0.55	1.82
1	6C	0.91	0.29	1.22	1	8	0.93	0.35	0.24	1	8	0.28	0.34	1.53
1	7	0.62	0.39	0.26	1	9	0.35	0.13	0.39	1	9	0.26	0.33	2.98
1	8	0.39	0.34	0.60	1	10	0.40	0.27	0.39	1	10	0.25	0.54	4.22
1	9	0.75	0.42	0.29	1	11	0.52	0.50	0.32	1	11	0.25	0.45	0.78
1	10	0.65	0.36	1.33	1	12	0.32	0.41	0.54	1	12	0.19	0.44	1.72
1	11	0.54	0.23	2.01	1	13	0.45	0.34	0.26	1	13	0.14	0.42	2.27
1	12	0.60	0.30	0.38	1	14	0.38	0.28	0.29	1	14	0.09	0.25	3.23
1	13	0.47	0.19	0.63	1	15	0.53	0.45	0.85	2	1	0.92	0.20	0.22
1	14	0.46	0.24	0.96	1	16	0.49	0.15	0.48	2	2	0.92	0.32	0.24
1	15	0.22	0.17	0.36	1	17	0.87	0.30	0.23	2	3	0.90	0.38	0.27
1	16	0.47	0.13	0.43	1	18	0.57	0.41	0.42	2	4	0.87	0.28	0.26
2	1	0.50	0.19	0.24	1	19	0.44	0.22	0.36	2	5	0.97	0.27	0.26
2	2	0.87	0.32	0.21	1	20	0.37	0.66	3.65	2	6	0.83	0.36	0.30
2	3	0.65	0.26	0.37	2	1	0.57	0.31	0.38	2	7	0.72	0.33	0.30
2	4	0.82	0.36	0.20	2	2	0.78	0.43	0.28	2	8	0.72	0.35	0.29
2	5	0.84	0.47	0.56	2	3	0.67	0.41	0.67	2	9	0.85	0.36	0.36
2	6	0.83	0.38	0.27	2	4	0.84	0.38	0.23	2	10	0.87	0.35	0.42
2	7	0.71	0.44	0.30	2	5	0.65	0.36	0.38	2	11	0.74	0.34	0.36
2	8	0.69	0.37	0.31	2	6	0.76	0.38	0.38	2	12	0.77	0.34	0.36
2	9	0.64	0.44	0.34	2	9	0.80	0.45	0.26	2	13	0.70	0.47	0.34
2	10	0.84	0.35	0.28	2	10	0.81	0.36	0.31	2	14	0.78	0.38	0.43
2	11	0.87	0.47	0.30	2	11	0.77	0.41	0.37	2	15	0.81	0.48	0.45
2	12	0.83	0.16	0.42	2	12	0.81	0.40	0.30	2	16	0.66	0.53	0.45
2	13	0.69	0.52	0.57	2	13	0.71	0.42	0.43	2	17	0.75	0.23	0.36
2	15	0.79	0.43	0.31	2	14	0.73	0.47	1.17	2	18	0.59	0.38	5.17
2	16	0.74	0.45	0.31	2	15	0.73	0.51	0.48	2	20	0.74	0.44	0.61
2	17	0.50	0.20	0.30	2	16	0.55	0.34	0.40	2	21	0.60	0.27	0.51
2	18	0.92	0.43	0.36	2	17	0.97	0.18	0.30	2	22	0.48	0.43	1.05

Table 6. 13: Item Statistics: Grade 8 (Cont'd)

Communication Arts					Mathematics					Science				
Session	Item	<i>p</i> -value	R _{it}	Omit Rate	Session	Item	<i>p</i> -value	R _{it}	Omit Rate	Session	Item	<i>p</i> -value	R _{it}	Omit Rate
2	19	0.55	0.36	0.82	2	18	0.78	0.34	0.28	2	23	0.92	0.26	0.73
2	20	0.87	0.42	0.33	2	19	0.37	0.44	0.36	2	25	0.52	0.26	0.48
2	21	0.74	0.42	0.43	2	20	0.73	0.42	0.41	2	26	0.88	0.38	0.35
2	22	0.66	0.38	0.32	2	21	0.78	0.35	0.72	2	27	0.92	0.35	0.37
2	23	0.88	0.36	0.33	2	22	0.58	0.46	1.44	2	28	0.86	0.32	0.35
2	24	0.81	0.36	0.47	2	23	0.78	0.37	0.35	2	29	0.81	0.39	0.40
2	25	0.80	0.42	0.32	2	24	0.60	0.42	0.50	2	30	0.61	0.34	0.55
2	26	0.60	0.41	0.47	2	25	0.54	0.42	0.83	2	31	0.56	0.41	0.63
2	27	0.60	0.39	0.54	2	26	0.51	0.52	0.97	2	32	0.69	0.48	0.40
2	28	0.81	0.43	0.30	2	27	0.59	0.33	0.52	2	33	0.66	0.32	0.49
2	29	0.91	0.39	0.37	2	28	0.51	0.55	0.44	2	34	0.49	0.13	0.46
2	30	0.75	0.52	0.35	2	29	0.44	0.39	0.58	2	35	0.62	0.36	0.62
2	31	0.84	0.50	0.43	2	31	0.49	0.38	0.60	2	36	0.73	0.29	0.54
2	32	0.70	0.51	1.23	3	1	0.40	0.58	2.31	2	37	0.49	0.27	0.74
2	33	0.85	0.50	1.71	3	2	0.52	0.53	0.63	2	38	0.59	0.14	0.38
2	34	0.94	0.36	0.40	3	3	0.61	0.60	1.85	2	39	0.52	0.36	4.23
2	35	0.95	0.37	0.48	3	4	0.34	0.48	0.52	2	40	0.63	0.38	0.48
2	36	0.81	0.49	0.50	3	5	0.38	0.35	0.41	2	41	0.78	0.40	0.77
2	37	0.65	0.38	0.63	3	6	0.34	0.50	0.44	2	42	0.69	0.42	0.87
2	38	0.68	0.36	0.70	3	7	0.37	0.45	0.60	2	43	0.74	0.36	0.54
2	39	0.55	0.42	0.84	3	8	0.50	0.58	0.52	2	44	0.36	0.17	0.72
3	1	0.59	0.35	0.28	3	9	0.43	0.46	0.44	2	45	0.33	0.32	0.84
3	2	0.61	0.36	0.32	3	10	0.74	0.47	0.34	3	1	0.34	0.08	0.80
3	3	0.80	0.37	0.33	3	11	0.60	0.39	0.43	3	2	0.36	0.38	0.92
3	4	0.79	0.44	0.34	3	12	0.48	0.60	0.53	3	3	0.25	0.46	1.61
					3	13	0.41	0.17	0.39	3	4	0.72	0.44	7.32
										3	5	0.33	0.29	1.99
										3	6	0.63	0.61	1.18
										3	7	0.69	0.48	1.44
										3	8	0.45	0.51	2.25
										3	9	0.56	0.49	4.35

Table 6. 14: Item Fit Statistics for Misfitting Items

Content	Grade	Session	Item	Chi-Square	DF	Total N	Z	Obs	Pred	Obs-Pred
CA	3	3	34	795.56	7	66235	210.75	0.80	0.80	-0.01
CA	4	2	30	976.20	7	66827	259.03	0.82	0.83	0.00
CA	7	3	28	806.50	7	65765	213.68	0.80	0.80	0.00
CA	8	2	12	959.19	7	65785	254.48	0.83	0.83	0.00
MA	3	3	2	1468.54	17	65686	248.94	0.56	0.57	0.00
MA	6	1	5	721.80	7	67062	191.04	0.85	0.85	0.00
MA	6	1	17	847.04	7	67047	224.51	0.69	0.69	0.00
MA	6	3	4	2189.43	17	66710	372.57	0.72	0.72	0.00
MA	7	1	12	699.29	7	65841	185.02	0.79	0.79	0.00

Table 6. 15: LOSS and HOSS Values by Grade and Content Area

Grade	Communication Arts		Mathematics		Science	
	LOSS	HOSS	LOSS	HOSS	LOSS	HOSS
3	455	790	450	780		
4	470	820	465	805		
5	485	840	480	830	470	855
6	505	855	495	845		
7	515	865	510	860		
8	530	875	525	885	540	895

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

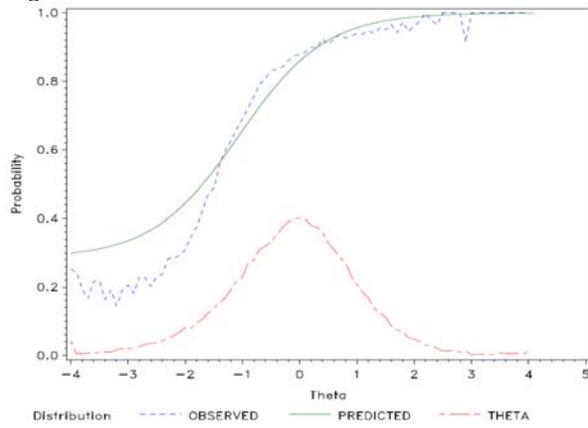


Figure 6. 2: Item characteristic curve for Grade 4 Communication Arts, Session 2 Item 30

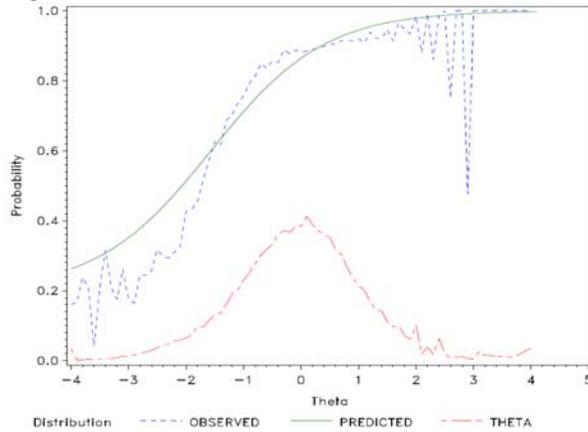


Figure 6. 3: Item characteristic curve for Grade 7 Communication Arts, Session 3 Item 28

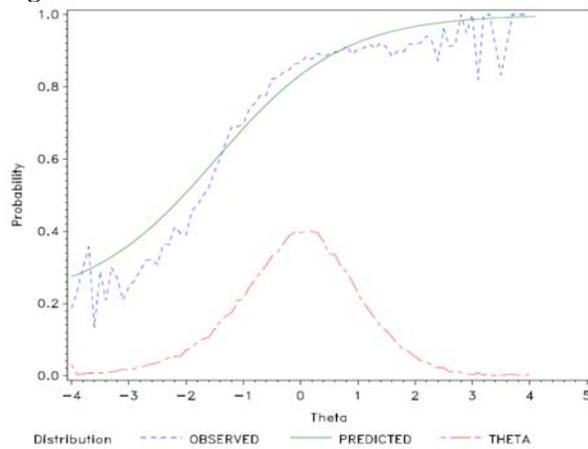


Figure 6. 4: Item characteristic curve for Grade 8 Communication Arts, Session 2 Item 12

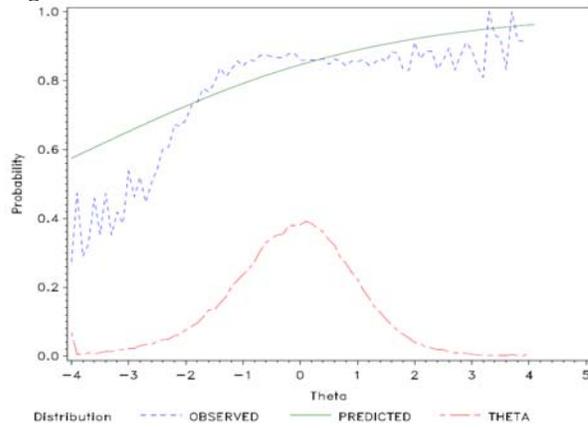


Figure 6. 5: Item characteristic curve for Grade 3 Mathematics, Session 3 Item 2

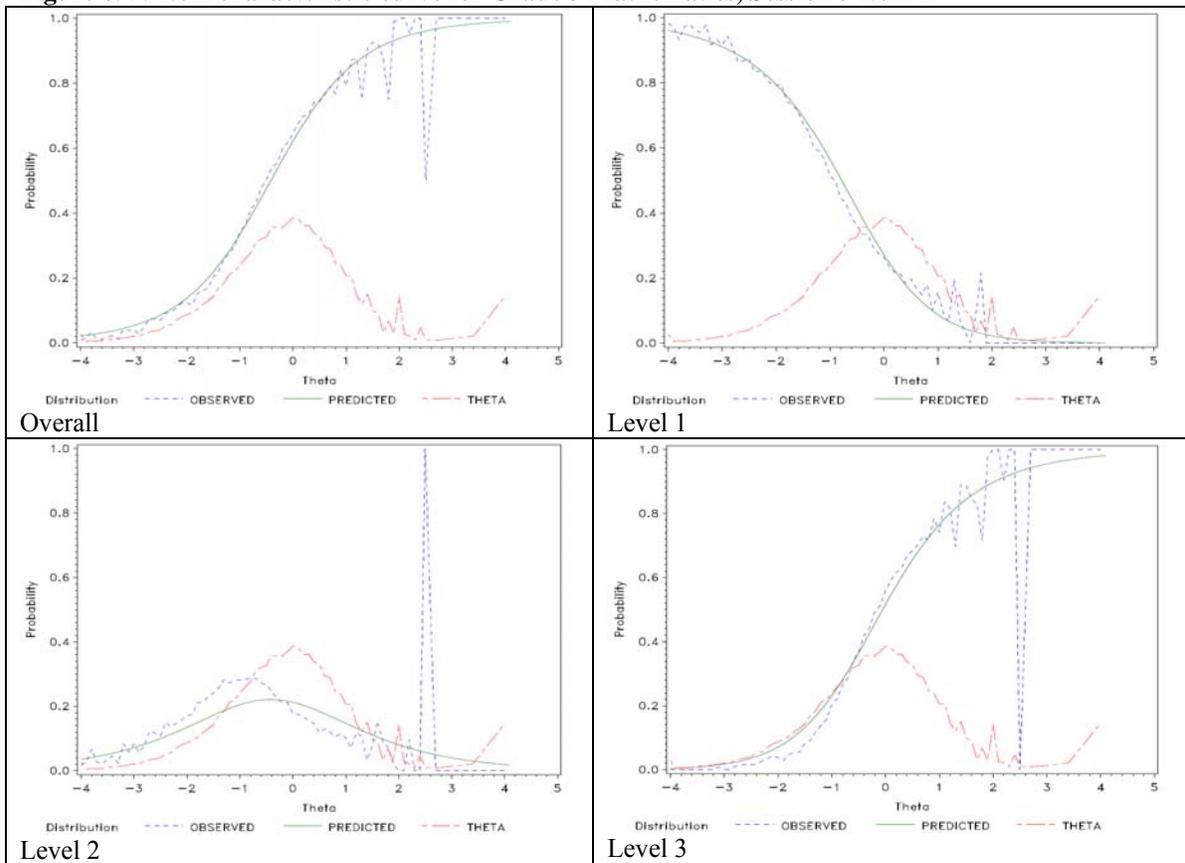


Figure 6. 6: Item characteristic curve for Grade 6 Mathematics, Session 1 Item 5

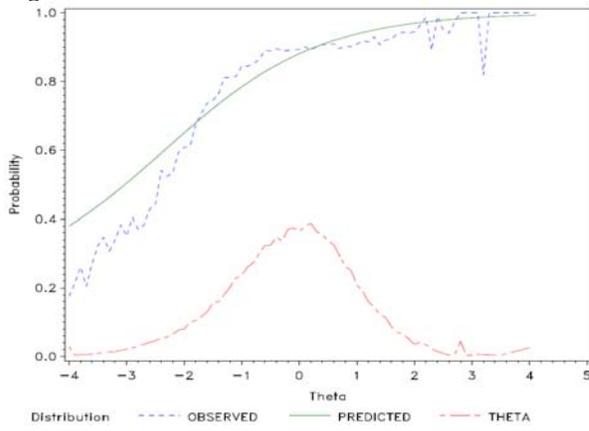


Figure 6. 7: Item characteristic curve for Grade 6 Mathematics, Session 1 Item 17

