Massachusetts Department of
ELEMENTARY & SECONDARY
EDUCATION

2012 MCAS and MCAS-Alt Technical Report

This document was prepared by the
Massachusetts Department of Elementary and Secondary Education
Mitchell D. Chester, Ed.D.
Commissioner

The Massachusetts Department of Elementary and Secondary Education, an affirmative action employer, is committed to ensuring that all of its programs and facilities are accessible to all members of the public. We do not discriminate on the basis of age, color, disability, gender identity, national origin, race, religion, sex, or sexual orientation. Inquiries regarding the Department's compliance with Title IX and other civil rights laws may be directed to the Human Resources Director, 75 Pleasant St., Malden, MA 02148, 781-338-6105.

Massachusetts Department of Elementary and Secondary Education
75 Pleasant Street, Malden, MA 02148-4906
Phone 781-338-3000 TTY: N.E.T. Relay 800-439-2370
www.doe.mass.edu

TABLE OF CONTENTS

Chapter 1. Overview

1.1 Purposes of the MCAS

The Massachusetts Education Reform Mandate

The Massachusetts Comprehensive Assessment System (MCAS) is the Commonwealth's program for student assessment developed in accordance with the Massachusetts Education Reform Act of 1993. The Massachusetts Education Reform Act specifies that the testing program must

- test all students who are educated with Massachusetts public funds, including students with disabilities and English language learner (ELL) students;
- measure performance based on the Massachusetts curriculum frameworks learning standards (the current Massachusetts curriculum frameworks are posted on the Massachusetts Department of Elementary and Secondary Education (ESE) website at www.doe.mass.edu/frameworks/current.html);
- report on the performance of individual students, schools, districts, and the state.

The Massachusetts Education Reform Act also stipulates that students earn a Competency Determination (CD) by passing grade 10 tests in English language Arts (ELA), mathematics, and science and technology/engineering (STE) as one condition of eligibility for a Massachusetts high school diploma.

To fulfill the requirements of the Massachusetts Education Reform Act, the MCAS is designed to

- measure student, school, and district performance in meeting the state's learning standards as detailed in the Massachusetts curriculum frameworks;
- provide measures of student achievement that will lead to improvements in student outcomes;
- help determine ELA, mathematics, and STE competency for the awarding of high school diplomas.

Additionally, MCAS results are used to fulfill federal requirements by contributing to school and district accountability determinations.

1.2 Purpose of This Report

The purpose of this *2012 MCAS and MCAS-Alt Technical Report* is to document the technical quality and characteristics of the 2012 MCAS operational tests, to present evidence of the validity and reliability of test score interpretations, and to describe modifications made to the program in 2012. Technical reports for 1998 to 2011 are available on the ESE website at www.doe.mass.edu/mcas/tech/?section=techreports. The *2012 MCAS and MCAS-Alt Technical Report* is designed to supplement the technical reports issued for previous MCAS administrations by providing information specific to the 2012 MCAS test administrations. Previous technical reports, as well as other documents referenced in this report, provide additional background information about the MCAS program and its development and administration.

This report is primarily intended for experts in psychometrics and educational measurement. It assumes a working knowledge of measurement concepts, such as reliability and validity, as well as statistical concepts of correlation and central tendency. For some sections, the reader is presumed to have basic familiarity with advanced topics in measurement and statistics, such as item response theory (IRT) and factor analysis.

1.3 Organization of This Report

This report provides detailed information regarding test design and development, scoring, and analysis and reporting of 2012 MCAS results at the student, school, district, and state levels. This detailed information includes, but is not limited to, the following:

- An explanation of test administration
- An explanation of equating and scaling of tests
- Statistical and psychometric summaries:
    a. Item analyses
    b. Reliability evidence
    c. Validity evidence

In addition, the technical appendices contain detailed item-level and summary statistics related to each 2012 MCAS test and its results.

Chapter 1 of this report provides a brief overview of what is documented within the report, including updates made to the MCAS program during 2012. Chapter 2 explains the guiding philosophy, purpose, uses, components, and validity of the state's assessment system. The next two chapters cover the test design and development, test administration, scoring, and analysis and reporting of results for the standard MCAS assessment (Chapter 3) and the MCAS Alternate Assessment (Chapter 4). These two chapters include information about the characteristics of the test items, how scores were calculated, the reliability of the scores, how scores were reported, and the validity of the results. Numerous appendices, which appear after Chapter 4, are referenced throughout the report.

1.4 Current Year Updates

In addition to changes detailed throughout this document, the following changes were made for the 2012 MCAS administrations.