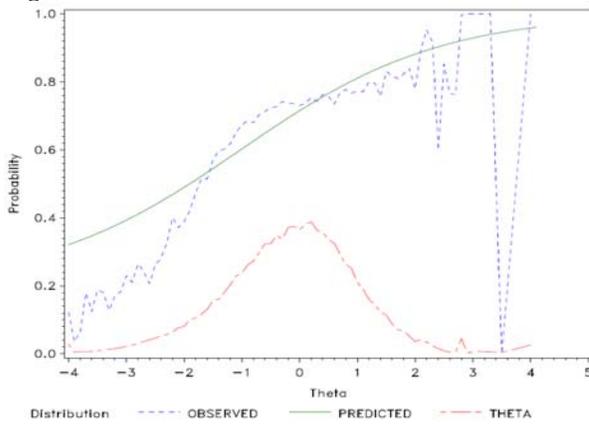


Figure 6. 8: Item characteristic curve for Grade 6 Mathematics, Session 3 Item 4

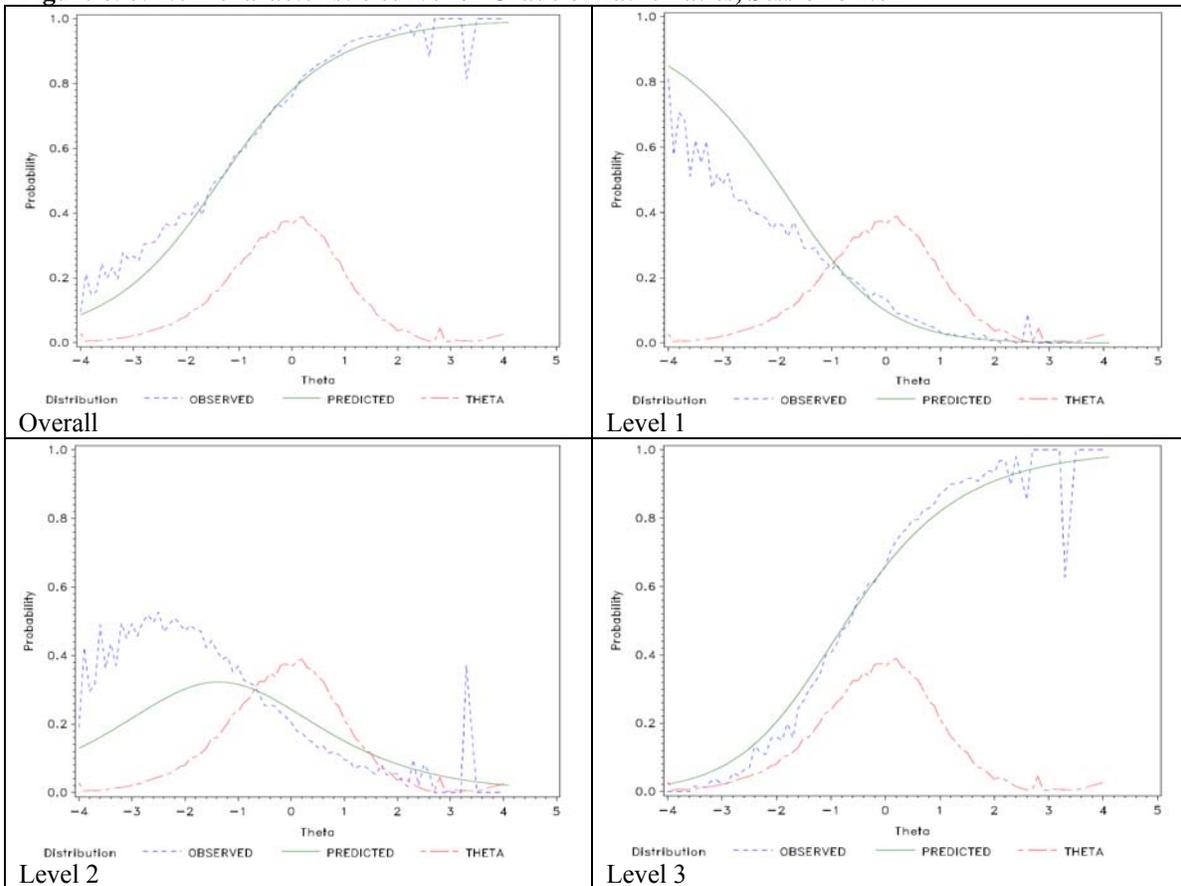


Figure 6. 9: Item characteristic curve for Grade 7 Mathematics, Session 1 Item 12

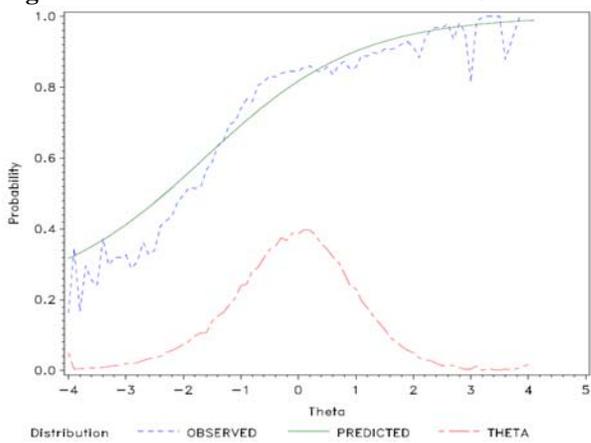


Figure 6. 10: Cross-Grade Articulation of Scale Scores at Selected Percentiles, Communication Arts MAP

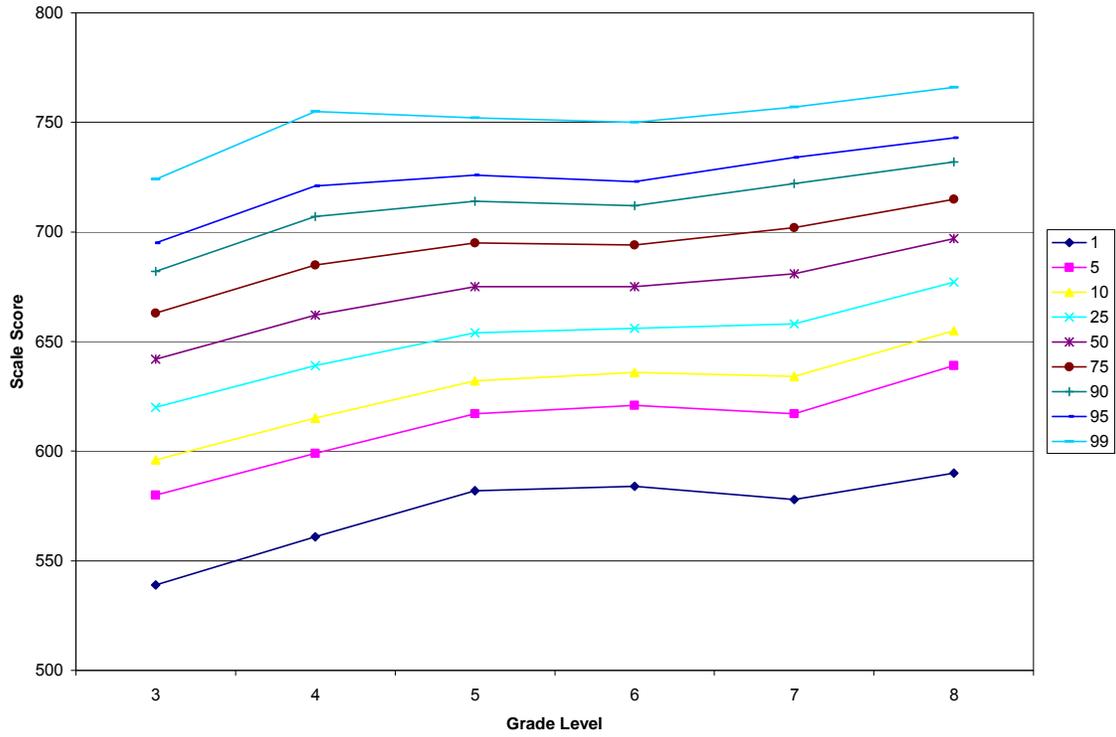


Figure 6. 11: Cross-Grade Articulation of Scale Scores at Selected Percentiles, Mathematics MAP

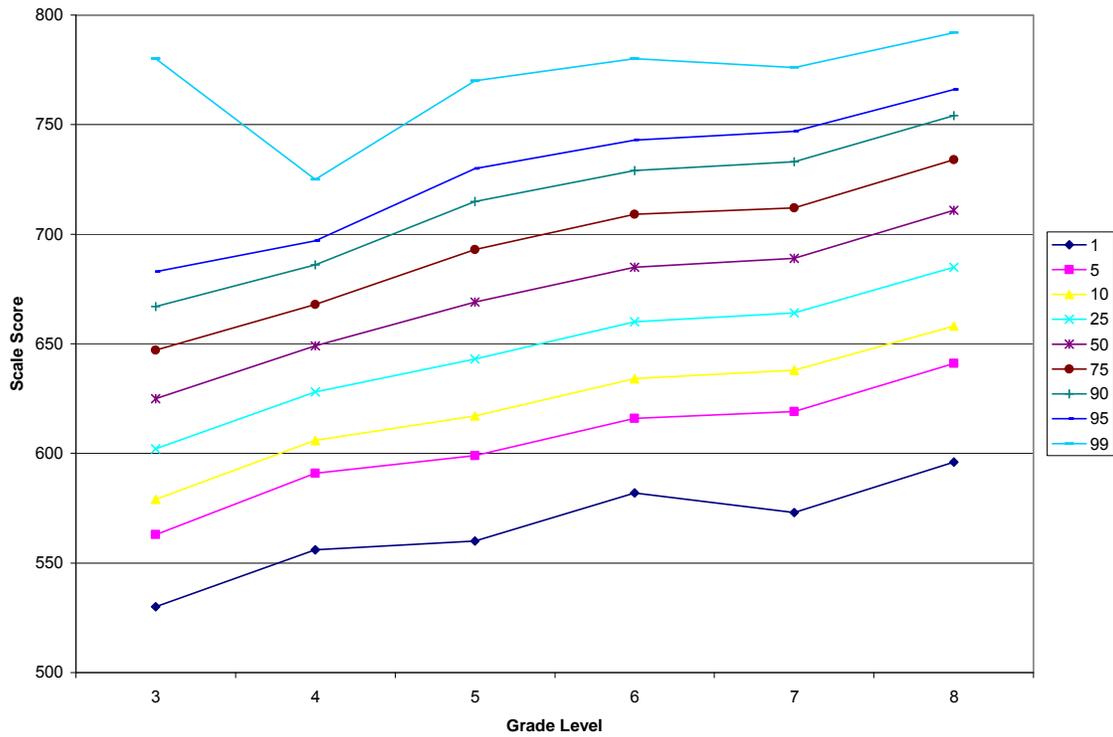


Figure 6. 12: Cross-Grade Articulation of Scale Scores at Selected Percentiles, Science MAP

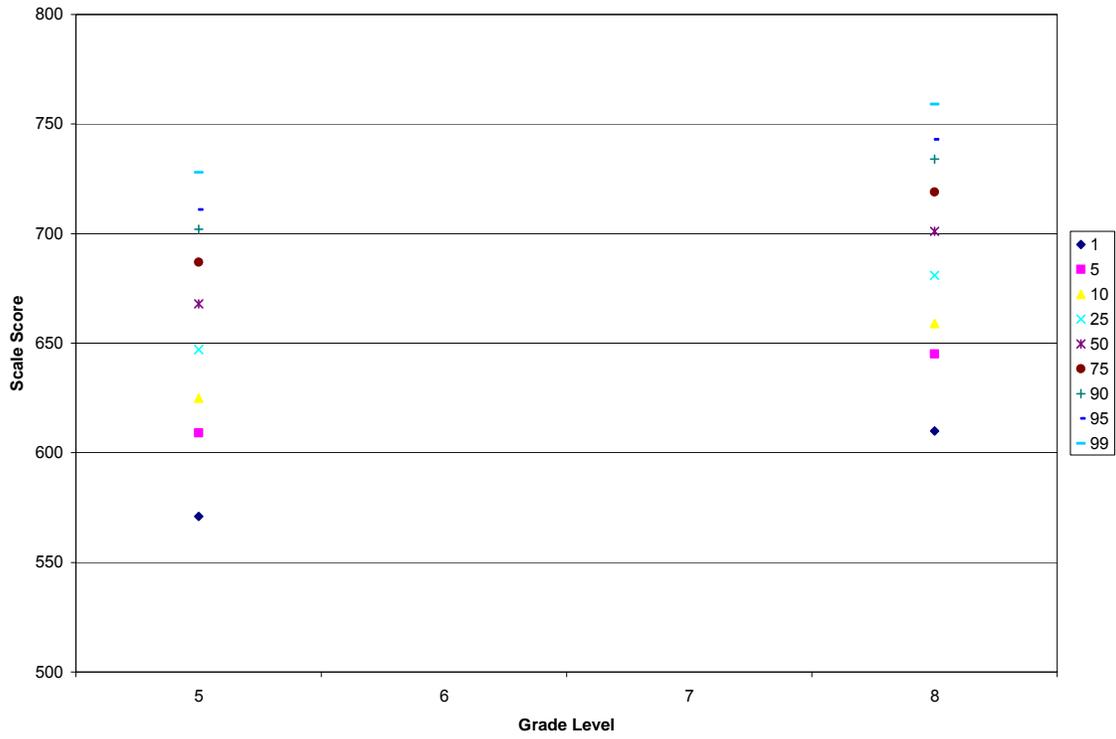


Figure 6. 13: Communication Arts Test Characteristic Curves by grade, 2010

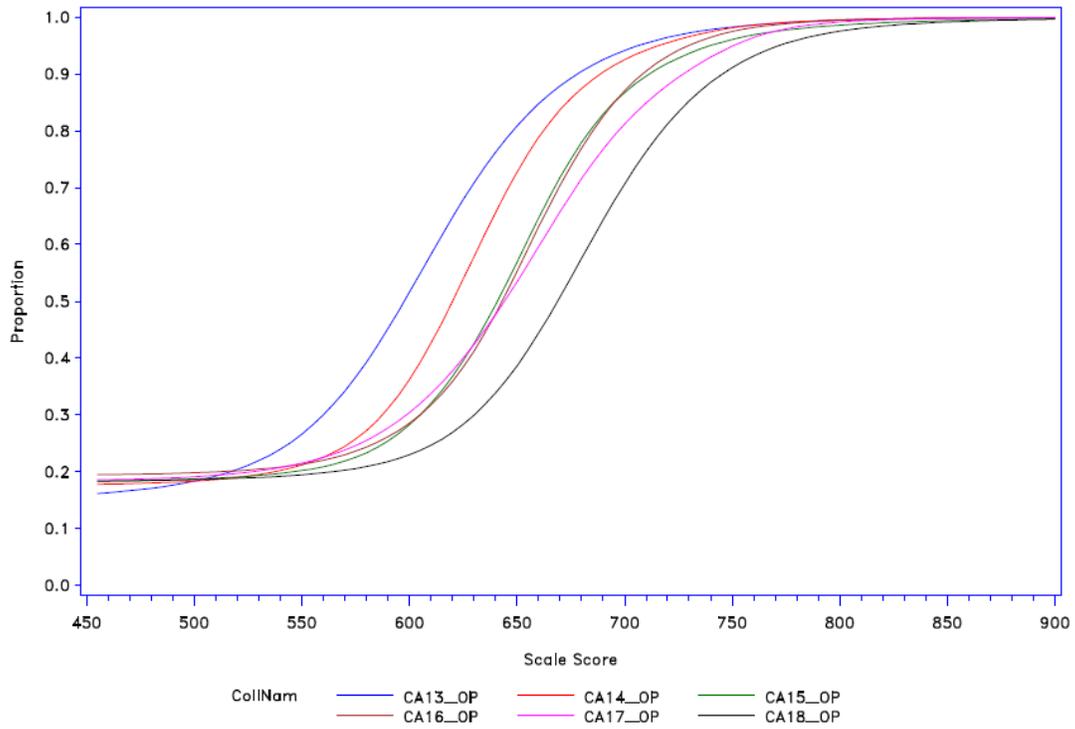


Figure 6. 14: Mathematics Test Characteristic Curves by grade, 2010

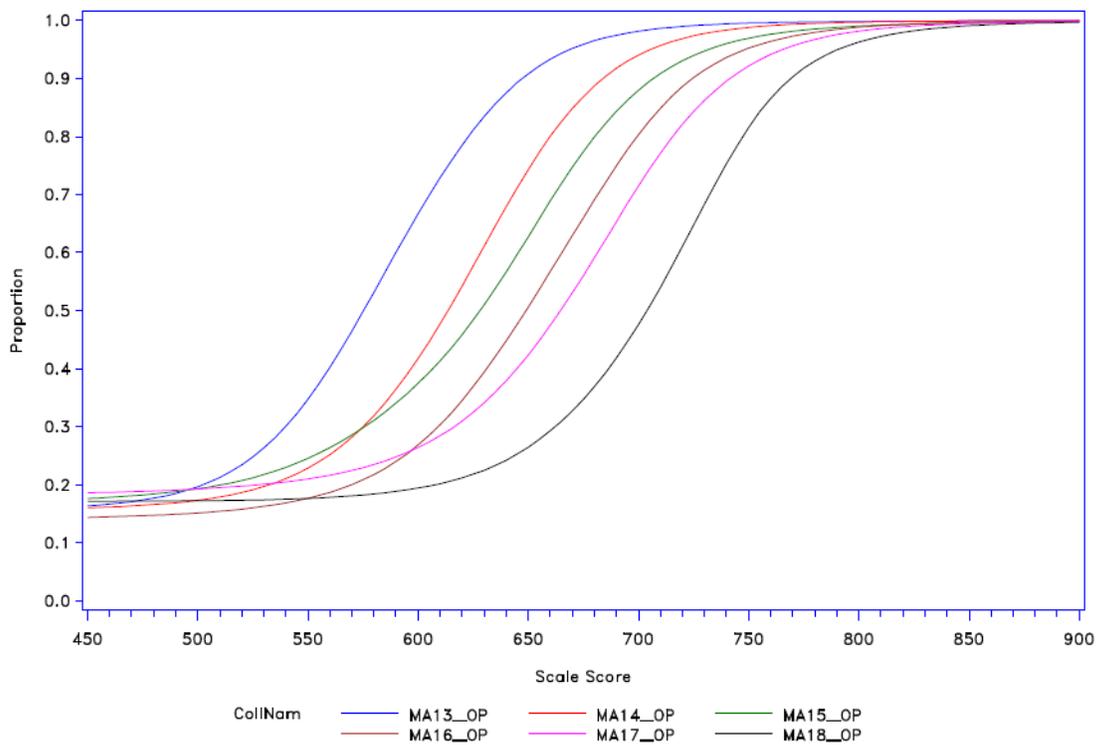
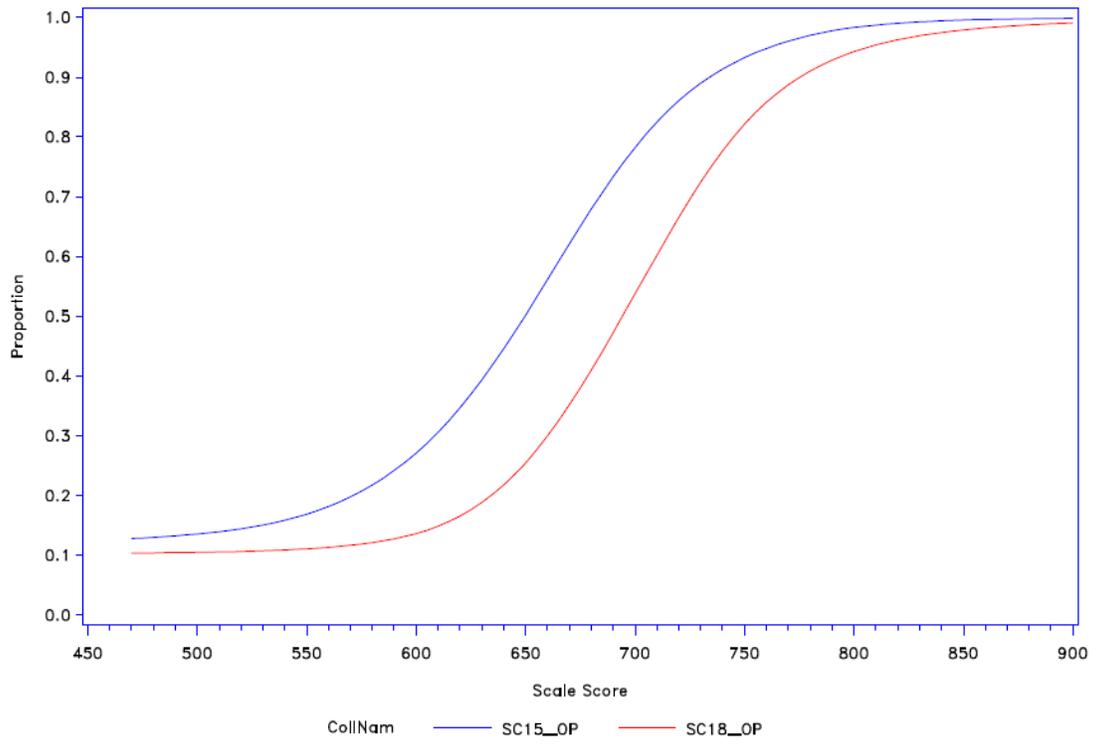


Figure 6. 15: Science Test Characteristic Curves by grade, 2010



CHAPTER 7: TEST RESULTS

This chapter of the Technical Report contains information on the results of the spring 2010 administration of the MAP. The scale score results are presented here. Performance level information is also provided. Presenting the results by performance level translates the quantitative scale provided through scale scores into a qualitative description of student performance: *Below Basic*, *Basic*, *Proficient*, and *Advanced*.

While the scale score provides an essential quantitative reference to student performance, the performance level information speaks directly to requirements of the NCLB Act, as well as plainly outlines the scores to parents, students, and educators. When combined, scale scores and performance levels provide a comprehensive set of tools to assess Missouri student performance by content and grade level.

This chapter also provides description of the score reports, data structure, and interpretive guide. The AERA, APA, & NCME (1999) Standards addressed in Chapter 7 are 4.1, 5.10, 6.2, and 13.19. Each Standard will be presented in the pertinent section of this chapter.

Results presented in the following section are based on census data. The results presented here may differ slightly from the official state summary report of all student populations due to ongoing resolution of test materials and student information. The results in the following tables are presented as evidence of reliability and validity of the scores from the MAP assessments and should not be used for state accountability purposes.

7.1 Student Participation

The following are subgroups reported during the administration of MAP (other demographic information is collected separately and merged into MAP data after CTB 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 percentage of students classified as reportable and the number of students classified as accountable. 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 2010 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 2010. 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 2010 for Communication Arts and Mathematics and from 2008 through 2010 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 2009 to 2010.

Table 7.14 shows the percentage of students in each achievement level in 2006 through 2010 on the Communication Arts test. The percentages at or above *Proficient* increased from 2009 to 2010.

Table 7.15 shows the percentage of students in each achievement level in 2006 through 2010 on the Mathematics test. As compared to 2009, increases in the percentage of students at or above *Proficient* were observed in all grades in 2010.

Table 7.16 shows the percentage of students in each achievement level in 2008 through 2010 on the Science test. In Grades 5 and 8, the percentage of students at or above *Proficient* increased from 2009 to 2010.

7.4 Reports

Score reports are the primary means of communicating test scores to relevant district personnel (i.e., test coordinators or superintendents), teachers, and parents. AERA, APA, & NCME (1999) Standard 5.10 states:

When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used.

Standard 4.1 is related in that it says:

Test documents should provide test users with clear explanations of the meaning and intended interpretations of derived score scales, as well as their limitations.

Interpretations related to the test scores are disseminated in two ways: (1) the individual score report, and (2) the *Guide to Interpreting Results* (CTB/McGraw-Hill, 2010).

In addition to providing interpretation, it is important that the information is understandable by the target audience. Standard 6.2 of the AERA, APA, & NCME (1999) Standards states:

Test documents should be complete, accurate, and clearly written so that the intended reader can readily understand the content.

The staffs at DESE and CTB strive to create documents that will be accessible to parents, teachers, and laypeople alike.

The individual student report is the primary means for sharing student test results with parents. As such, it should be a stand-alone document from which parents can glean relevant information so they understand their child's test score. In 2008, the individual MAP student reports were redesigned so that they were more parent-friendly. These changes include improved interpretations of the MAP scale score, *TerraNova* scale score, and Lexile score. Starting in 2010, Lexile scores are no longer reported. In addition, the state mean score is now provided, as are activities that parents may engage in to help their children improve their skills within the content area in accordance with the Missouri Curriculum Framework. The new score reports also simplify the way in which the scale score and performance level are presented and interpreted. Starting in 2008, parents no longer receive scores for content/knowledge standards or for process/performance standards.

The *Guide to Interpreting Results* is intended for use by school and district personnel so that they can interpret their score reports. It provides a context for the score reports in that it outlines the history and purpose of MAP. It also overviews the Missouri Show-Me

Standards and GLE Strands. It provides greater detail on the types of scores reported on the individual student report, and it provides all of the abbreviated achievement-level descriptors (ALDs), as well as the web location of the detailed ALDs. Finally, it outlines each piece of the individual student report and overviews the student label. The *Guide to Interpreting Results* is located on DESE’s website at:

<http://dese.mo.gov/divimprove/assess/map/>

7.4.1 Description of Each Type of Report

In this section, descriptions are provided for the following reports: Individual Student Report, Student Score Label, online Crystal Reports, and School/District Performance Summary Reports. Table 7.17 shows each report type and for whom the report is intended.

Individual Student Report

One copy of the Individual Student Report (ISR) is provided to schools to be sent home to the parents. On the left side of the page, results for a given content area are shown, including the student’s MAP scale score, the state mean MAP scale score, the National Percentile score from the *TerraNova* section of MAP, and a brief definition of the National Percentile.

In the middle of the page, the student’s scale score is shown again along with the achievement level associated with that scale score. This is followed by a brief explanation of what the achievement level means. When a student does not receive a scale score, he or she will receive either “Level Not Determined” (LND) or “Invalidated” in place of the MAP scale score. No achievement level is assigned for the LND students. Invalidated students are assigned to the LOSS and to the *Below Basic* achievement level. A brief explanation accompanies the meaning of LND or Invalidated.

On the right side of the page are recommended activities based on the child’s achievement level. These are generic activities that are targeted to all students within an achievement level, not specific activities targeted at the individual student. A sample report is provided in Appendix C, Figure C.1.

Student Score Label

The Student Score Label is designed so that each student’s test results can be placed in the student’s permanent record. A label is provided for every student who participated in the spring 2010 administration of the MAP. Each label has a self-adhesive backing so that it can be peeled from the sheet and placed in the student’s cumulative school record. The label presents a snapshot of the student’s results on the MAP. Separate labels are generated for each grade and content area; thus, a student will have multiple labels for each of the content areas administered within a grade. The label lists the student’s scale score and National Percentile for each content area.

CTB/McGraw-Hill provided multiple labels for each student. The labels are provided in print only. A sample report is provided in Appendix C, Figure C.2.

Online Crystal Reports

Schools and districts are able to access summary-level reports through the online Crystal Reports tool. This tool allows district and school administrators to create on-the-fly reports containing information relevant to their data needs. There are several reporting options available through the Crystal Reports tool, including administrative reports, AYP reports, achievement-level reports, content standard reports, and item analysis reports. Table 7.18 lists each of the major report headings and the sub-reports found under each reporting type.

For each sub-report, a user selects various filters such as year, grade level/content area, and level of reporting (state, district, school) in order to create the desired report. For the Content Standard Reports, the user may also disaggregate results by various subgroups (e.g. race, disability).

A detailed discussion of all reports is beyond the scope of this document. Only those reports that are first-level analyses of MAP data will be discussed. The AYP reports also will not be discussed nor will some of the Administrative Reports, including the Level Not Determined and Map Alternate reports. Examples of all reports discussed are provided in Appendix C.

The Crystal Reports tool is accessed through DESE's website. Each school and/or district is assigned a user name and password so that it can access the site.

Administrative Reports

These reports provide student-level test data. Based on only the MAP test results, four reports are generated: MAP Scale Score Summary, MAP Student Demographic, Student Achievement Level, and Participation Invalidation.

MAP Scale Score Summary: This report lists each student in the school or district along with his/her MOSIS ID, testing year, content area, grade level, MAP scale score, achievement level, and *TerraNova* National Percentile. An example is included in Appendix C, Figure C.3.

MAP Student Demographic: This report lists each student in the school or district along with their date of birth (DOB), content area, CTB number, MOSIS ID, district ID, and relevant demographic information, including the student's race; the student's disability diagnosis; if the student has been in the district for less than a year; if the student has been in the building for less than a year; if the student is Limited English Proficient (LEP); if the student qualifies for free and reduced lunch (SES); if the student has an individualized education plan (IEP); if the student is an English-language learner (ELL)/LEP who has been in the school for less than one year and in the country for less than three years; if the student is an LEP/ELL Title 3, the number of months the LEP/ELL student has been in the U.S.; if the student took MAP-A; and if the student is Title I. An example is included in Appendix C, Figure C.4.

Student Achievement Level: This report lists all of the students in a school or district along with the year of testing, content area, grade level, achievement level, and MOSIS ID. An example is included in Appendix C, Figure C.5.

Participation Invalidation Report: This report lists all of the students in a school or district who were invalidated on the test. It gives each student's full name, MOSIS ID, and marks the reason for the invalidation. It also tells the student's assigned scale score and achievement level. An example is included in Appendix C, Figure C.6.

Achievement Level—4 Levels

These reports contain summary information on school or district performance in terms of the four MAP achievement levels.

Achievement Level 4 Report: This report summarizes the number and percentage of students in each achievement level. This report is comprised of 19 columns: Total; content area; grade; year; number of accountable (ACC) students; number of reportable (REP) students; number and percentage of students whose achievement level was not determined (LND); number and percentage of students classified in the *Below Basic* (BB) achievement level; number and percentage of students classified in the *Basic* (B) achievement level; number and percentage of students classified in the *Proficient* (P) achievement level; number and percentage of students classified in the *Advanced* (A) achievement level; MAP index score; mean MAP scale score; and the median *TerraNova* national percentile. The first column, Total, shows if aggregated or disaggregated information is being shown. A key to the abbreviations is found in the bottom left corner, as is the computation details for the MAP Index score. An example is included in Appendix C, Figure C.7.

Content Standard

The content standard reports summarize information about the Content Standards (CSs).

Content Standards Report: This report has 14 columns: content area; grade level; category/type; year; percentage of points earned on content standard 1 (CS-1); points possible (PP) on CS-1; percentage of points earned on CS-2; PP on CS-2; percentage of points earned on CS-3; PP on CS-3; percentage of points earned on CS-4; PP on CS-4; percentage of points earned on CS-5; and PP on CS-5. The category/type column indicates if the data is aggregated or disaggregated. An example is included in Appendix C, Figure C.8.

Content Standards Detail: This report shows the percentage of points each student achieved on each content standard within a particular content area. An example is included in Appendix C, Figure C.9.

Item Analysis Expanded

This set of reports provides detailed item-level results for the school or district aggregated either by the content standard or process standard.

Content Standard IBD EX: The Content Standard Item Benchmark Descriptions (IBD) Extended (EX) report contains item-level detail aggregated by content standard. The report is comprised of 11 columns: school code (SC); grade level (GR); standard number and description (desc.); code for the grade-level expectation (GLE); description of the GLE; depth of knowledge (DOK) of the item; session/item number where the item was in the operational test; question type (QT); points possible for the item; average points (avg pts) earned by students in the district on that item; and percentage of points earned by the students in the district on that item. An example is included in Appendix C, Figure C.10.

Goal Process Standard IBD EX: The Goal Process Standard Item Benchmark Descriptions (IBD) Extended (EX) report contains item-level detail aggregated by the goal process standard. The report is comprised of 12 columns: school code (SC); grade level (GR); goal; standard description (desc.); code for the grade-level expectation (GLE); description of the GLE; depth of knowledge (DOK) of the item; session/item number where the item was in the operational test; question type (QT); points possible for the item; the average points (avg pts) earned by students in the district on that item; and percentage of points earned by the students in the district on that item. An example is included in Appendix C, Figure C.11.

School/District Summary Reports

CTB provides DESE with school and district summary reports for each school and district in the state. These reports are intended for the sole use of DESE and are not distributed to schools and districts. These reports provide performance information for all students within a school or district who took the MAP.

The school or district is listed in the left-most column along with the purpose of the report. The main section of the summary report consists of a table that divides students from the school or district into achievement levels. The *Reportable* column shows the number of students with valid MAP scale scores. The *Accountable* column should equal the grade-level enrollment at the time the MAP was administered.

Within both the *Reportable* and *Accountable* columns, students are categorized as *Below Basic*, *Basic*, *Proficient*, or *Advanced*. The number and percentage of students falling into each achievement level is reported. A short description of the knowledge, skills and abilities associated with each achievement level is also reported. Students who are not assigned to an achievement level will be classified as *Level Not Determined*. A short descriptor is also associated with this categorization.

Below the table, the norm-referenced summary statistics are reported for each school or district. The norm-referenced information includes the National Percentile (NP)

associated with the Mean Normal Curve Equivalent, the median NP, and the number of students with *TerraNova* scores.

On the back of these reports, the terms *Reportable* and *Accountable* are defined. A sample of the School/District Summary Report is provided in Appendix C, Figure C.12.

7.5 Data Structures

A data file referred to as a General Research File (GRF) was provided to DESE by CTB/McGraw-Hill. It contains one record for every test book submitted; each record contains demographic information for each student, as well as item responses, raw score, content and process standard raw scores, and scale score data for each content area. The layout for a state-level GRF is included in Appendix C.

7.6 Interpreting Test Results

Individual Student Reports and Student Labels

The *Guide to Interpreting Results* was written for Missouri teachers and administrators who receive score reports from the 2010 administration of the MAP. The *Guide to Interpreting Results* was developed collaboratively by CTB/McGraw-Hill and DESE staff. DESE staff has opportunities to review, provide feedback, and give final approval.

This guide has six sections. The first section presents an overview of key terms and test-related concepts. The second section presents the Show-Me Content Standards/GLE Strands. The third section presents the Show-Me Performance Standards. The fourth section discusses assessment terms and the types of scores that will be presented on the score reports. The fifth section presents the achievement-level descriptors for all grade levels/content areas. Finally, the sixth section presents sample score reports.

The 2010 edition is available on the DESE website at:

<http://dese.mo.gov/divimprove/assess/map/mapgenresources.html>

Crystal Reports

Training for the Crystal Report tool is provided through online help tools.

7.7 Summary

In summary, the overall purpose of reporting test results is to communicate various aggregations of student performance to stakeholders. These results are presented in the context of score reports that aid the user in understanding the meaning of the test scores. The reports and ancillary information developed by CTB/McGraw-Hill address multiple best practices of the testing industry but in particular are related to the following Standards (AERA, APA, & NCME, 1999):

- Standard 4.1—Test documents should provide test users with clear explanations of the meaning and intended interpretations of derived score scales, as well as their limitations.
- Standard 5.10—When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used.
- Standard 6.2—Test documents should be complete, accurate, and clearly written so that the intended reader can readily understand the content.
- Standard 13.19—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.