1.4.1 Updated information about MCAS Test Participation Requirements

Complete, updated student participation requirements for all spring 2012 MCAS tests can be found in the *Spring 2012 Principal's Administration Manual*. The updates to this document included the requirement that ELL students participate in the Massachusetts English Proficiency Assessment (MEPA) to qualify as full participants in the spring 2012 ELA tests. Student participation requirements for the November 2011 ELA and Mathematics retests, February 2012 Biology test, and March 2012 ELA and Mathematics retests can be found in the *Fall 2011/Winter 2012 Principal's Administration Manual*. For a copy of either document, please call Student Assessment Services at 781-338-3625.

1.4.2 Transition to the *2011 Massachusetts ELA and Mathematics Curriculum Frameworks*

On December 21, 2010, the Board of Elementary and Secondary Education and the Board of Early Education and Care adopted the new *2011 Massachusetts Curriculum Frameworks for English Language Arts and Literacy* and the new *2011 Massachusetts Curriculum Framework for Mathematics*, both based on the *Common Core State Standards*. These new frameworks are posted at www.doe.mass.edu/candi/commoncore/ along with side-by-side comparisons of the new and the previous standards. The new frameworks have also been posted, along with the other current frameworks, on the ESE's website at www.doe.mass.edu/frameworks/current.html. Documents summarizing the ESE's plan for transitioning MCAS to the 2011 frameworks are posted at www.doe.mass.edu/mcas/transition/.

Chapter 2. The State Assessment System

2.1 Introduction

MCAS is designed to meet the requirements of the Massachusetts Education Reform Act of 1993. This law specifies that the testing program must

- test all public school students in Massachusetts, including students with disabilities and English language learner (ELL) students;
- measure performance based on the Massachusetts curriculum frameworks learning standards;
- report on the performance of individual students, schools, and districts.

As required by The Massachusetts Education Reform Act, students must pass the grade 10 tests in ELA, mathematics, and STE as one condition of eligibility for a high school diploma (in addition to fulfilling local requirements).

2.2 Guiding Philosophy

The MCAS and MCAS Alternate Assessment (MCAS-Alt) programs play a central role in helping all stakeholders in the Commonwealth's education system—students, parents, teachers, administrators, policy leaders, and the public—understand the successes and challenges in preparing students for higher education, work, and engaged citizenship.

In the decade since the first administration of the MCAS tests, the ESE has gathered evidence from many sources suggesting that the assessment reforms introduced in response to the Massachusetts Education Reform Act of 1993 have been an important factor in raising the academic expectations of all students in the Commonwealth and in making the educational system in Massachusetts one of the country's best.

The MCAS testing program has been an important component of education reform in Massachusetts for over a decade. The program continues to evolve, with recent and current improvements that

- respond to stakeholders' interests;
- reflect the vision and goals outlined by the governor's Readiness Project;
- respond to the Board of Elementary and Secondary Education's 21st Century Skills Task Force by developing an assessment system that is viewed by teachers as integral to their daily instructional activities;
- ensure that the MCAS measures the knowledge and skills students need to meet the challenges of the 21st century.

2.3 Purpose of the State Assessment System

The MCAS is a custom-designed program owned in its entirety by the Commonwealth of Massachusetts. All items included on the MCAS tests are written to measure standards contained in the Massachusetts curriculum frameworks. Equally important, virtually all standards contained in the curriculum frameworks are measured by items on the MCAS tests.[1] All MCAS tests are designed to measure MCAS performance levels based on performance-level descriptors derived from the Massachusetts curriculum frameworks. Therefore, the primary inferences drawn from the MCAS test results are about the level of students' mastery of the standards contained in the Massachusetts curriculum frameworks.

2.4 Uses of the State Assessment System

MCAS results are used for a variety of purposes. Official uses of MCAS results include the following:

- determining school and district progress toward the goals set by the state and federal accountability systems
- determining whether high school students have demonstrated the knowledge and skills required to earn a Competency Determination (CD)—one requirement for earning a high school diploma in Massachusetts
- providing information to support program evaluation at the school and district levels
- helping to determine the recipients of scholarships, including the John and Abigail Adams Scholarship
- providing diagnostic information to help all students reach higher levels of performance

2.5 Validity of the State Assessment System

Validity information for the state assessment system is provided throughout this technical report. Validity evidence includes information on test design and development; administration; scoring; technical evidence of test quality (classical item statistics, differential item functioning, item response theory statistics, reliability, dimensionality, decision accuracy and consistency); and reporting. Information is described in detail in sections of this report and summarized for each of the assessment components in their respective Validity subsections (Section 3.9 for MCAS and 4.9 for MCAS-Alt).

[1] A small number of standards in the current curriculum frameworks have been classified as not appropriate for large-scale paper-and-pencil assessments such as the MCAS tests. Examples of those standards from the 2004 ELA framework include Language Standard 3, which requires students to make oral presentations, and Composition Standard 24, which requires students to conduct a research project. Standards such as those are to be assessed at the local level. See http://www.doe.mass.edu/frameworks/current.html for information about scheduled updates to the curriculum frameworks.