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	66947	99.7%	66947	99.8%		
4	67510	99.7%	67510	99.8%		
5	66730	99.7%	66730	99.8%	66730	99.7%
6	67476	99.7%	67476	99.8%		
7	66279	99.6%	66279	99.7%		
8	66463	99.5%	66463	99.6%	66463	99.5%

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	33912	99.7%	33912	99.8%		
4	34654	99.7%	34654	99.8%		
5	34175	99.7%	34175	99.7%	34175	99.7%
6	34624	99.6%	34624	99.7%		
7	33779	99.6%	33779	99.6%		
8	33842	99.4%	33842	99.5%	33842	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	32956	99.7%	32956	99.8%		
4	32764	99.7%	32764	99.8%		
5	32469	99.7%	32469	99.8%	32469	99.8%
6	32772	99.8%	32772	99.8%		
7	32419	99.7%	32419	99.7%		
8	32520	99.6%	32520	99.7%	32520	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	50108	99.8%	50108	99.8%		
4	50566	99.8%	50566	99.9%		
5	50237	99.8%	50237	99.8%	50237	99.8%
6	50866	99.8%	50866	99.8%		
7	50434	99.8%	50434	99.7%		
8	51080	99.6%	51080	99.6%	51080	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	11873	99.7%	11873	99.7%		
4	12160	99.7%	12160	99.7%		
5	11958	99.5%	11958	99.6%	11958	99.5%
6	12120	99.6%	12120	99.6%		
7	11627	99.5%	11627	99.5%		
8	11361	99.4%	11361	99.2%	11361	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	3125	99.1%	3125	99.8%		
4	3017	98.5%	3017	99.8%		
5	2805	98.9%	2805	99.5%	2805	99.5%
6	2771	99.1%	2771	99.8%		
7	2539	99.1%	2539	99.7%		
8	2312	98.4%	2312	99.5%	2312	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	1462	96.6%	1462	99.7%		
4	1398	97.4%	1398	99.6%		
5	1324	96.9%	1324	99.8%	1324	99.8%
6	1293	97.4%	1293	99.8%		
7	1266	97.0%	1266	99.8%		
8	1256	97.4%	1256	99.9%	1256	99.8%

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	313	98.7%	313	99.0%		
4	283	99.3%	283	99.6%		
5	316	99.4%	316	99.7%	316	99.4%
6	339	99.7%	339	99.7%		
7	329	99.1%	329	98.8%		
8	346	99.4%	346	99.7%	346	99.7%

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	6336	99.9%	6570	99.9%		
4	7084	99.8%	7343	99.9%		
5	7102	99.8%	7355	99.9%	7027	99.9%
6	7235	99.8%	7492	99.9%		
7	6830	99.7%	7050	99.7%		
8	6517	99.7%	6817	99.7%	6536	99.5%

Table 7. 10: Summary Statistics for Communication Arts

Grade	N	Mean	Std. Dev.	Scale Scores by Percentiles				
				10	25	50	75	90
3	66,751	640.27	36.63	596	620	642	663	682
4	67,301	661.34	38.95	615	639	662	685	707
5	66,500	673.65	35.33	632	654	675	695	714
6	67,260	674.18	33.12	636	656	675	694	712
7	66,034	678.85	36.25	634	658	681	702	722
8	66,139	694.28	34.01	655	677	697	715	732

Table 7. 11: Summary Statistics for Mathematics

Grade	N	Mean	Std. Dev.	Scale Scores by Percentiles				
				10	25	50	75	90
3	66,814	624.89	39.28	579	602	625	647	667
4	67,394	647.59	34.01	606	628	649	668	686
5	66,580	667.70	41.74	617	643	669	693	715
6	67,315	683.36	39.48	634	660	685	709	729
7	66,052	686.51	40.28	638	664	689	712	733
8	66,166	707.98	40.04	658	685	711	734	754

Table 7. 12: Summary Statistics for Science

Grade	N	Mean	Std. Dev.	Scale Scores by Percentiles				
				10	25	50	75	90
5	66,558	664.76	32.48	625	647	668	687	702
8	66,101	698.28	31.07	659	681	701	719	734

Table 7. 13: Comparison of State-Level Means, 2006 through 2010 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			
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			
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
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			
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			
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

Table 7. 14: Comparison of Percentage of Students in each Achievement Level, Communication Arts 2006 through 2010 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
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
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
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
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
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

Table 7. 15: Comparison of Percentage of Students in each Achievement Level, Mathematics 2006 through 2010 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
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
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
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
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
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

Table 7. 16: Comparison of Percentage of Students in each Achievement Level, Science 2008 through 2010 Census Data

Grade	Year	N	No Level	Below Basic	Basic	Proficient	Advanced	Prof & Adv
5	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
8	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

Table 7. 17: Summary of Score Reports for Spring 2010

Score Report		Paper Report		Electronic Report		
		Parent	Teacher	Principal	System	DESE
Student Score Labels			X			
Individual Student Report		X				
Performance Summary Report	School Performance Summary Report					X
	District Performance Summary Report					X
	Crystal Reports			X	X	

Table 7. 18: Types of Reports Available to Districts through Crystal Reports

Crystal Report	Sub Reports
Administrative Report	Level Not Determined MAP Alternate MAP Scale Score Summary MAP Student Demographic Participation Invalidation Student Achievement Level
AYP	AYP AYP Additional Indicator AYP Growth Target Met AYP Growth Trajectory AYP Summary
Achievement Level-4 Levels	Achievement Level 4 Report
Content Standards	Content Standards Content Standards Detail
Item Analysis Expanded	Content Standard IBD EX Goal Process Standard IBD EX

CHAPTER 8: ACHIEVEMENT-LEVEL SETTING

A Bookmark standard setting was held in 2005 to establish cut scores for the Communication Arts and Mathematics MAP tests. Another Bookmark standard setting was held in 2008 to establish cut scores for the Science MAP. In this chapter, we briefly describe the MAP achievement-level setting, and we present the cut scores established and the achievement-level descriptors derived from the achievement-level setting.

A detailed discussion of the Communication Arts and Mathematics achievement-level setting may be found in the *Missouri Assessment Program Final Bookmark Standard Setting Technical Report* (2005). A detailed discussion of the Science achievement-level setting may be found in the *Missouri Assessment Program Bookmark Standard Setting Technical Report 2008 for Missouri Achievement-Level Setting Grades 5, 8, and 11 Science* (2008). These Technical Reports address AERA, APA, & NCME (1999) Standard 4.19:

When proposed score interpretation involves one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented.

We briefly overview the rationale and procedures used for MAP standard setting in the following section.

In terms of the validity of the MAP scores, it is essential to understand that descriptors and cut scores are established in a collaborative, participatory process, largely driven by the input of Missouri teachers and educators. The descriptors clearly establish, in plain language, the proper frame of reference for understanding how to interpret test scores, cut scores in particular.

8.1 Legislation Affecting MAP Standard Setting

A modified Bookmark Standard Setting Procedure (BSSP) was used to establish cut scores for the Communication Arts and Mathematics MAP tests for Grades 3 through 8 and high school and Science for Grades 5, 8, and 11. A modification of the Bookmark was used to meet the requirements of Senate Bill 1080, which requires that cut scores be established for the MAP tests that are like the cut scores established for the National Assessment of Educational Progress (NAEP).

Senate Bill 1080 was interpreted such that the *Proficient* achievement level met, but did not exceed, the NAEP performance standards. In other words, the percentage of students who attain *Proficient* on the MAP should be similar to or slightly higher than the percentage attaining *Proficient* on NAEP. The percentage of students in the other three achievement levels would be allowed to vary between NAEP and the MAP.

For the purposes of the MAP standard setting, participants were allowed to recommend *Proficient* cut scores within a pre-specified range. This range was based on the percentage of students who could be classified as either *Proficient* or *Advanced*. For

Communication Arts and Mathematics, no fewer than 26% and no more than 44% of students could be classified as *Proficient* or *Advanced*. For Science, no fewer than 27% and no more than 48% of students could be classified as *Proficient* or *Advanced*.

The pre-specified range was determined using the results from NAEP and MAP. For all three content areas, the high end of the range (in terms of scale score points) was based on NAEP results. This was the lowest percentage of students classified as *Proficient* or *Advanced* on the NAEP test for Grades 4 and 8 Reading, Mathematics, and Science using both national and state data.

The low end of the range (in terms of scale score points) was based on the 2005 MAP results for the Communication Arts and Mathematics standard setting and on the 2007 MAP results for Science. This was the highest percentage of students classified as *Proficient* or *Advanced* on the previous years' tests.

8.2 Bookmark Standard Setting Procedure

A modified BSSP was used to establish cut scores on the Communication Arts, Mathematics, and Science MAP. At both workshops, the BSSP involved three rounds of discussion and voting. AERA, APA, & NCME (1999) Standard 4.21 says:

When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way.

The Technical Reports associated with each standard setting give detailed reports of the standard setting design and procedure. Here, we discuss the major activities of the three rounds.

Round 1: In this round, panelists discussed target students (the students for whom they were placing cut scores), took the test, studied and discussed the test items in order of difficulty, and made initial recommendations of cut scores.

Round 2: In this round, panelists were shown their Round 1 recommendations and the percentage of students in each achievement level as a result of their Round 1 recommendations. They discussed their Round 1 recommendations for cut scores and made another recommendation based on their Round 2 discussions.

Round 3: In this round, panelists were shown their Round 2 recommendations and the percentage of students in each achievement level as a result of their Round 2 recommendations. They discussed their Round 2 recommendations for cut scores and made another recommendation based on their Round 3 discussions.

Following Round 3, panelists wrote draft achievement-level descriptors which were later edited by CTB and DESE staff.

The Missouri State Board of Education approved the cut scores as recommended by the standard setting panelists.

8.3 Cut Scores

In this section, we present the cut scores for each grade level/content area of MAP. Tables 8.1 through 8.3 show the cut scores for Communication Arts, Mathematics, and Science, respectively. Please note that we only present the cut scores for Grades 3 through 8. The high school MAP tests are no longer part of the assessment system.

8.4 Achievement-Level Descriptors

In Appendix D of this report, we present the short achievement-level descriptors that were drafted during the standard setting and finalized between CTB and DESE staff after the standard setting. We only present the short achievement-level descriptors for those grades that are currently part of the MAP.

8.5 Summary

This chapter presented a brief overview of the standard setting process used for the grade-level MAP tests, as well as the rationale behind the standard setting. The standard settings are addressed in more detail in the relevant Technical Reports. The standard settings undertaken by CTB/McGraw-Hill address the following Standards (AERA, APA, & NCME, 1999):

- Standard 4.19—When proposed score interpretation involves one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented.
- Standard 4.21—When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way.

Table 8. 1: Communication Arts Cut Scores

Grade	Cut Scores		
	Basic	Proficient	Advanced
3	592	648	673
4	612	662	691
5	625	675	702
6	631	676	704
7	634	680	712
8	639	696	723

Table 8. 2: Mathematics Cut Scores

Grade	Cut Scores		
	Basic	Proficient	Advanced
3	568	628	667
4	596	651	688
5	605	668	706
6	628	681	721
7	640	685	724
8	670	710	741

Table 8. 3: Science Cut Scores

Grade	Cut Scores		
	Basic	Proficient	Advanced
5	626	669	692
8	671	703	735

CHAPTER 9: EVIDENCE OF CONSTRUCT-RELATED VALIDITY

Evidence for construct-related validity—the meaning of test scores and the inferences they support—is the central concept underlying the MAP validation process. In this section, CTB presents evidence of construct-related validity through studies of test reliability, convergent validity, and divergent validity. All analyses in this section are based on census data.

Chapter 9 of this report demonstrates adherence to AERA, APA, & NCME (1999) Standards 1.11, 1.18, 2.1, 2.2, 2.4, 2.14, and 2.15. Each standard will be discussed in the pertinent section of this chapter.

9.1 Minimization of Construct-Irrelevant Variance and Construct Under-Representation

Minimization of construct-irrelevant variance and construct under-representation is addressed in the following steps of the test development process: 1) specification, 2) item writing, 3) review, 4) field testing, 5) test construction, and 6) calibration (see Chapter 3 for more information on steps 1 through 5 and Chapter 6 for more information on calibration).

Construct-irrelevant variance refers to error variance that is caused by factors unrelated to the constructs measured by the test. For example, when tests are not administered under standardized conditions (e.g., one administration may be timed, but another administration may be untimed), differences in student performance related to different administration conditions may result. Careful specification of content and review of the items representing that content are first steps in minimizing construct-irrelevant variance. Then, empirical evidence, especially item-level data, is used to infer construct irrelevance.

Construct under-representation occurs when the content of the assessment does not reflect the full range of content that the assessment is expected to cover. MAP is designed to represent the Show-Me Standards/GLE strands. Specification and review, in which test blueprints are developed and reviewed, are primary steps in the development process designed to ensure that content is appropriately represented.

9.2 Reliability

Reliability refers to the consistency of the students' test scores on parallel forms of a test. A reliable test is one that produces scores that are expected to be relatively stable if the test is administered repeatedly under similar conditions. Often, however, it is impractical to administer multiple forms of the test, and reliability is estimated on a single administration of the test. This type of reliability, known as internal consistency, provides an estimate of how consistently examinees perform across items within a test during a single test administration (Crocker & Algina, 1986). Reliability is a necessary, but not sufficient, condition of validity.

The AERA, APA, & NCME (1999) Standards indicate:

. . . reliability evidence may be reported in terms of variances or standard deviations of measurement errors, in terms of one or more coefficients, or in terms of IRT-based test information functions (27).

In accordance with the AERA, APA, & NCME (1999) Standards and developing and maintaining tests of the highest quality, CTB has calculated the reliability of each MAP test in a variety of ways: reliability of raw scores, overall standard error of measurement, IRT-based conditional standard error of measurement, and decision consistency of achievement-level classifications.

There are several specific AERA, APA, & NCME (1999) Standards that this chapter addresses. These include Standards 2.1, 2.2, and 2.4, each articulated below.

Standard 2.1 *For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported.*

The total score reliabilities are discussed in Section 9.2.1. of this chapter. The subscore reliabilities and SEMs are presented in Section 9.4.3. The SEM of the total score is discussed in Section 9.2.2.

Standard 2.2: The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original scale units and in units of each derived score recommended for use in test interpretation.

The overall SEM is discussed in Section 9.2.2 and is reported in scale score units. The conditional SEM is discussed in Section 9.2.3.

Standard 2.4: Each method of quantifying the precision or consistency of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select examinees for reliability analyses and descriptive statistics on these samples should be reported.

Section 9.2 discusses different ways of measuring test reliability, including reliability of raw scores, overall SEM, IRT-based conditional SEM, and decision consistency of achievement-level classifications. The sample on which these statistics are computed is discussed in Section 6.1 of Chapter 6.

9.2.1 Test Reliability

The reliability of raw scores on the MAP tests was evaluated using Cronbach's (1951) coefficient alpha, which is a lower-bound estimate of test reliability. The reliability coefficient is a ratio of the variance of true test scores to those of the observed scores, with the values ranging from 0 to 1. The closer the value of the reliability coefficient is

to 1, the more consistent the scores, where 1 refers to a perfectly consistent test. As a rule of thumb, reliability coefficients that are equal to or greater than 0.8 are considered acceptable for tests of moderate lengths.

Cronbach's coefficient alpha was computed using the formula

$$\alpha = \frac{n}{n-1} \left[1 - \frac{\sum_{i=1}^n \sigma_i^2}{\sigma_x^2} \right], \quad (9.1)$$

where n is the number of items on the test, σ_i^2 is the variance of item i and σ_x^2 is the variance of the total test score.

Total test reliability measures, such as Cronbach's coefficient alpha and SEM, consider the consistency (reliability) of performance over all test items in a given form, the results of which imply how well the items measure the content domain and could continue to do so over repeated administrations. The number of items in the test influences these statistics; a longer test can be expected to be more reliable than a shorter test.

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 reliability of reported test scores can be characterized by the standard errors associated with the scores. The SEM may be used to determine the range within which a student's true score is likely to fall. An observed score should be regarded not as a student's true score, but as an estimate of a student's true score. It is expected that 68% of the time a student's score obtained from a single test administration would fall within one SEM of the student's true score and that 95% of the time the obtained score would fall within approximately two standard errors of the true score. The SEM is an index of the random variability in test scores and is defined as follows:

$$SEM = SD\sqrt{1 - R_{xx'}}, \quad (9.2)$$

where SD represents standard deviation of the raw score distribution and $R_{xx'}$, is estimated by $\hat{\alpha}$ as expressed in Equation 9.1.

The overall 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

In contrast to SEM, the conditional standard errors of measurement (CSEMs) express the degree of measurement error in scale score units and are conditioned on the ability of the student. We report the CSEM in support of AERA, APA, & NCME (1999) Standard 2.14, which states:

Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of the cut scores.

In further compliance with Standard 2.14, the CSEM of each cut score is reported in Table 9.5.

The CSEMs are defined as the reciprocal of the square root of the test information function and can be estimated across all points of the ability continuum (Hambleton & Swaminathan, 1985):

$$\text{CSEM}(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}}, \quad (9.3)$$

where $I(\theta_i)$ is the test information function, as a sum of item information function 2, obtained as

$$I(\theta_i) = \sum_j \frac{p'_{ij}(\theta_i)^2}{p_{ij}(\theta_i)q_{ij}(\theta_i)}, \quad (9.4)$$

where $p'_{ij}(\theta_i)$ is the derivative of $p_{ij}(\theta_i)$ and $q_{ij}(\theta_i) = 1 - p_{ij}(\theta_i)$.

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 17 scale score points.

Figures 9.1 through 9.3 display the CSEM curves and cut scores for each grade level/content area. The estimates of measurement error tend to be higher at the low and high ends of the scale score range. The measurement error increases when there are few observations at a particular ability level. Generally, there are few students with extreme scores, and these score levels cannot be estimated as accurately as levels toward the middle of the ability range. Figures 9.1 through 9.3 demonstrate that the tests are designed so that measurement error is minimized in the middle of the scale range where the majority of students are located.

9.2.4 Classification Accuracy and Consistency

The *Standards* (AERA, APA, & NCME, 1999) also make reference to an additional measurement concern that bears on evidence for validity:

Some authorities have proposed that a semantic distinction be made between “reliability of scores” and “degree of agreement in classification.” The former term would be reserved for analysis of score variation under repeated measurement. The term *classification consistency . . .*, rather than reliability, would be used in discussions of consistency of classification. Adoption of such usage would make it clear that the importance of an error of any given size depends on the proximity of the examinee’s score to the cut score.

Classification Consistency: Classification consistency (also known as decision consistency) is defined as the extent to which the classifications of students agree on the basis of two independent administrations of the test or one administration of two parallel test forms. It is difficult, however, to obtain data from repeated administrations of the same form because of cost, time, and students’ recall of the first administration. Also, it is difficult to construct two parallel forms. A common practice, therefore, is to estimate decision consistency from one administration of a test. These analyses directly address AERA, APA, & NCME (1999) Standard 2.15, which states:

When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure, using the same form or alternate forms of the instrument.

Classification Accuracy: Classification accuracy is defined as the extent to which the actual classifications of test takers agree with classifications that would be made on the basis of their true scores (Livingston & Lewis, 1995). It is common to estimate classification accuracy by utilizing a psychometric model to find true scores corresponding to observed scores.

In other words, classification *consistency* refers to the agreement between two observed scores, while classification *accuracy* refers to the agreement between the observed score and the true score. A straightforward approach to classification consistency estimation can be expressed in terms of a contingency table representing the probability of a particular classification outcome under specific scenarios. For example, the following table is a contingency table of $(H+1) \times (H+1)$, where H is the number of cut scores, such that two cut scores yield a 3×3 contingency table.

Example of Contingency Table with Two Cut Scores

	Level 1	Level 2	Level 3	Sum
Level 1	P_{11}	P_{21}	P_{31}	$P_{.1}$
Level 2	P_{12}	P_{22}	P_{32}	$P_{.2}$
Level 3	P_{13}	P_{23}	P_{33}	$P_{.3}$
Sum	$P_{1.}$	$P_{2.}$	$P_{3.}$	1.0

CTB used a method suggested by Kolen and Kim (2005) for estimating consistency and accuracy that involves the generation of item responses using item parameters based on the IRT model (see also Kim, Choi, Um, & Kim, 2006; Kim, Barton, & Kim, 2007). Two sets of item responses are generated using a set of item parameters and an examinee’s ability distribution from a single test administration. These two sets of item responses are considered as an examinee’s responses on two administrations of the same form. The procedure is described below and is implemented with KKCLASS software (Kim, 2005).

- Step 1: Obtain item parameters (\mathbf{I}) and ability distribution weight ($\hat{g}(\theta)$) at each quadrature point from a single test.
- Step 2: Compute two raw scores at each quadrature point. At a given quadrature point θ_i , generate two sets of item responses using the item parameters from a test form, assuming that the same test form was administered twice to an examinee with the true ability θ_i .
- Step 3: Construct a classification matrix at each quadrature point. Determine the joint event for the cells in the table above using the raw scores obtained from Step 2.
- Step 4: Repeat Steps 2 and 3 R times and get average values over R replications.
- Step 5: Multiply distribution weight ($\hat{g}(\theta)$) by average values in Step 4 for each quadrature point and sum across all quadrature points. From this final contingency table, decision consistency indices, such as consistency agreement and kappa, can be computed.
- Step 6: Because examinee ability is estimated at each quadrature point, this quadrature point can be considered the true score. Therefore, decision accuracy is computed using both examinee estimated ability (observed scores) and quadrature point (true score).

Tables 9.6 and 9.7 show the results for the 2010 MAP classification analyses. Classification consistency and classification accuracy conditioned on level of achievement (Table 9.6) and on cut score (Table 9.7) are presented. As can be seen in Table 9.6, classification accuracy conditioned on achievement level ranged from 0.65 to 0.91, and classification consistency conditioned on achievement level ranged from 0.54 to 0.84. The magnitude of classification consistency and accuracy measures is influenced by key features of the test design including the number of items, number of cut scores, and the reliability and associated SEM.

Perhaps the most important indices for accountability systems are those for the accuracy and consistency of classification decisions made at specific cut points. To evaluate decisions at specific cut points, the joint distribution of all the performance levels is collapsed into a dichotomized distribution around that specific cut point. As an example, the dichotomization at the cut point between the *Basic* and *Proficient* classifications was formed. The proportion of correct classifications below this particular cut point is equal to the sum of all the cells at the levels *Below Basic* and *Basic*, and the proportion of correct classifications above that particular cut point is equal to the sum of all the cells at the levels *Proficient* and *Advanced*. Table 9.7 shows the classification accuracy and consistency estimates when conditioned on MAP cut scores. The classification accuracy statistics are at or above 0.90 while the classification consistency statistics are at or above 0.87. These results suggest that consistent and accurate performance-level classifications are being made for students in Missouri based on the MAP.

9.2.5 Convergent Validity

Convergent validity is a subtype of construct validity that can be estimated by the extent to which measures of constructs that theoretically should be related to each other are, in fact, observed as related to each other. Analyses of the internal structure of a test can indicate the extent to which the relationships among test items conform to the construct the test purports to measure. For example, the MAP Mathematics test is designed to measure a single overall construct—Mathematics achievement; therefore, the items comprising the Mathematics MAP should only measure Mathematics, not Science, Language, or Reading.

This Technical Report summarizes additional statistics that contribute to construct validity (Cronbach's coefficient alpha reported previously in this section and item fit reported in Chapter 6). The internal consistency coefficient (Cronbach's alpha) reported above is a measure of item homogeneity. In order for a group of items to be homogeneous, they must measure the same construct (construct validity) or represent the same content domain (content validity). Because IRT models were used to calibrate test items and to report student scores, item fit is also relevant to construct validity. The extent to which test items function as the IRT model prescribes is relevant to the validation of test scores. As shown in Chapter 6, only nine items were flagged for poor model/data fit across all 14 grade level/content area MAP tests.

9.3 Principal Components Analysis

As another measure of construct validity, CTB examined the unidimensionality of each grade level/content area MAP test. One of the underlying assumptions of the IRT models used to scale the MAP is that the tests being calibrated are unidimensional, that is, items comprising the MAP in each grade level/content area measure a single content domain. For example, Mathematics items should measure Mathematics ability and not measure Reading skills. Standard 1.11 of the AERA, APA, & NCME (1999) Standards says:

If the rationale for a test use or interpretation depends on premises about the relationship among parts of the test, evidence concerning internal structure should be provided.

In this section, we examine the internal structure by evaluating the unidimensionality assumption through Principal Components Analysis (PCA). This analysis seeks evidence that there exists a single primary factor, the first principal component, which accounts for much of the relationship between items. The presence of a single or dominant factor suggests that a test is sufficiently unidimensional (i.e., measures one underlying construct).

A PCA was conducted on each grade level/content area MAP. A large first principal component is evident in each analysis. It is common to have additional eigenvalues greater than 1.0, which may suggest the presence of other factors.

For all grade level/content area MAP tests, the ratio of the variance accounted for by the first factor to the second and third is sufficiently large to support the claim that these tests are unidimensional. All of the MAP content area tests exhibit first principal components accounting for more than 15% of the test variance (see Tables 9.8 through 9.10). To further investigate the unidimensionality of the Communication Arts, Mathematics, and Science tests, the ratio of the first eigenvalue to the second eigenvalue was explored (see Tables 9.8 through 9.10). These ratios show that the first eigenvalue is at least five times as large as the second eigenvalue for most of the grade levels/content areas. This substantial difference in magnitude indicates that one factor appears to be dominant and that the Communication Arts, Mathematics, and Science tests are essentially unidimensional.

This evidence supports the claim that there is a dominant dimension underlying the items/tasks in each test and that scores from each test represent performance primarily determined by that ability. Construct-irrelevant variance, such as factual knowledge irrelevant to doing well in a subject, does not appear to create significant nuisance factors.

9.4 Analyses by Content Standard

Three sets of analyses were conducted for the content standard level in another attempt to assess the construct validity of the MAP. First, the reliability of each Content Standard was computed. Second, correlation coefficients that measure the relationship between the

Content Standards were computed. Finally, the SEM was computed for each reportable content standard.

9.4.1 Reliability of Content Standards

Cronbach's (1951) coefficient alpha was computed for each of the Content Standards by grade level/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 level/content area. 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 previously. In each table, 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.

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. As noted above, the value of the correlation coefficients will be affected by the limited number of items measuring each Content Standard. We expect to see a more modest relationship (smaller correlation coefficients) reported between the Content Standards as a consequence of the lower number of items measuring each content standard (e.g., Writing Formally and Informally). The PPM between two content standard subscores may be artificially low because of measurement error.

AERA, APA, & NCME (1999) Standard 1.18, states:

When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coefficients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported.

We can correct for the attenuation of the PPM statistically using Spearman's formula:

$$CAPP = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}}, \quad (9.5)$$

where r_{xy} is the PPM between two content standards, r_{xx} is the reliability of one of those content standards, and r_{yy} is the reliability for the other content standard.

Across all tables, the CAPPM indicate strong relationships between the content standards. In some cases, the CAPPM is greater than 1.0. “Disattenuated values greater than 1.00 indicate that measurement error is not randomly distributed” (Schumacker, 1996). The strong relationships suggested by the CAPPM in Tables 9.11 through 9.13 are further evidence of the validity of the test construct. Since the overall content area is comprised of the content standard subscores and the content area is expected to measure a single dimension, then we would expect that these subscores are also highly related.

9.4.3 Standard Error of Measurement of Content Standards

In this chapter, we report the SEM associated with each of the Content Standards in Tables 9.14 through 9.16 for Communication Arts, Mathematics, and Science, respectively. These SEMs are reported in the percent correct metric as content standards are reported in that metric.

9.5 Divergent (Discriminant) Validity

Measures of different constructs should not be highly correlated with each other. Divergent validity is a subtype of construct validity that can be assessed by the extent to which measures of constructs that theoretically should not be related to each other are, in fact, observed as not related to each other. Typically, correlation coefficients among measures of unrelated or distantly related constructs are examined in support of divergent validity.

To assess the divergent validity of the MAP tests, correlations were computed between the Communication Arts, Mathematics, and Science scale scores for students who took more than one MAP subject area test in 2010. These correlations are based on the census data and the results are shown in Table 9.17. The correlation coefficients ranged from 0.71 (between Communication Arts and Mathematics in Grade 3) to 0.82 (between Mathematics and Science in Grade 8). The correlation coefficients suggest that individual student scores for Communication Arts, Mathematics, and Science are moderately to highly related. The correlation coefficients between Science and the other two content areas in grade 8 suggest that the Science grade 8 MAP is highly related to the Communication Arts and Mathematics grade 8 MAP. The tests are not perfectly related to each other, suggesting that different constructs are being tapped; however, the test scores do appear at least moderately related to one another, suggesting they are tapping into a similar knowledge base. This is especially true of the Science grade 8 test. The Science MAP is comprised of many constructed-response items, which may help account for its relationship with the Communication Arts test. The Science MAP tests similar thinking skills and item types as are found in the Mathematics MAP, which may help account for the strong correlation between the Science and Mathematics grade 8 test scores.

9.6 Summary

In summary, the overall purpose of each of the test administration workshops and the ancillary materials is to keep districts informed about policies and procedures related to testing in general and the MAP in particular. The information imparted is clearly related to standardizing the administration of the MAP, maintaining the security of the assessment, allowing access to the assessments for special populations by clearly delineating appropriate accommodations, and by providing guidance on appropriate interpretations of the test results. These communication and training efforts by DESE and the ancillary information developed by CTB/McGraw-Hill address multiple best practices of the testing industry but in particular are related to the following Standards (AERA, APA, & NCME, 1999):

- Standard 1.11—If the rationale for a test use or interpretation depends on premises about the relationship among parts of the test, evidence concerning internal structure should be provided.
- Standard 1.18—When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coefficients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported.
- Standard 2.1—For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported.
- Standard 2.2—The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original scale units and in units of each derived score recommended for use in test interpretation.
- Standard 2.4—Each method of quantifying the precision or consistency of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select examinees for reliability analyses and descriptive statistics on these samples should be reported.
- Standard 2.14—Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of the cut scores.
- Standard 2.15—When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure, using the same form or alternate forms of the instrument.

Table 9. 1: Reliability in Communication Arts

Grade	Number of Items	Number of Score Points	Cronbach's Alpha
3	56	63	0.91
4	58	62	0.93
5	56	61	0.91
6	56	60	0.91
7	63	70	0.91
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	0.92
4	62	69	0.92
5	58	62	0.91
6	58	62	0.92
7	61	65	0.92
8	61	68	0.93

Table 9. 3: Reliability in Science

Grade	Number of Items	Number of Score Points	Cronbach's Alpha
5	63	82	0.90
8	66	86	0.93

Table 9. 4: SEM by Subgroup

Grade	Category	Group	CA SEM	MA SEM	SC SEM
3	Overall		11.05	11.45	
	Ethnicity	White (not Hispanic)	10.99	11.89	
		Black (not Hispanic)	11.36	10.25	
		Hispanic	11.00	10.97	
		Asian/Pacific Islander	11.40	13.01	
		Native American/Alaskan	12.79	10.82	
	Gender	Male	11.21	11.32	
Female		11.14	11.55		
Accommodations	No	10.99	11.89		
	Yes	12.01	10.91		
4	Overall		10.67	9.62	
	Ethnicity	White (not Hispanic)	11.16	9.44	
		Black (not Hispanic)	10.38	9.69	
		Hispanic	10.43	9.38	
		Asian/Pacific Islander	11.49	10.77	
		Native American/Alaskan	11.24	9.12	
	Gender	Male	10.47	9.79	
Female		11.26	9.42		
Accommodations	No	11.16	9.50		
	Yes	11.83	10.55		
5	Overall		10.66	12.24	10.12
	Ethnicity	White (not Hispanic)	10.60	12.48	9.98
		Black (not Hispanic)	10.88	12.49	11.13
		Hispanic	10.64	11.69	10.39
		Asian/Pacific Islander	12.46	13.69	10.23
		Native American/Alaskan	10.68	11.96	10.85
	Gender	Male	10.90	12.25	10.16
Female		10.69	11.99	9.82	
Accommodations	No	10.36	12.35	9.97	
	Yes	12.71	12.95	12.20	
6	Overall		10.05	10.88	
	Ethnicity	White (not Hispanic)	10.07	11.04	
		Black (not Hispanic)	10.29	11.45	
		Hispanic	9.69	10.36	
		Asian/Pacific Islander	10.80	11.92	
		Native American/Alaskan	10.69	10.95	
	Gender	Male	10.14	10.69	
Female		10.02	10.86		
Accommodations	No	9.75	10.92		
	Yes	12.11	11.72		

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

Grade	Category	Group	CA SEM	MA SEM	SC SEM
7	Overall		11.11	11.25	
	Ethnicity	White (not Hispanic)	10.92	10.69	
		Black (not Hispanic)	11.33	12.02	
		Hispanic	11.07	11.45	
		Asian/Pacific Islander	12.46	11.86	
		Native American/Alaskan	11.70	11.95	
	Gender	Male	11.25	11.08	
		Female	11.12	10.88	
	Accommodations	No	11.21	11.13	
		Yes	12.99	14.15	
8	Overall		10.15	10.97	8.51
	Ethnicity	White (not Hispanic)	10.06	10.53	8.42
		Black (not Hispanic)	10.86	12.42	9.21
		Hispanic	10.34	11.29	8.74
		Asian/Pacific Islander	11.65	11.76	8.67
		Native American/Alaskan	10.60	10.98	8.72
	Gender	Male	10.77	11.00	8.66
		Female	9.86	10.86	8.27
	Accommodations	No	9.89	10.35	8.06
		Yes	13.47	14.82	10.60

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	9	648	10	673	14
	4	612	8	662	9	691	14
	5	625	9	675	8	702	12
	6	631	8	676	8	704	11
	7	634	10	680	9	712	12
	8	639	10	696	8	723	10
Mathematics	3	568	9	628	10	667	17
	4	596	9	651	8	688	12
	5	605	13	668	9	706	13
	6	628	10	681	9	721	12
	7	640	12	685	8	724	10
	8	670	11	710	8	741	7
Science	5	626	9	669	8	692	9
	8	671	8	703	7	735	8

Table 9. 6: Decision Accuracy and Consistency Conditioned on Level of Achievement

Content Area	Grade	Accuracy				Consistency			
		Below Basic	Basic	Prof.	Adv.	Below Basic	Basic	Prof.	Adv.
Communication Arts	3	0.89	0.85	0.65	0.82	0.81	0.83	0.54	0.68
	4	0.89	0.88	0.68	0.85	0.83	0.81	0.57	0.75
	5	0.86	0.85	0.73	0.84	0.80	0.81	0.62	0.74
	6	0.86	0.83	0.74	0.85	0.77	0.80	0.65	0.70
	7	0.84	0.85	0.75	0.82	0.80	0.77	0.67	0.72
	8	0.82	0.86	0.76	0.82	0.74	0.84	0.67	0.74
Mathematics	3	0.84	0.85	0.77	0.82	0.77	0.82	0.69	0.64
	4	0.85	0.88	0.80	0.83	0.79	0.83	0.74	0.69
	5	0.85	0.87	0.77	0.86	0.74	0.82	0.72	0.75
	6	0.84	0.85	0.83	0.84	0.79	0.80	0.74	0.75
	7	0.86	0.84	0.82	0.90	0.79	0.77	0.78	0.80
	8	0.85	0.81	0.83	0.91	0.79	0.75	0.75	0.82
Science	5	0.83	0.82	0.71	0.87	0.78	0.78	0.61	0.76
	8	0.88	0.82	0.86	0.83	0.83	0.76	0.77	0.74