Chapter 3. MCAS

3.1 Overview

MCAS tests have been administered to students in Massachusetts since 1998. In 1998, ELA, mathematics, and STE were assessed at grades 4, 8, and 10. In subsequent years, additional grades and content areas were added to the testing program. Following the initial administration of each new test, performance standards were set.

Public school students in the graduating class of 2003 were the first students required to earn a CD in ELA and mathematics as a condition for receiving a high school diploma. To fulfill the requirements of the NCLB Act, tests for several new grades and content areas were added to the MCAS in 2006. As a result, all students in grades 3–8 and 10 are assessed in both ELA and mathematics.

The program is managed by ESE staff with assistance and support from the assessment contractor, Measured Progress (MP). Massachusetts educators play a key role in the MCAS through service on a variety of committees related to the development of MCAS test items, the development of MCAS performance-level descriptors, and the setting of performance standards. The program is supported by a five-member national Technical Advisory Committee as well as measurement specialists from the University of Massachusetts–Amherst.

More information about the MCAS program is available at www.doe.mass.edu/mcas.

3.2 Test Design and Development

The 2012 MCAS test administration included operational tests in the following grades and content areas:

- grades 3–8 and grade 10 ELA, including a composition component at grades 4, 7, and 10
- grades 3–8 and grade 10 Mathematics
- grades 5 and 8 STE
- high school STE end-of-course tests in Biology, Chemistry, Introductory Physics, and Technology/Engineering

The 2012 MCAS administration also included retest opportunities in ELA and mathematics in November 2011 and March 2012 for students beyond grade 10 who had not yet passed the standard grade 10 test. A February Biology test was also administered.

3.2.1 Test Specifications

3.2.1.1 Criterion-Referenced Test

Items used on the MCAS are developed specifically for Massachusetts and are directly linked to Massachusetts content standards. These content standards are the basis for the reporting categories developed for each content area and are used to help guide the development of test items. No content or process other than those described in the Massachusetts curriculum frameworks is subject to statewide assessment. An item, depending on its type, may address one, all, or several of the indicators within a standard. In 2011, Massachusetts adopted new curriculum standards in mathematics and ELA. In 2012, all common test items for ELA and mathematics were coded, when possible, to both the new standards and the previous standards.

3.2.1.2 Item Types

Massachusetts educators and students are familiar with the types of items used in the assessment program. The types of items and their functions are described below.

- **Multiple-choice** items are used to provide breadth of coverage within a content area. Because they require no more than a minute for most students to answer, multiple-choice items make efficient use of limited testing time and allow for coverage of a wide range of knowledge and skills. Multiple-choice items appear on every MCAS test except the ELA composition. Each multiple-choice item requires that students select the single best answer from four response options. Multiple-choice items are aligned to one primary standard. They are machine-scored; correct responses are worth one score point, and incorrect and blank responses are assigned zero score points. The blanks are reported separately from the incorrect responses.
- **One-point short-answer** mathematics items are used to assess students' skills and abilities to work with brief, well-structured problems that have one or a very limited number of solutions (e.g., mathematical computations). Short-answer items require approximately two minutes for most students to answer. The advantage of this type of item is that it requires students to demonstrate knowledge and skills by generating, rather than selecting, an answer. One-point short-answer items are hand-scored and assigned one point (correct) or zero points (blank or incorrect). The blanks are distinguished from the incorrect responses.
- **Two-point open-response** items are used in the grade 3 Mathematics test. Students are expected to generate one or two sentences of text in response to a word problem. The student responses are hand-scored with a range of score points from zero to two. Two-point responses are totally correct, one-point responses are partially correct, and responses with a score of zero are completely incorrect. Blank responses receive a score of zero. The blanks are distinguished from the incorrect responses.
- **Two-point short-response** items are used in the grade 3 ELA Reading Comprehension test. Students are expected to generate one or two sentences of text in response to a passage-driven prompt. The student responses are hand-scored with a range of score points from zero to two. Two-point responses are totally correct, one-point responses are partially correct, and responses with a score of zero are completely incorrect. Blank responses receive a score of zero. The blanks are distinguished from the incorrect responses.
- **Four-point open-response** items typically require students to use higher-order thinking skills—such as evaluation, analysis, and summarization—to construct satisfactory responses. Open-response items take most students approximately 5 to 10 minutes to complete. Open-response items are hand-scored by readers trained in the specific requirements of each question scored. Students may receive up to four points per open-response item. Totally incorrect or blank responses receive a score of zero. The blanks are distinguished from the incorrect responses.
- **Writing prompts** are administered to all students in grades 4, 7, and 10 as part of the ELA test. The assessment consists of two sessions separated by a ten-minute break. During the first session, students write a draft composition. In the second session, students write a final composition based on that draft. Each composition is hand-scored by trained scorers. Students receive two scores: one for topic development (0 to 6 points), and the other for

standard English conventions (0 to 4 points). Student reports include a score for each of these dimensions. Each student composition is scored by two different scorers; the final score is a combination of both sets of scores, so students may receive up to 20 points for their compositions. These 20 composition points amount to 28% of a student's overall ELA score.

3.2.1.3 Description of Test Design

The MCAS is structured using both common and matrix items. Common items are administered to all students in a given grade level. Student scores are based only on common items. Matrix items are either new items included on the test for field-test purposes or equating items used to link one year's results to those of previous years. In addition, field-test and equating items are divided among the multiple forms of the test for each grade and content area. The number of test forms varies by grade and content area but ranges between 5 and 32 forms. Each student takes only one form of the test and therefore answers a subset of the field-test and equating items. Equating and field-test items are not distinguishable to test takers. Because all students participate in the field test, an adequate sample size (approximately 1,800 students per item) is provided to produce reliable data that can be used to inform item selection for future tests.

3.2.2 ELA Test Specifications

3.2.2.1 Standards

The reading comprehension portion of the ELA test measured the following learning standards from the *2001 Massachusetts English Language Arts Curriculum Framework* and the *2004 Supplement to the Massachusetts English Language Arts Curriculum Framework*:

- Language Strand
    - Standard 4: Vocabulary and Concept Development
    - Standard 5: Structure and Origins of Modern English
    - Standard 6: Formal and Informal English
- Reading and Literature Strand
    - Standard 8: Understanding a Text
    - Standard 9: Making Connections
    - Standard 10: Genre
    - Standard 11: Theme
    - Standard 12: Fiction
    - Standard 13: Nonfiction
    - Standard 14: Poetry
    - Standard 15: Style and Language
    - Standard 16: Myth, Traditional Narrative, and Classical Literature
    - Standard 17: Dramatic Literature

The composition portion of the ELA test measured the following learning standards from the *Massachusetts English Language Arts Curriculum Framework*:

- Composition Strand
    - Standard 19: Writing
    - Standard 20: Consideration of Audience and Purpose
    - Standard 21: Revising
    - Standard 22: Standard English Conventions
    - Standard 23: Organizing Ideas in Writing

The following standards cannot be assessed on a large-scale paper-and-pencil test and are to be locally assessed:

- Language Strand
    - Standard 1: Discussion
    - Standard 2: Questioning, Listening, and Contributing
    - Standard 3: Oral Presentation
- Reading and Literature Strand
    - Standard 7: Beginning Reading
    - Standard 18: Dramatic Reading and Performance
- Composition Strand
    - Standard 24: Research
    - Standard 25: Evaluating Writing and Presentations
- Media Strand
    - Standard 26: Analysis of Media
    - Standard 27: Media Production

For grade-level articulation of these standards, please refer to the *Massachusetts English Language Arts Curriculum Framework*. 2012 is the last year that the MCAS ELA assessment will be aligned to the 2001/2004 standards. The 2013 test will be aligned to standards found in the *2011 Massachusetts Curriculum Framework for English Language Arts and Literacy*. Items field-tested in the 2012 assessment for future use are aligned to the 2011 Massachusetts ELA standards. The November 2012 and March 2013 English Language Arts retests for the high school CD will be aligned to the 2001/2004 Massachusetts ELA standards.

3.2.2.2 Item Types

The reading comprehension portion of the ELA tests includes a mix of multiple-choice and open-response items. Two-point short-response items are included in the grade 3 test only. A writing prompt is administered to students in grades 4, 7, and 10. Each type of item is worth a specific number of points in a student's total score. Table 3-1 indicates the possible number of raw score points for each item type.

Table 3-1. 2012 MCAS: ELA Item Types and Score Points

Item Type	*Possible Raw Score Points*
Multiple Choice	0 or 1
Short Response	0, 1, or 2
Open-Response	0, 1, 2, 3, or 4
Writing Prompt	0 to 20

3.2.2.3 Test Design

In 2010, as part of an effort to reduce testing time, the ELA reading comprehension tests in grades 3–8 were shortened by eliminating one session, going from three sessions to two sessions. The 2011 MCAS administration used this shorter test design and the 2012 administration continued with the two-session design. Table 3-2 shows the current design.

Table 3-2. 2012 MCAS: ELA Reading Comprehension Test Designs

Grade	*# of Sessions*	*Minutes per Session*	*Common Points*	*Matrix Points*
3	2	60	48	14
4–8	2	60	52	14

Grade 3 ELA Reading Comprehension Test

The common portion of this test includes two long passages and three short passages. Each long passage is typically accompanied by 10 multiple-choice items and either one 4-point open-response item or two 2-point short-response items. Each short passage is accompanied by five or six multiple-choice items and one or no short-response items. The grade 3 reading comprehension test contains a total of 48 common points and 14 matrix points distributed across two testing sessions.