Table 9. 7: Decision Accuracy and Consistency at Achievement Cut Points

Content Area	Grade	Accuracy			Consistency		
		Below Basic/ Basic	Basic/ Prof.	Prof./Adv.	Below Basic/ Basic	Basic/ Prof.	Prof./Adv.
Communication Arts	3	0.98	0.90	0.91	0.97	0.87	0.88
	4	0.98	0.92	0.91	0.97	0.88	0.88
	5	0.98	0.92	0.92	0.97	0.88	0.88
	6	0.97	0.91	0.92	0.96	0.87	0.89
	7	0.97	0.91	0.93	0.96	0.87	0.90
	8	0.98	0.92	0.92	0.98	0.89	0.90
Mathematics	3	0.98	0.91	0.93	0.97	0.87	0.91
	4	0.98	0.91	0.95	0.97	0.87	0.93
	5	0.98	0.93	0.93	0.96	0.90	0.91
	6	0.97	0.93	0.94	0.96	0.90	0.91
	7	0.97	0.93	0.95	0.95	0.90	0.93
	8	0.96	0.93	0.95	0.94	0.91	0.93
Science	5	0.96	0.91	0.92	0.95	0.88	0.89
	8	0.96	0.93	0.96	0.94	0.89	0.94

Table 9. 8: Principal Component Analysis for Communication Arts

Grade	Eigenvalue	Percent of Variance Explained	Cumulative Percent of Variance Explained
Grade 3			
First Component	10.55	18.84	18.84
Second Component	1.56	2.79	21.63
Ratio (First/Second)	6.76		
Grade 4			
First Component	12.46	21.48	21.48
Second Component	1.92	3.31	24.79
Ratio (First/Second)	6.49		
Grade 5			
First Component	10.64	19.01	19.01
Second Component	1.68	3.01	22.01
Ratio (First/Second)	6.32		
Grade 6			
First Component	10.03	17.91	17.91
Second Component	1.49	2.67	20.58
Ratio (First/Second)	6.71		
Grade 7			
First Component	10.33	16.40	16.40
Second Component	1.72	2.73	19.13
Ratio (First/Second)	6.02		
Grade 8			
First Component	10.72	17.86	17.86
Second Component	1.70	2.83	20.69
Ratio (First/Second)	6.32		

Table 9.9: Principal Component Analysis for Mathematics

Grade	Eigenvalue	Percent of Variance Explained	Cumulative Percent of Variance Explained
Grade 3			
First Component	10.86	19.74	19.74
Second Component	1.67	3.03	22.77
Ratio (First/Second)	6.52		
Grade 4			
First Component	11.50	18.55	18.55
Second Component	1.70	2.74	21.30
Ratio (First/Second)	6.76		
Grade 5			
First Component	10.66	18.38	18.38
Second Component	1.81	3.13	21.50
Ratio (First/Second)	5.88		
Grade 6			
First Component	11.42	19.69	19.69
Second Component	1.81	3.12	22.81
Ratio (First/Second)	6.30		
Grade 7			
First Component	11.34	18.60	18.60
Second Component	1.75	2.87	21.47
Ratio (First/Second)	6.48		
Grade 8			
First Component	12.05	19.76	19.76
Second Component	1.84	3.02	22.79
Ratio (First/Second)	6.54		

Table 9.10: Principal Component Analysis for Science

Grade	Eigenvalue	Percent of Variance Explained	Cumulative Percent of Variance Explained
Grade 5			
First Component	9.85	15.63	15.63
Second Component	1.65	2.62	18.25
Ratio (First/Second)	5.97		
Grade 8			
First Component	11.99	18.17	18.17
Second Component	1.86	2.81	20.98
Ratio (First/Second)	6.46		

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	12	0.68	0.85	0.82		0.85
	2	Reading Fiction/Poetry/Drama	23	0.62	0.79	0.93		1.11
	3	Reading Nonfiction	18	0.60	0.74	0.80		1.12
	4	Writing Formally/Informally	NR*					
	5	Combined Reading	41	0.66	0.92	0.94		0.88
4	1	Speaking/Writing Standard English	12	0.63	0.74	0.78		0.78
	2	Reading Fiction/Poetry/Drama	20	0.54	0.85	0.91		1.04
	3	Reading Nonfiction	24	0.57	0.77	0.85		1.09
	4	Writing Formally/Informally	NR					
	5	Combined Reading	44	0.59	0.92	0.96		0.91
5	1	Speaking/Writing Standard English	12	0.57	0.85	0.87		0.88
	2	Reading Fiction/Poetry/Drama	22	0.57	0.79	0.89		1.10
	3	Reading Nonfiction	21	0.60	0.72	0.84		1.06
	4	Writing Formally/Informally	NR					
	5	Combined Reading	43	0.63	0.93	0.92		0.90
6	1	Speaking/Writing Standard English	12	0.63	0.84	0.88		0.88
	2	Reading Fiction/Poetry/Drama	21	0.60	0.81	0.92		1.11
	3	Reading Nonfiction	22	0.63	0.75	0.81		1.09
	4	Writing Formally/Informally	NR					
	5	Combined Reading	43	0.66	0.94	0.93		0.89
7	1	Speaking/Writing Standard English	16	0.65	0.83	0.90		0.89
	2	Reading Fiction/Poetry/Drama	24	0.59	0.78	0.89		1.13
	3	Reading Nonfiction	20	0.65	0.71	0.81		1.08
	4	Writing Formally/Informally	NR					
	5	Combined Reading	44	0.67	0.94	0.91		0.88
8	1	Speaking/Writing Standard English	16	0.68	0.82	0.86		0.86
	2	Reading Fiction/Poetry/Drama	21	0.60	0.79	0.94		1.13
	3	Reading Nonfiction	21	0.65	0.76	0.84		1.08
	4	Writing Formally/Informally	NR					
	5	Combined Reading	42	0.67	0.95	0.93		0.89

*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
3	1	Number and Operations	18	0.78	0.95	0.78	0.92	
	2	Algebraic Relationship	11	0.70	0.70	0.78	0.96	
	3	Geometric and Spatial	12	0.54	0.52	0.63	0.78	
	4	Measurement	9	0.67	0.66	0.50	0.67	
	5	Data and Probability	NR*					
4	1	Number and Operations	23	0.84	0.98	0.73	0.93	0.76
	2	Algebraic Relationship	13	0.73	0.66	0.77	0.93	0.79
	3	Geometric and Spatial	9	0.52	0.48	0.60	0.81	0.65
	4	Measurement	11	0.71	0.63	0.52	0.69	0.82
	5	Data and Probability	6	0.50	0.46	0.36	0.49	0.52
5	1	Number and Operations	15	0.77	0.96	0.81	1.00	0.88
	2	Algebraic Relationship	13	0.74	0.76	0.86	0.98	0.97
	3	Geometric and Spatial	10	0.55	0.58	0.60	0.86	0.86
	4	Measurement	10	0.68	0.66	0.52	0.60	0.96
	5	Data and Probability	10	0.63	0.68	0.54	0.61	0.66
6	1	Number and Operations	17	0.81	0.98	0.90	0.99	0.95
	2	Algebraic Relationship	11	0.73	0.69	0.96	0.96	0.94
	3	Geometric and Spatial	8	0.62	0.61	0.58	0.91	0.93
	4	Measurement	8	0.68	0.61	0.53	0.58	0.95
	5	Data and Probability	14	0.74	0.68	0.61	0.63	0.75
7	1	Number and Operations	14	0.79	0.91	0.91	0.92	0.90
	2	Algebraic Relationship	19	0.72	0.80	0.95	0.89	0.94
	3	Geometric and Spatial	11	0.64	0.67	0.62	0.93	0.93
	4	Measurement	7	0.62	0.61	0.56	0.58	0.92
	5	Data and Probability	10	0.64	0.67	0.59	0.56	0.64
8	1	Number and Operations	13	0.73	0.91	0.93	0.94	0.95
	2	Algebraic Relationship	18	0.71	0.82	0.92	0.94	1.04
	3	Geometric and Spatial	15	0.66	0.69	0.68	0.97	1.00
	4	Measurement	6	0.61	0.64	0.60	0.57	0.98
	5	Data and Probability	9	0.65	0.76	0.67	0.60	0.65

*NR=Not Reported

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	1.01	0.91	0.96	1.04	0.93	0.84	0.85
	2	Force and Motion	5	0.52	0.45	0.87	1.00	1.02	0.89	0.89	0.95
	3	Characteristics of Living Organisms	7	0.50	0.42	0.52	0.91	0.94	0.83	0.77	0.84
	4	Interactions of Organisms	7	0.56	0.51	0.51	0.59	0.99	0.86	0.82	0.96
	5	Earth's Processes	7	0.62	0.54	0.53	0.59	0.61	0.97	0.89	0.98
	6	The Universe	6	0.53	0.44	0.45	0.49	0.57	0.56	0.81	0.81
	7	Scientific Inquiry	16	0.51	0.48	0.44	0.50	0.55	0.49	0.64	0.80
	8	Technology and the Environment	6	0.39	0.38	0.36	0.44	0.46	0.36	0.38	0.36
8	1	Matter and Energy	9	0.66	0.94	0.98	0.94	1.02	0.87	0.90	0.93
	2	Force and Motion	6	0.46	0.37	0.93	0.90	0.94	0.79	0.89	0.95
	3	Characteristics of Living Organisms	6	0.56	0.39	0.49	1.00	1.05	0.85	0.89	1.05
	4	Interactions of Organisms	6	0.57	0.41	0.53	0.56	0.97	0.84	0.84	0.88
	5	Earth's Processes	8	0.62	0.43	0.56	0.55	0.57	0.85	0.88	1.03
	6	The Universe	6	0.54	0.37	0.45	0.48	0.49	0.59	0.74	0.77
	7	Scientific Inquiry	19	0.66	0.49	0.56	0.57	0.60	0.51	0.81	0.90
	8	Technology and the Environment	6	0.56	0.43	0.55	0.49	0.58	0.44	0.61	0.56

Table 9. 14: Mean, Standard Deviation, and Standard Error of Measurement (SEM) of Communication Arts Content Standards

Grade	Content Standard	Mean	Std. Deviation	SEM
3	1	72.51	19.74	11.17
	2	76.95	16.66	7.63
	3	71.61	19.95	8.92
	5	74.34	17.19	5.95
4	1	76.58	16.29	9.91
	2	81.03	19.64	7.61
	3	71.44	19.99	7.74
	5	75.43	18.98	5.69
5	1	67.63	17.42	11.42
	2	70.39	17.62	8.07
	3	74.53	21.17	8.47
	5	72.24	18.09	5.72
6	1	67.64	19.92	12.12
	2	71.96	18.75	8.17
	3	72.22	19.43	8.47
	5	72.08	17.98	5.96
7	1	62.26	17.82	10.54
	2	72.25	16.83	7.89
	3	72.82	19.86	8.66
	5	72.49	17.04	5.90
8	1	55.90	19.58	11.08
	2	67.23	19.06	8.73
	3	77.08	19.96	7.98
	5	71.73	18.31	6.07

Table 9. 15: Mean, Standard Deviation, and Standard Error of Measurement (SEM) of Mathematics Content Standards

Grade	Content Standard	Mean	Std. Deviation	SEM
3	1	77.11	17.87	8.38
	2	77.10	20.85	11.42
	3	81.39	16.45	10.01
	4	70.60	22.21	12.76
4	1	74.01	19.08	7.63
	2	72.67	19.35	11.28
	3	74.23	19.10	12.08
	4	62.15	22.62	12.59
	5	78.36	20.66	14.31
5	1	66.37	22.09	10.59
	2	69.14	23.01	11.27
	3	83.46	17.07	10.80
	4	66.21	22.22	14.05
	5	78.95	19.52	11.38
6	1	67.77	23.05	10.05
	2	65.95	19.84	11.05
	3	76.48	19.89	12.89
	4	69.81	22.64	14.67
	5	71.69	20.89	10.45
7	1	68.59	23.42	10.73
	2	56.24	20.97	9.38
	3	63.01	20.41	12.58
	4	68.40	24.33	15.77
	5	68.24	18.88	11.33
8	1	68.21	21.97	11.42
	2	52.07	24.25	10.29
	3	57.43	19.84	11.22
	4	54.95	28.34	18.58
	5	53.09	23.81	14.09

Table 9. 16: Mean, Standard Deviation, and Standard Error of Measurement (SEM) of Science Content Standards

Grade	Content Standard	Mean	Std. Deviation	SEM
5	1	52.07	20.87	13.53
	2	57.28	22.32	16.55
	3	73.16	20.32	14.08
	4	58.75	23.36	14.96
	5	53.06	25.46	15.90
	6	60.08	22.53	14.94
	7	65.76	15.76	9.46
	8	61.09	20.09	16.07
8	1	50.92	20.73	12.09
	2	53.65	22.07	17.52
	3	45.68	18.07	12.90
	4	54.34	25.86	17.15
	5	59.51	21.31	13.97
	6	38.45	22.45	14.38
	7	60.20	20.04	8.74
	8	54.40	20.85	13.83

Table 9. 17: Inter-Correlation of Communication Arts, Mathematics, and Science Scale Scores

Grade	CA/MA	CA/SC	MA/SC
3	0.71	-	-
4	0.73	-	-
5	0.74	0.77	0.78
6	0.75	-	-
7	0.77	-	-
8	0.75	0.80	0.82