Grades 4–8 ELA Reading Comprehension Tests

The common portion of each of these tests includes two long passages and three short passages. Each long passage is typically accompanied by 10 multiple-choice items and one 4-point open-response item. A total of 16 multiple-choice items and two 4-point open-response items accompany three short passages. The grades 4–8 reading comprehension tests contain 52 common points and 14 matrix points per form distributed across two testing sessions.

Grade 10 ELA Reading Comprehension Test

The common portion of the grade 10 reading comprehension test consists of three long passages and three short passages with a total of 52 common points. Each long passage is accompanied by eight multiple-choice items and one 4-point open-response item. The three short passages are accompanied by a total of 12 multiple-choice items and one 4-point open-response item. The grade 10 reading comprehension test is divided into three testing sessions.

ELA Composition

Students in grades 4, 7, and 10 must also complete the composition portion of the MCAS. The composition portion of the ELA test consists of one writing prompt with a total value of 20 points (12 points for topic development and 8 points for standard English conventions). At the three grades tested, the composition score accounts for 28% of a student’s total raw score for ELA.

The ELA composition tests have historically assessed narrative writing at grade 4, expository writing at grade 7, and literary analysis at grade 10. In 2013 and beyond, the composition assessments for grades 4 and 7 may include any modes listed in the 2011 frameworks. In 2013, grade 10 composition will continue to be assessed through literary analysis.

ELA Retests

Retests were offered to students beyond grade 10 who had not yet met the ELA requirement for earning a CD by passing the grade 10 ELA test. Retests were available to students in their junior and senior years in November and March. The reading comprehension portion of the retests consists of common items only.

Table 3-3. 2012 MCAS: Distribution of ELA Common and Matrix Items by Grade and Item Type

Grade and Test			*Items per Form*								*Total Matrix Items Across Forms*							
Grade	*Test*	*# of Forms*	*Common*				*Matrix*				*Equating Positions*				*Field-Test Positions*			
			MC	*SR*	*OR*	*WP*	*MC*	*SR*	*OR*	*WP*	*MC*	*SR*	*OR*	*WP*	*MC*	*SR*	*OR*	*WP*
3	Reading Comprehension	15	36	4	1		10	2[a]	1[a]		30	6[a]	3[a]		120	24	12	
4	Reading Comprehension	15	36		4		10		1		30		3		120		12	
	Composition	2[b]				1												
5	Reading Comprehension	15	36		4		10		1		30		3		120		12	
6	Reading Comprehension	15	36		4		10		1		30		3		120		12	
7	Reading Comprehension	15	36		4		10		1		30		3		120		12	
	Composition	2[b]				1												
8	Reading Comprehension	15	36		4		10		1		30		3		120		12	
10	Reading Comprehension	24	36		4		12		2		[c]		[c]		288		48	
	Composition	2[b]				1												
Retest[d]	Reading Comprehension	1	36		4													
	Composition	1				1												
	Reading Comprehension	1	36		4													
	Composition	1				1												

[a] The grade 3 matrix form has space for either one 4-point OR or two 2-point SR items.
[b] The ELA composition is field-tested out of state.
[c] The grade 10 ELA test is pre-equated; therefore, the entire set of matrix positions is available for field-testing.
[d] ELA retests consist of common items only.

3.2.2.4 Blueprints

Table 3-4 shows the test specifications—the percentage of common item points aligned to the Massachusetts ELA curriculum framework strands—for the MCAS 2012 ELA tests.

Table 3-4. 2012 MCAS: Distribution of ELA Item Points across Strands by Grade

Framework Strand	*Percent of Raw Score Points at Each Grade*						
	3	*4*	*5*	*6*	*7*	*8*	*10*
Language	15	8	12	12	8	12	8
Reading and Literature	85	64	88	88	64	88	64
Composition		28			28		28
Total	100	100	100	100	100	100	100

3.2.2.5 Cognitive Levels

Each item on the ELA test is assigned a cognitive level according to the cognitive demand of the item. Cognitive levels are not synonymous with item difficulty. The cognitive level rates each item based on the complexity of the mental processing a student must use to answer the item correctly. Each of the three cognitive levels used in ELA is described below.

- Level I (Identify/Recall) – Level I items require that the test taker recognize basic information presented in the text.
- Level II (Infer/Analyze) – Level II items require that the test taker understand a given text by making inferences and drawing conclusions related to the text.
- Level III (Evaluate/Apply) – Level III items require that the test taker understand multiple points of view and be able to project his or her own judgments or perspectives on the text.

Each cognitive level is represented in the reading comprehension portion of the ELA test.

3.2.2.6 Reference Materials

At least one English-language dictionary per classroom was provided for student use during ELA composition tests. The use of bilingual word-to-word dictionaries was allowed for current and former English language learner (ELL) students only, during both the ELA composition and ELA reading comprehension tests. No other reference materials were allowed during the ELA composition or ELA reading comprehension tests.

3.2.2.7 Passage Types

The reading comprehension tests include both long and short passages. Long passages range in length from approximately 1,000 to 1,500 words; short passages are generally under 1,000 words. Word counts are slightly reduced at lower grades. Dramas, myths, fables, and folktales are treated as short passages regardless of length.

Passages were selected from published works; no passages were specifically written for the ELA tests. Passages are categorized into one of two types:

- Literary passages – Literary passages represent a variety of genres: poetry, drama, fiction, biographies, memoirs, folktales, fairy tales, myths, legends, narratives, diaries, journal entries, speeches, and essays. Literary passages are not necessarily fictional.
- Informational passages – Informational passages are reference materials, editorials, encyclopedia articles, and general nonfiction. Informational passages are drawn from sources such as magazines, newspapers, and books.

In grades 3–8, the common form of the ELA test includes one long and two short literary passages and one long and one short informational passage. In grade 10, the common form includes one long and three short literary passages and two long informational passages.

The reading comprehension portion of the MCAS ELA test is designed to include a set of passages with a balanced representation of male and female characters; races and ethnicities; and urban, suburban, and rural settings. It is important that passages be of interest to the age group being tested. Approximately 50% of the passages are written by authors found in Appendices A and B of the *English Language Arts Curriculum Framework*.

The main difference among the passages used for grades 3–8 and 10 is their degree of complexity, which results from increasing levels of sophistication in language and concepts, as well as passage length. MP uses a variety of readability formulas to aid in the selection of passages appropriate for the intended audience. In addition, Massachusetts teachers use their grade-level expertise when participating in the selection of passages as members of the Assessment Development Committees (ADCs).

Items based on ELA reading passages require students to demonstrate skills in both literal comprehension, in which the answer is stated explicitly in the text, and inferential comprehension, in which the answer is implied by the text or the text must be connected to relevant prior knowledge to determine an answer. Items focus on the reading skills reflected in the content standards and require students to use reading skills and strategies to answer correctly.

Items coded to the language standards use the passage as a stimulus for the items. There are no stand-alone multiple-choice, short-response, or open-response items on the MCAS ELA assessments. All vocabulary, grammar, and mechanics questions on the MCAS are derived from a passage. The writing prompt is not associated with a specific reading passage.

3.2.3 Mathematics Test Specifications

3.2.3.1 Standards

The MCAS Mathematics tests at grades 3–8 and 10 measured the learning standards of the five strands of the *2000 Massachusetts Mathematics Curriculum Framework* and the *2004 Supplement to the Massachusetts Mathematics Curriculum Framework*:

- Number Sense and Operations
- Patterns, Relations, and Algebra
- Geometry
- Measurement
- Data Analysis, Statistics, and Probability

Massachusetts is transitioning to the 2011 Curriculum Frameworks for mathematics. While the 2012 tests were aligned to the 2000/2004 mathematics standards, the focus of the tests was the 2000/2004 standards that were connected to the 2011 standards. By the 2014 test administration, the MCAS mathematics assessments will be completely assessing the 2011 Massachusetts Curriculum Frameworks for mathematics. All new development, including those items field-tested in the 2012 matrix positions, aligns to the 2011 standards.

3.2.3.2 Item Types

The mathematics tests include multiple-choice, short-answer, and open-response items. Short-answer items require students to perform a computation or solve a simple problem. Open-response items are more complex, requiring 5–10 minutes of response time. Each type of item is worth a specific number of points in the student's total mathematics score, as shown in Table 3-5.

Table 3-5. 2012 MCAS: Mathematics Item Types and Score Points

Item Type	*Possible Raw Score Points*
MC	0 or 1
SA	0 or 1
2-point OR*	0, 1, or 2
OR	0, 1, 2, 3, or 4

* Only grade 3 mathematics uses 2-point OR items.

3.2.3.3 Test Design

In 2010, as part of an effort to reduce testing time, the mathematics tests in grades 3–8 were shortened by eliminating some of the matrix slots. The 2011 test used the 2010 design, and the 2012 test continues to use the 2010 test design.

Table 3-6. 2012 MCAS: Mathematics Test Designs

Grade	*# of Sessions*	*Minutes per Session*	*Common Points*	*Matrix Points*
3	2	45	40	7
4–6	2	45	54	7
7–8	2	50	54	12

The tests are composed of common and matrix items. The matrix slots in each test form are used to field-test potential items or to equate the current year's test to that of previous years by using previously administered items. Table 3-7 shows the distribution of these items on the mathematics tests.