Figure 9. 1: SEM Plot Communication Arts, Grades 3 – 8

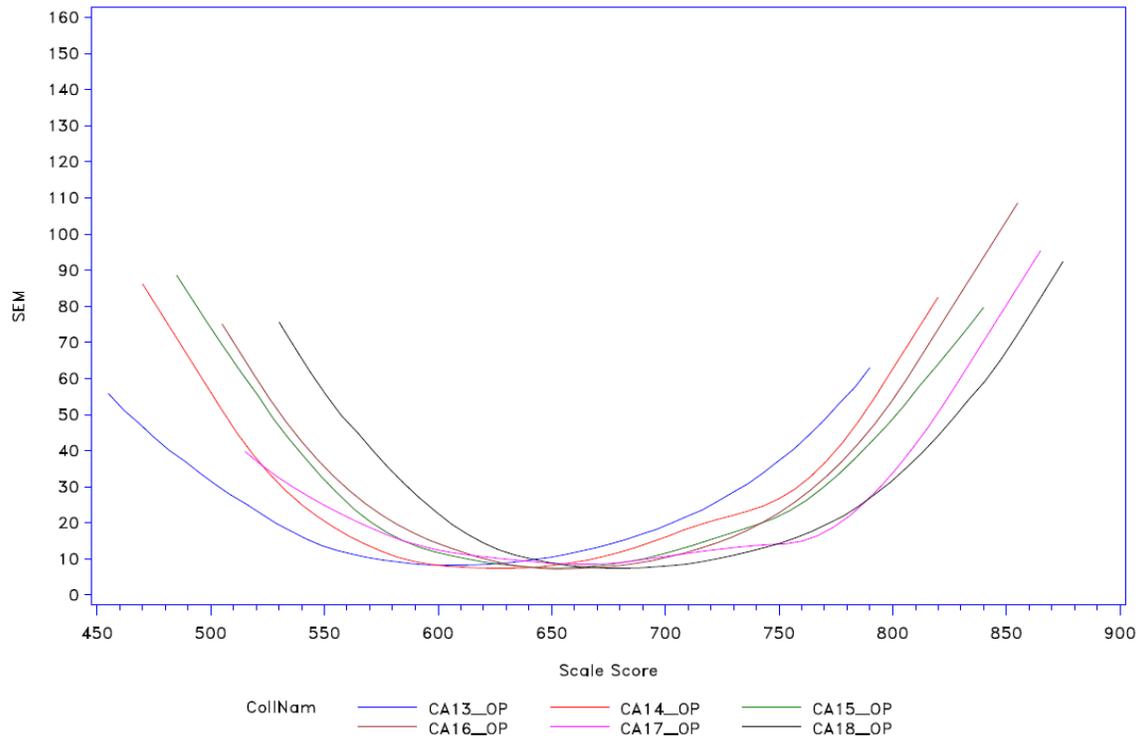


Figure 9. 2: SEM Plot Mathematics, Grades 3 – 8

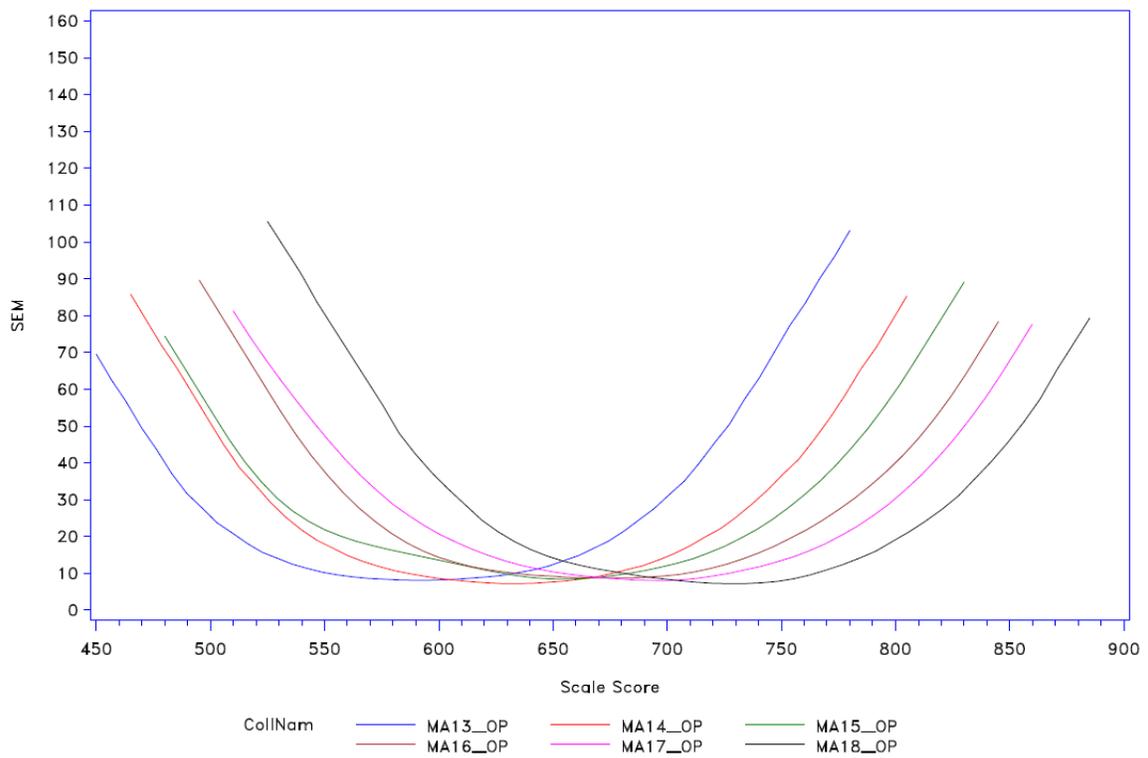
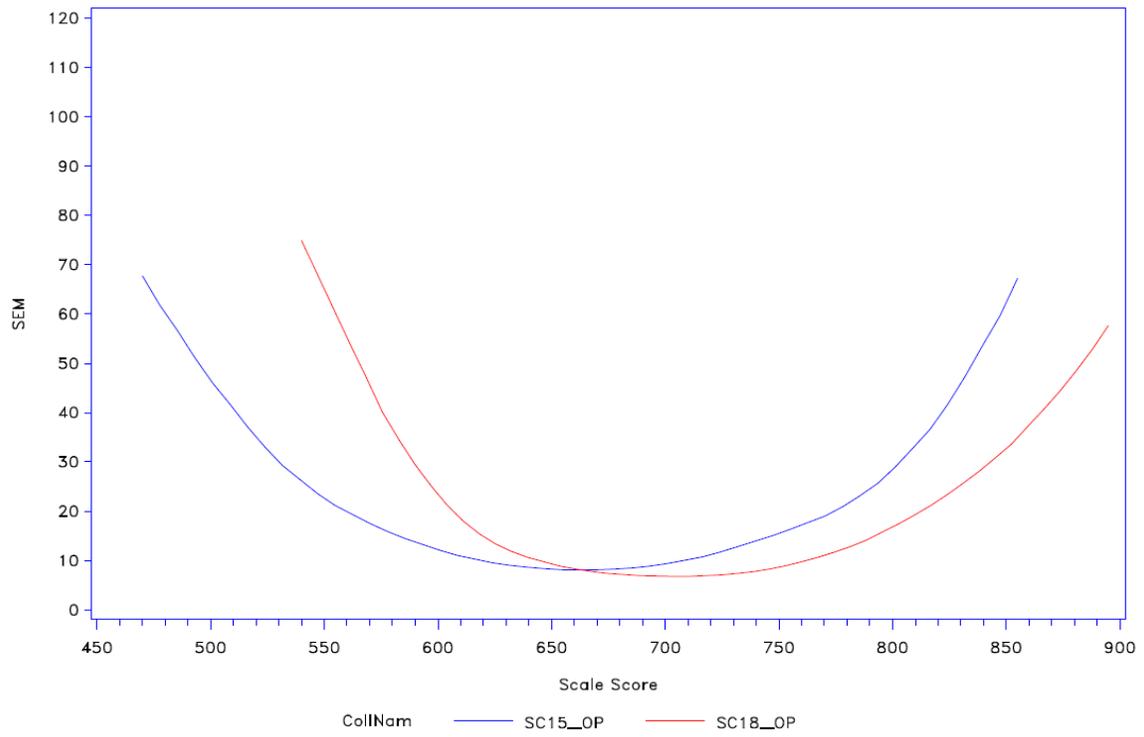


Figure 9. 3: SEM Plot Science, Grade 5 and 8



CHAPTER 10: FAIRNESS

As noted in the Standards (AERA, APA, & NCME, 1999), there are varying definitions of fairness. In this chapter, we examine fairness as it relates to minimizing bias on a test. We then look at test performance among varying subgroups assessed by the MAP. It should be noted that differences in test performance among subgroups does not mean that a test is unfair—it simply means that groups perform differentially on the test. Even when a test is carefully and properly constructed, differences may exist among subgroups as a result of differences in curriculum or learning by students in the subgroups.

This chapter is particularly relevant to AERA, APA, & NCME (1999) Standards 7.1, 7.2, 7.3, and 7.4. Standards 7.1 through 7.4 are from Chapter 7 of the AERA, APA, & NCME (1999) Standards, which is titled “Fairness in Testing and Test Use.” Each of these Standards will be presented, as will the way the Standard is addressed in this chapter.

Standard 7.1 *When credible research reports that test scores differ in meaning across examinee subgroups for the type of test in question, then to the extent feasible, the same forms of validity evidence collected for the examinee population as a whole should also be collected for each relevant subgroup. Subgroups may be found to differ with respect to appropriateness of test content, internal structure of test responses, the relation of test scores to other variables, or the response processes employed by individual examinees. Any such findings should receive due consideration in the interpretation and use of scores as well as in the subsequent test revisions.*

There is no particular research on the MAP showing that the test scores of examinee subgroups differ in meaning; however, this is an ongoing concern in any large-scale testing program. To lessen the possibility of differences in test score meaning, CTB has several steps that are followed in item development and selections as is explicated in Section 10.1 of this chapter. Also, DESE conducts content and bias reviews on items as is explained in Chapter 3.

Standard 7.2 *When credible research reports differences in the effects of construct-irrelevant variance across subgroups of test takers on performance on some part of the test, the test should be used if at all only for those subgroups for which evidence indicates that valid inferences can be drawn from test scores.*

Again, there is no research on the MAP showing differences in the effects of construct-irrelevant variance across subgroups; however, DESE and CTB undertake steps to minimize construct-irrelevant variance through the test development process outlined in Section 10.1 of this chapter and explained in detail in Chapter 3.

Standard 7.3 *When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability, and/or linguistic groups in the population of test takers in the content domain measured by the test, test developers*

should conduct appropriate studies when feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups.

CTB conducts DIF studies following the field test and the operational administration of the MAP. During the field test phase of the project, items flagged for DIF will be further examined for possible bias. Items flagged for bias will be removed from the item pool. Section 10.2 of this chapter explains the steps taken to evaluate MAP items through the use of DIF.

Standard 7.4 *Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the domain.*

Section 10.1 of this chapter is directly relevant to Standard 7.4. In this section, we explain the steps taken by CTB to minimize words, phrases, and content that may be regarded as offensive by members of particular demographic subgroups. Section 3.2.5 of Chapter 3 discusses the Content and Bias Review conducted for the MAP. This review is also critical in fulfilling Standard 7.4.

10.1 Minimizing Bias through Careful Test Development

The development of a test that is fair for all examinees begins in the early stages of planning and development. The item and test development processes that were used to minimize bias are summarized below.

First, careful attention was paid to content validity during the item development and item selection processes. Bias can occur only if the test is measuring different things for different groups. By eliminating irrelevant skills or knowledge from the items, the possibility of bias is reduced.

Second, item writers and test developers followed several published guidelines for reducing or eliminating bias. These included *Guidelines for Bias-Free Publishing* (Macmillan/McGraw-Hill, 1993a) and *Reflecting Diversity: Multicultural Guidelines for Educational Publishing Professionals* (Macmillan/McGraw-Hill, 1993b). Test developers reviewed the items and other testing materials with these guidelines in mind. Internal editorial reviews were conducted by at least three different people: a content editor who directly supervised the item writers; a style editor; and a content supervisor. The final test was again reviewed by at least these same people, and was also subjected to an independent review by a quality assurance editor.

Third, careful attention is given to item statistics throughout the test development process. As part of the test assembly process, attempts are made to avoid using or reusing items with poor statistical fit or distractors with positive point biserial correlations, since this may indicate that an item is tapping an ability that is irrelevant to the construct being

measured. DIF statistics are also examined during test construction. Items that have exhibited significant DIF against one or more subgroups are removed from further consideration unless it is essential to include them in order to meet content specifications.

Additional steps to reduce bias, including the use of Bias Review committees comprised of Missouri participants, are described in more detail in Chapter 3 of this report.

10.2 Evaluating Bias through Differential Item Functioning Statistics

After administering the test, an empirical approach known as DIF was used to examine the items. The DIF statistics indicate the degree to which members of a particular subgroup performs better or worse than expected on each item as compared to the reference group. The DIF procedures used and the results of these analyses are detailed in this section.

The position of CTB/McGraw-Hill concerning test bias is based on two general propositions. First, students may differ in their background knowledge, cognitive and academic skills, language, attitudes, and values. To the degree that these differences are large, no one curriculum and no one set of instructional materials will be equally suitable for all. Therefore, no one test will be equally appropriate for all. Furthermore, it is difficult to specify what amount of difference can be called large and to determine how these differences will affect the outcome of a particular test. Second, schools have been assigned the tasks of developing certain basic cognitive skills and supporting development of these skills equitably among all students. Therefore, there is a need for tests that measure the common skills and bodies of knowledge that are common to all learners. The test publisher's task is to develop assessments that measure these key cognitive skills without introducing extraneous or construct-irrelevant elements into the performances on which the measurement is based. If these tests require that students have culturally-specific knowledge and skills not taught in school, differences in performance among students can occur because of differences in student background and out-of-school learning. Such tests are measuring different things for different groups and can be called biased (Camilli & Shepard, 1994; Green, 1975).

In order to lessen this bias, CTB/McGraw-Hill strives to minimize the role of the extraneous elements, thereby increasing the number of students for whom the test is appropriate. As discussed above and in Chapter 3 of this report, careful attention is given during the test development and test construction processes to lessen the influence of these elements for large numbers of students (including the use of Bias Review committees). Unfortunately, in some cases these elements may continue to play a substantial role. To assess the extent to which items may be performing differently for various subgroups of interest, DIF analyses are conducted after each operational test administration.

DIF statistics are used to quantify differences in item performance between two groups after controlling for examinees' overall achievement level. Two DIF statistics that are commonly used for this purpose are the Mantel-Haenszel (MH) statistic (1959) and the

Standardized Mean Difference (SMD) between the reference and focal groups, proposed by Dorans and Schmitt (1991).

The MH statistic (Zwick, Donoghue, & Grima, 1993) is computed as:

$$\text{Mantel } \chi^2 = \frac{\left(\sum_k F_k - \sum_k E(F_k) \right)^2}{\sum_k \text{Var}(F_k)},$$

where F_k is the sum of scores for the focal group at the k^{th} level of the matching variable. Note that the MH statistic is sensitive to sample size such that larger sample sizes increase the value of chi-square.

In addition to the MH chi-square statistic, the delta statistic (MH-D DIF) was computed for all items. Educational Testing Service (ETS) first developed the MH-D DIF statistic. To compute delta, alpha (the odds ratio) is first computed as:

$$\alpha_{MH} = \frac{\sum_{k=1}^K N_{r1k}N_{f0k} / N_k}{\sum_{k=1}^K N_{f1k}N_{r0k} / N_k},$$

where N_{r1k} is the number of correct responses in the reference group at ability level k , N_{f0k} is the number of incorrect responses in the focal group at ability level k , N_k is the total number of responses, N_{f1k} is the number of correct responses in the focal group at ability level k , and N_{r0k} is the number of incorrect responses in the reference group at ability level k . MH-D DIF is then computed as:

$$\text{MH-D DIF} = -2.35 \ln(\alpha_{MH}).$$

For SR items, the MH (χ^2_{MH}) statistic was used to evaluate potential DIF items. In the MH procedure, subgroups are matched by their raw total test score using a contingency table with k ability levels. When applying the MH procedure, the log-odds ratio α is assumed to be constant across the k matched levels. The χ^2_{MH} , then, estimates a pooled common-odds ratio. Taking the natural logarithm of the common-odds ratio and its confidence limits and multiplying these with the constant -2.35 , the resulting values may then be placed on the MH delta metric (Δ_{MH}) for interpretive purposes. Items were flagged for DIF using the following criteria:

- Moderate DIF: Significant MH chi-square statistic ($p < 0.05$) and $1.0 \leq |\text{MH-D DIF}| < 1.5$
- Large DIF: Significant Mantel-Haenszel chi-square statistic ($p < 0.05$) and $|\text{MH-D DIF}| \geq 1.5$

For CR items, an effect size (ES) statistic based on the MH chi-square will be used. The ES is obtained by dividing the SMD statistics by the standard deviation of the item. The SMD is an effect size index of DIF, which is relatively easy to interpret (Zwick et al., 1993). The SMD compares the means of the reference and focal group, adjusting for the distribution of reference and focal group members on the conditioning variable (Zwick et al., 1993), which for these analyses is the MAP raw score. SMD is computed as (Zwick et al., 1993):

$$SMD = p_{fk} \left(\sum_k m_{Fk} - \sum_k m_{Rk} \right),$$

where p_{fk} = proportion of the focal group members at the k th level of the matching variable, $m_{Fk} = 1/N_{F1k}$, and $m_{Rk} = 1/N_{R1k}$. Items are flagged using the same rules that are used in NAEP:

- Moderate DIF: If the MH statistic is significant ($p < .05$) and $|ES|$ is between 0.17 and 0.25.
- Large DIF: If the MH statistic is significant ($p < .05$) and $|ES| \geq 0.25$.

A positive DIF value indicates that the item favors the focal group, while a negative value indicates that the item disadvantages the focal group. Tables 10.1, 10.2, and 10.3 show the DIF results for the following subgroups:

- **Gender:** Focal group is Females; reference group is Males.
- **Ethnicity:** Focal groups are Black, Hispanic, Asian/Pacific Islander, Native American/Alaskan; reference group is White.
- **Accommodations:** Focal group is students who received one or more testing accommodations; reference group is all others.

A negative SMD value implies that the focal group has a lower mean item score than the reference group, whereas a positive value implies that the focal group has a higher mean item score than the reference group, conditioned on the matching test score.

The DIF analyses are not performed for subgroups of less than 100. In these cases, the statistical procedures do not have sufficient power to detect differences should they exist.

Tables 10.1, 10.2, and 10.3 summarize the number of DIF flags by grade for each focal group. They also show the number of items on each test, as well as the sample size of each subgroup. For example, in Grade 6 Communication Arts (see Table 10.1) there was one item flagged for DIF for the accommodated subgroup. In this case, the flagged item exhibited moderate negative DIF. Three items were flagged for DIF for the female subgroup: two items exhibited moderate negative DIF while the one exhibited moderate positive DIF. One item was flagged for moderate negative DIF for the Hispanic

subgroup. One item was flagged for moderate negative DIF for the Asian/Pacific Islander subgroup. Finally, one item was flagged for moderate positive DIF for of the Native American/Alaskan subgroup.

Again, any items included on the MAP (including those items flagged for DIF) have been thoroughly reviewed for content and bias by Missouri teachers, DESE staff, and CTB Content Development staff. Further, these items were reviewed for possible DIF flags during the field test stage of test development. The DIF flags found on the operational assessment do not necessarily indicate that an item is biased; rather, DIF flags indicate that the item functions differently for equally able members of different groups (Camilli & Shepard, 1994). All items flagged for DIF in the tables stated above had been thoroughly reviewed before inclusion on the operational MAP to insure that they do not tap knowledge or specific ability irrelevant to the construct the test intends to measure. Items are not necessarily suppressed from operational scoring if they are flagged for DIF.

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 scale score means and standard deviations, numbers of students, 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 of 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:

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

where \overline{x}_a is the mean score of group A, \overline{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.

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 grades. There is a small difference between the mean test scores of Hispanic and White students, where Hispanics underperform White students on the Communication Arts tests. 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 Grades 4, 5, and 8. There is a small difference in mean Communication Arts test scores, where Asian/Pacific Islander students outperform White students in all grades except Grade 3.

There is a medium difference between the mean Mathematics test scores of Black and White students, where Black students underperform White students in all grades, except Grade 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 compared to White students, where Native American students underperform White students in all grades. 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.

There is a large difference between the mean Science test scores of Black students compared to 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 compared to 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 compared to White students in Grades 5 and 8, where Native American students underperform White students.