Table 3-7. 2012 MCAS: Distribution of Mathematics Common and Matrix Items by Grade and Item Type

Grade	*# of Forms*	*Items per Form*						*Total Matrix Items Across Forms*											
		Common			*Matrix*			*Total Slots*			*Equating Slots*			*Field-Test Slots (available)*			*Unique FT Items*[a]		
		MC	*SA*	*OR*	*MC*	*SA*	*OR*	*MC*	*SA*	*OR*	*MC*	*SA*[a]	*OR*[b]	*MC*	*SA*	*OR*	*MC*	*SA*	*OR*
3	18	26	6	4[b]	2	1	1[b]	36	18	18[b]	13	3	2[b]	23	15	16[b]	23	9	6[b]
4	22	32	6	4	2	1	1	44	22	22	15	3	2	29	19	20	29	9	6
5	21	32	6	4	2	1	1	42	21	21	16	3	2	26	18	19	26	9	6
6	21	32	6	4	2	1	1	42	21	21	16	3	2	26	18	19	26	9	6
7	21	32	6	6	2	2	2	42	42	42	16	3	2	26	39	40	26	15	6
8	21	32	6	6	2	2	2	42	42	42	16	3	2	26	39	40	26	15	6
10	32	32	4	6	7	1	2	224	32	64	[c]	[c]	[c]	161	23	46	212	23	30
Retest[d]	1	32	4	6				36	18	18							23	9	6
	1	32	4	6				42	21	21							26	9	6

[a] The numbers represented in the field-test positions are unique field-test items. There are more field-test slots than unique items, so items are repeated. Therefore, at grade 4, there were actually 22 SA slots and 22 OR slots, while 9 unique SA items were assessed and 6 unique OR items were assessed.

[b] OR items at grade 3 are worth 2 points.

[c] The grade 10 test is pre-equated; therefore, the entire set of matrix slots is available for field-testing.

[d] Mathematics retests consist of common items only.

3.2.3.4 Blueprints

Table 3-8 shows the test specifications—the distribution of common item points across the Massachusetts Mathematics curriculum framework strands—for the 2012 MCAS Mathematics tests.

Table 3-8. 2012 MCAS: Mathematics Common Point Distribution by Strand and Grade

Framework Strand	*Percent of Raw Score Points at Each Grade*						
	3	*4*	*5*	*6*	*7*	*8*	*10*
Number Sense and Operations	40	40	41	35	30	26	20
Patterns, Relations, and Algebra	20	15	26	30	30	35	30
Geometry	10	15	9	9	15	15	15
Measurement	15	15	15	15	10	9	15
Data Analysis, Statistics, and Probability	15	15	9	11	15	15	20
Total	100	100	100	100	100	100	100

3.2.3.5 Cognitive Levels

Each item on the Mathematics test is assigned a cognitive level according to the cognitive demand of the item. Cognitive levels are not synonymous with difficulty. The cognitive level rates each item based on the complexity of the mental processing a student must use to answer the item correctly. Each of the three cognitive levels used in the Mathematics tests is listed and described below.

- Level I (Recall and Recognition) – Level I items require students to recall mathematical definitions, notations, simple concepts, and procedures, as well as to apply common, routine procedures or algorithms (that may involve multiple steps) to solve a well-defined problem.
- Level II (Analysis and Interpretation) – Level II items require students to engage in mathematical reasoning beyond simple recall, in a more flexible thought process, and in enhanced organization of thinking skills. The items demand that students make a decision about the approach needed, represent or model a situation, or use one or more nonroutine procedures to solve a well-defined problem.
- Level III (Judgment and Synthesis) – Level III items require students to perform more abstract reasoning, planning, and evidence-gathering. In order to answer these types of questions, students must engage in reasoning about an open-ended situation with multiple decision points to represent or model unfamiliar mathematical situations and solve more complex, nonroutine, or less well-defined problems.

Cognitive levels I and II are represented by items in all grades. Level III is best represented by OR items. An attempt is made to include cognitive level III items at each grade.

3.2.3.6 Use of Calculators, Reference Sheets, Tool Kits, and Rulers

The second session of the grade 7, 8, and 10 Mathematics tests is a calculator session. All items included in this session are either calculator neutral (calculators are permitted but not required to answer the question) or calculator active (students are expected to use a calculator to answer the question). Each student taking the grade 7, 8, or 10 Mathematics test had access to a calculator with at least four functions and a square root key.

Reference sheets are provided to students at grades 5–8 and 10. These sheets contain information, such as formulas, that students may need to answer certain items. The reference sheets are published each year with the released items and have remained the same for several years over the various test administrations.

Tool kits are provided to students at grades 3 and 4. The tool kits contain manipulatives designed for use when answering specific questions. Because the tool kits are designed for specific items, they change annually. The parts of the tool kits used to answer common questions are published with the released items.