10.4 Summary

In summary, the overall purpose of this chapter is to address fairness concerns that are relevant to the administration of the MAP. The information in this chapter addresses multiple best practices of the testing industry, and in particular are related to the following Standards (AERA, APA, & NCME, 1999):

- Standard 7.1—When credible research reports that test scores differ in meaning across examinee subgroups for the type of test in question, then to the extent feasible, the same forms of validity evidence collected for the examinee population as a whole should also be collected for each relevant subgroup. Subgroups may be found to differ with respect to appropriateness of test content, internal structure of test responses, the relation of test scores to other variables, or the response processes employed by individual examinees. Any such findings should receive due consideration in the interpretation and use of scores as well as in the subsequent test revisions.
- Standard 7.2—When credible research reports differences in the effects of construct-irrelevant variance across subgroups of test takers on performance on some part of the test, the test should be used if at all only for those subgroups for which evidence indicates that valid inferences can be drawn from test scores.
- Standard 7.3—When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability, and/or linguistic groups in the population of test takers in the content domain measured by the test, test developers should conduct appropriate studies when feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups.
- Standard 7.4—Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the domain.

Table 10. 1: 2010 MAP DIF Statistics: Number of Flagged Items, Communication Arts

Grade	Group	Sample Size	Large Negative	Moderate Negative	Moderate Positive	Large Positive	Number of Items
3	Accommodated	6271	1	1			56
	Asian/Pacific Islander	1411	1	4	1		56
	Native American/Alaskan	308					56
	Black (not Hispanic)	11819	1	1	1		56
	Hispanic	3093		1			56
	Female	32841		1		1	56
4	Accommodated	7006		4			58
	Asian/Pacific Islander	1361	1	2	2	1	58
	Native American/Alaskan	281			1	1	58
	Black (not Hispanic)	12103			1		58
	Hispanic	2963		1			58
	Female	32648				1	58
5	Accommodated	7032		1	1		56
	Asian/Pacific Islander	1279		4	2		56
	Native American/Alaskan	313					56
	Black (not Hispanic)	11884					56
	Hispanic	2772	1				56
	Female	32334		3			56
6	Accommodated	7167		1			56
	Asian/Pacific Islander	1259		1			56
	Native American/Alaskan	337			1		56
	Black (not Hispanic)	12056					56
	Hispanic	2744		1			56
	Female	32674		2	1		56
7	Accommodated	6778		1			63
	Asian/Pacific Islander	1227	1	1	1		63
	Native American/Alaskan	326					63
	Black (not Hispanic)	11553		1			63
	Hispanic	2515					63
	Female	32318		2	1	1	63
8	Accommodated	6437		1			60
	Asian/Pacific Islander	1220	2	6	4		60
	Native American/Alaskan	344					60
	Black (not Hispanic)	11273	1	1		1	60
	Hispanic	2274		1	2		60
	Female	32388	2	2	4		60

Table 10. 2: 2010 MAP DIF Statistics: Number of Flagged Items, Mathematics

Grade	Group	Sample Size	Large Negative	Moderate Negative	Moderate Positive	Large Positive	Number of Items
3	Accommodated	6552	1	1	1		55
	Asian/Pacific Islander	1458		2	4		55
	Native American/Alaskan	310	1				55
	Black (not Hispanic)	11841		2	2		55
	Hispanic	3118		1			55
	Female	32897		2			55
4	Accommodated	7333		3			62
	Asian/Pacific Islander	1393		1			62
	Native American/Alaskan	282					62
	Black (not Hispanic)	12123			1		62
	Hispanic	3010		2			62
	Female	32707					62
5	Accommodated	7340		2	1		58
	Asian/Pacific Islander	1322	2	4	3	1	58
	Native American/Alaskan	315			2		58
	Black (not Hispanic)	11909		2	3		58
	Hispanic	2790		2			58
	Female	32401		1			58
6	Accommodated	7480			2	2	58
	Asian/Pacific Islander	1290		4	2		58
	Native American/Alaskan	338					58
	Black (not Hispanic)	12070	1	2	2		58
	Hispanic	2765		1			58
	Female	32711		3	2		58
7	Accommodated	7025	1		1	3	61
	Asian/Pacific Islander	1263	2	2	2		61
	Native American/Alaskan	325					61
	Black (not Hispanic)	11565	1	3	1		61
	Hispanic	2530		1			61
	Female	32330		2	1		61
8	Accommodated	6787	1	1	1		61
	Asian/Pacific Islander	1255	2	1	2		61
	Native American/Alaskan	345					61
	Black (not Hispanic)	11272		2			61
	Hispanic	2300		1			61
	Female	32403		2			61

Table 10. 3: 2010 MAP DIF Statistics: Number of Flagged Items, Science

Grade	Group	Sample Size	Large Negative	Moderate Negative	Moderate Positive	Large Positive	Number of Items
5	Accommodated	7014					63
	Asian/Pacific Islander	1321		5	2		63
	Native American/Alaskan	314					63
	Black (not Hispanic)	11904		2	1		63
	Hispanic	2791		1	1		63
	Female	32395		2	3		63
8	Accommodated	6498		2			66
	Asian/Pacific Islander	1253	1	2	3		66
	Native American/Alaskan	345		1			66
	Black (not Hispanic)	11226			1		66
	Hispanic	2298			3		66
	Female	32389		6	4		66

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)	50030	645.12	34.74		0.90
		Black (not Hispanic)	11842	621.73	37.87	0.66	0.91
		Hispanic	3096	628.53	34.78	0.48	0.90
		Asian/Pacific Islander	1413	649.93	38.01	-0.14	0.91
		Native American/Alaskan	309	640.31	40.45	0.14	0.90
	Gender	Male	33811	635.76	37.37		0.91
		Female	32867	644.93	35.24	-0.25	0.90
	Accommodations	No	60425	644.57	33.15		0.89
		Yes	6329	599.12	42.46	1.33	0.92
Mathematics	Ethnicity	White (not Hispanic)	50022	630.02	37.61		0.90
		Black (not Hispanic)	11841	604.12	38.75	0.68	0.93
		Hispanic	3119	614.72	36.57	0.41	0.91
		Asian/Pacific Islander	1458	640.57	43.35	-0.28	0.91
		Native American/Alaskan	310	622.00	36.08	0.21	0.91
	Gender	Male	33837	624.56	40.02		0.92
		Female	32901	625.26	38.49	-0.02	0.91
	Accommodations	No	60277	628.53	37.60		0.90
		Yes	6538	591.30	38.56	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)	50485	666.30	37.21		0.91
		Black (not Hispanic)	12119	642.42	39.24	0.63	0.93
		Hispanic	2972	649.50	36.86	0.45	0.92
		Asian/Pacific Islander	1361	674.24	43.43	-0.21	0.93
		Native American/Alaskan	281	655.83	42.50	0.28	0.93
	Gender	Male	34542	656.26	39.56		0.93
		Female	32674	666.77	37.52	-0.27	0.91
	Accommodations	No	60229	666.41	35.29		0.90
		Yes	7072	618.17	41.84	1.34	0.92
Mathematics	Ethnicity	White (not Hispanic)	50498	652.32	31.47		0.91
		Black (not Hispanic)	12127	628.26	36.63	0.74	0.93
		Hispanic	3010	639.55	31.26	0.41	0.91
		Asian/Pacific Islander	1393	663.64	38.07	-0.36	0.92
		Native American/Alaskan	282	643.24	32.23	0.29	0.92
	Gender	Male	34596	647.08	34.61		0.92
		Female	32708	648.18	33.29	-0.03	0.92
	Accommodations	No	60094	651.27	31.66		0.91
		Yes	7300	617.22	37.31	1.05	0.92

Table 10. 6: Impact Analysis, Grade 5

Content Area	Category	Group	N	Mean	Std. Dev.	Effect Size	Coefficient Alpha
Communication Arts	Ethnicity	White (not Hispanic)	50145	677.96	33.51		0.90
		Black (not Hispanic)	11900	656.02	36.25	0.64	0.91
		Hispanic	2773	665.65	33.64	0.37	0.90
		Asian/Pacific Islander	1283	688.35	41.52	-0.31	0.91
		Native American/Alaskan	314	668.75	35.61	0.27	0.91
	Gender	Male	34060	670.16	36.32		0.91
		Female	32359	677.36	33.82	-0.20	0.90
	Accommodations	No	59414	678.34	31.25		0.89
		Yes	7088	634.24	42.38	1.35	0.91
Mathematics	Ethnicity	White (not Hispanic)	50151	673.38	39.47		0.90
		Black (not Hispanic)	11914	643.45	41.62	0.75	0.91
		Hispanic	2790	660.19	38.96	0.33	0.91
		Asian/Pacific Islander	1322	689.35	45.62	-0.40	0.91
		Native American/Alaskan	315	661.40	42.27	0.30	0.92
	Gender	Male	34089	667.91	43.30		0.92
		Female	32407	667.54	39.97	0.01	0.91
	Accommodations	No	59253	672.42	39.05		0.90
		Yes	7327	629.54	43.15	1.08	0.91
Science	Ethnicity	White (not Hispanic)	50146	670.93	28.81		0.88
		Black (not Hispanic)	11906	640.29	35.20	1.02	0.90
		Hispanic	2792	655.39	31.32	0.54	0.89
		Asian/Pacific Islander	1321	672.06	34.09	-0.04	0.91
		Native American/Alaskan	314	660.93	34.31	0.35	0.90
	Gender	Male	34086	665.60	33.86		0.91
		Female	32398	663.88	31.05	0.05	0.90
	Accommodations	No	59574	668.09	30.05		0.89
		Yes	6993	636.22	38.57	1.03	0.90

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)	50768	678.33	31.84		0.90
		Black (not Hispanic)	12070	657.58	32.54	0.65	0.90
		Hispanic	2746	665.79	30.64	0.39	0.90
		Asian/Pacific Islander	1259	685.67	38.20	-0.23	0.92
		Native American/Alaskan	338	672.48	32.24	0.18	0.89
	Gender	Male	34490	669.76	33.79		0.91
		Female	32698	678.88	31.68	-0.28	0.90
	Accommodations	No	60044	678.86	29.41		0.89
Yes		7216	635.25	36.52	1.44	0.89	
Mathematics	Ethnicity	White (not Hispanic)	50768	689.18	36.81		0.91
		Black (not Hispanic)	12074	659.50	40.48	0.79	0.92
		Hispanic	2766	673.22	36.63	0.43	0.92
		Asian/Pacific Islander	1290	702.65	45.05	-0.36	0.93
		Native American/Alaskan	338	677.21	36.50	0.33	0.91
	Gender	Male	34529	682.51	40.42		0.93
		Female	32714	684.32	38.40	-0.05	0.92
	Accommodations	No	59872	688.49	36.41		0.91
Yes		7444	642.12	39.05	1.26	0.91	

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)	50325	683.56	34.54		0.90
		Black (not Hispanic)	11571	659.11	35.84	0.70	0.90
		Hispanic	2515	670.58	33.38	0.38	0.89
		Asian/Pacific Islander	1228	690.74	44.04	-0.21	0.92
		Native American/Alaskan	326	677.10	38.99	0.19	0.91
	Gender	Male	33641	672.38	37.51		0.91
		Female	32328	685.62	33.54	-0.37	0.89
	Accommodations	No	59227	683.98	32.37		0.88
Yes		6807	634.21	37.50	1.51	0.88	
Mathematics	Ethnicity	White (not Hispanic)	50301	692.17	37.78		0.92
		Black (not Hispanic)	11566	662.09	40.08	0.79	0.91
		Hispanic	2531	676.43	38.17	0.42	0.91
		Asian/Pacific Islander	1263	707.13	48.41	-0.39	0.94
		Native American/Alaskan	325	683.87	42.26	0.22	0.92
	Gender	Male	33656	685.34	41.88		0.93
		Female	32332	687.78	38.46	-0.06	0.92
	Accommodations	No	59077	691.56	37.09		0.91
Yes		6976	643.69	40.84	1.28	0.88	

Table 10. 9: Impact Analysis, Grade 8

Content Area	Category	Group	N	Mean	Std. Dev.	Effect Size	Coefficient Alpha
Communication Arts	Ethnicity	White (not Hispanic)	50901	698.46	31.82		0.90
		Black (not Hispanic)	11291	676.10	36.20	0.68	0.91
		Hispanic	2276	686.29	32.70	0.38	0.90
		Asian/Pacific Islander	1223	705.83	41.19	-0.23	0.92
		Native American/Alaskan	344	690.57	35.33	0.25	0.91
	Gender	Male	33637	689.76	35.89		0.91
		Female	32405	699.04	31.17	-0.28	0.90
	Accommodations	No	59643	698.86	29.83		0.89
		Yes	6496	652.24	40.60	1.50	0.89
Mathematics	Ethnicity	White (not Hispanic)	50887	713.48	37.23		0.92
		Black (not Hispanic)	11275	683.24	41.41	0.80	0.91
		Hispanic	2301	698.05	37.63	0.41	0.91
		Asian/Pacific Islander	1255	729.44	48.01	-0.43	0.94
		Native American/Alaskan	345	702.07	38.81	0.31	0.92
	Gender	Male	33660	707.68	41.56		0.93
		Female	32409	708.37	38.38	-0.02	0.92
	Accommodations	No	59431	712.91	36.59		0.92
		Yes	6737	664.47	42.78	1.30	0.88
Science	Ethnicity	White (not Hispanic)	50881	703.69	28.07		0.91
		Black (not Hispanic)	11229	674.91	32.57	0.99	0.92
		Hispanic	2298	688.75	29.14	0.53	0.91
		Asian/Pacific Islander	1254	707.19	35.41	-0.12	0.94
		Native American/Alaskan	345	696.35	29.07	0.26	0.91
	Gender	Male	33615	698.93	32.73		0.93
		Female	32397	697.64	29.25	0.04	0.92
	Accommodations	No	59685	701.76	28.49		0.92
		Yes	6422	665.79	35.33	1.23	0.91

References

- Altman, D.G. (1991). *Practical statistics for medical research*. London: Chapman and Hall.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association, Inc.
- Assessment Resource Center (2008). *MAP and Missouri schools: A consequential validity study*. Columbia, MO: Author.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, 46, 443–459.
- Burket, G. R. (1995). PARDUX [Computer program]. Monterey, CA: CTB/McGraw-Hill.
- Camilli, G., & Shepard, A. L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publication.
- Candell, G.L., & Drasgow, F. (1988). An iterative procedure for linking metrics bias in item response theory. *Applied Psychological Measurement*, 12(3), 253–260.
- Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245–276.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- CTB/McGraw-Hill (2003). *TerraNova the 2nd edition: CAT technical report*. Monterey, CA: Author.
- CTB/McGraw-Hill (2005). *Missouri assessment program final bookmark standard setting technical report*. Monterey, CA: Author.

- CTB/McGraw-Hill (2008). *Missouri assessment program bookmark standard setting technical report 2008 for Missouri achievement-level setting grades 5, 8, and 11 science*. Monterey, CA: Author.
- CTB/McGraw-Hill (2009). *TerraNova 3rd edition technical addendum: Forms E and F*. Monterey, CA: Author.
- CTB/McGraw-Hill (2010). *Guide to interpreting results*. Monterey, CA: Author.
- Dorans, N.J., & Schmitt, M.P. (1991). *Constructed response and differential item functioning: A pragmatic approach*. Princeton: Educational Testing Service.
- Green, D.R. (1975). Procedures for assessing bias in achievement tests. Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Hingham, MA: Kluwer-Nijhoff Publishing.
- Karkee, T., & Choi, S. (2005). Impact of eliminating anchor items flagged from statistical criteria on test score classifications in common item equating. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada, April 15, 2005.
- Kim, D. (2005). KKCLASS [Computer program]. Unpublished.
- Kim, D., Barton, K., & Kim, J. (2007). Estimating classification consistency and classification accuracy with pattern scoring. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Kim, D., Choi, S., Um, K., & Kim, J. (2006). A comparison of methods for estimating classification consistency. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, CA.
- Kolen, M. J., & Brennan, R.L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.
- Kolen, M. J., & Kim, D. (2005). Personal correspondence.
- Landis, J.R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179–197.

- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Macmillan/McGraw-Hill (1993a). *Guidelines for bias-free publishing*. New York, NY: Author.
- Macmillan/McGraw-Hill (1993b). *Reflecting diversity: Multicultural guidelines for educational publishing professionals*. New York, NY: Author.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–748.
- Michaelides, M.P., & Haertel, E.H. (2004). Sampling of common items: An unrecognized source of error in test equating. Los Angeles, CA: Center for the Study of Evaluation.
- Schumacker, R.E. (1996). Disattenuating correlation coefficients. *Rasch Measurement Transactions, 10*, 479.
- Swaminathan, H., Hambleton, R.K., & Algina, J. (1974). Reliability of criterion-reference tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11*(4), 263–267.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.
- Thissen, D. (1982). Marginal maximum-likelihood estimation for the one-parameter logistic model. *Psychometrika, 47*, 175–186.
- Thompson, S., & Thurlow, M. (2002). Universally designed assessments: Better tests for everyone! (Policy Directions. No. 14). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [November 9, 2009], from: <http://www.cehd.umn.edu/NCEO/OnlinePUBs/Policy14.htm>
- Voelkle, M., Schwarz, R., Arenson, E., & Ito, K. (2002). An investigation of factors affecting Stocking & Lord equating. (Paper in progress.)
- Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245–262.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

- Yen, W.M., & Candell, G.L. (1991). Increasing score reliability with item-pattern scoring: An empirical study in five score metrics. *Applied Measurement in Education, 4*, 209–228.
- Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233–251.