Rulers are provided to students in grades 3–8.

3.2.4 STE Test Specifications

3.2.4.1 Standards

Grades 5 and 8

The STE tests at grades 5 and 8 measured the learning standards of the four strands of the *2006 Massachusetts Science and Technology/Engineering Curriculum Framework*:

- Earth and Space Science
- Life Science
- Physical Sciences
- Technology/Engineering

High School

Each of the four end-of-course high school STE tests focuses on one subject (biology, chemistry, introductory physics, or technology/engineering). Students in grade 9 who are enrolled in a course that corresponds to one of the tests are eligible but not required to take the test in the course they studied. All students are required to take one of the four tests by the time they complete grade 10. Grade 10 students who took an STE test in grade 9 but did not pass are required to take an STE test again. If a student is enrolled in or has completed more than one STE course, he or she may select which STE test to take (with consultation from parents/guardians and school personnel). Any grade 11 or grade 12 student who has not yet earned a CD in STE is eligible to take any of the four STE tests. Testing opportunities are provided in February (Biology only) and June (Biology, Chemistry, Introductory Physics, and Technology/Engineering).

The high school STE tests measure the learning standards of the strands listed in Tables 3-12 through 3-15.

3.2.4.2 Item Types

The STE tests include multiple-choice and open-response items. Open-response items are more complex, requiring 5–10 minutes of response time. Each type of item is worth a specific number of points in the student’s total test score, as shown in Table 3-9.

Table 3-9. 2012 MCAS: STE Item Types and Score Points

Item Type	*Possible Raw Score Points*
MC	0 or 1
OR	0, 1, 2, 3, or 4

The high school Biology test includes one common module per test. A module is composed of a stimulus (e.g., a graphic or a written scenario) and a group of associated items (four multiple-choice items and one open-response item).

3.2.4.3 Test Design

The STE tests are composed of common and matrix items. Each form includes the full complement of common items, which are taken by all students, and a set of matrix items. Table 3-10 shows the number of unique items field-tested. Often, there are fewer unique items than field-test positions. When this happens, field-test items are repeated across two or more forms.

Table 3-10. 2012 MCAS: Distribution of STE Common and Matrix Items by Grade and Item Type

Grade	*Test*	*# of Forms*	*Items per Form*				*Total Matrix Items Across Forms*			
			Common		*Matrix*		*Equating Positions*		*Field-Test Positions*	
			MC	*OR*	*MC*	*OR*	*MC*	*OR*	*MC*	*OR*
5	STE	22	38	4	3	1	19	2	47	12
8	STE	22	38	4	3	1	19	2	47	12
HS	Biology	15	40[a]	5[a]	12[b]	2[b]	NA[c]	NA[c]	180	30
	Chemistry	5	40	5	20	2	NA[c]	NA[c]	100	10
	Introductory Physics	10	40	5	12	2	NA[c]	NA[c]	120	20
	Technology/Engineering	5	40	5	20	2	NA[c]	NA[c]	100	10

[a] The common items on each high school Biology form include a module consisting of 4 MC items and 1 OR item that are included in the overall counts.

[b] High school Biology matrix items may include one matrix module per form consisting of 4 MC items and 1 OR item. These are included in the overall matrix counts. If a module is not field-tested in a specific form, the spaces are used for stand-alone items.

[c] High school STE tests are pre-equated; therefore, the entire set of matrix slots is available for field-testing.

3.2.4.4 Blueprints

Grades 5 and 8

Table 3-11 shows the distribution of common items across the four strands of the *Massachusetts Science and Technology/Engineering Curriculum Framework*.

Table 3-11. 2012 MCAS: STE Common Point Distribution by Strand and Grade

Framework Strand	*Percent for Grade 5*	*Percent for Grade 8*
Earth and Space Science	30	25
Life Science	30	25
Physical Sciences	25	25
Technology/Engineering	15	25
Total	100	100

High School

Tables 3-12 through 3-15 show the distribution of common items across the various content strands for the MCAS high school STE tests.

Table 3-12. 2012 MCAS: High School Biology Common Point Distribution by Strand

MCAS Reporting Category	*Percent of Raw Score Points(±5%)*	*Related Framework Strand(s)*
Biochemistry and Cell Biology	25	• The Chemistry of Life • Cell Biology
Genetics	20	• Genetics
Anatomy and Physiology	15	• Anatomy and Physiology
Evolution and Biodiversity	20	• Evolution and Biodiversity
Ecology	20	• Ecology
Total	100	

Table 3-13. 2012 MCAS: High School Chemistry Common Point Distribution by Strand

MCAS Reporting Category	*Percent of Raw Score Points*	*Related Framework Strand(s)*
Atomic Structure and Periodicity	25	• Atomic Structure and Nuclear Chemistry • Periodicity
Bonding and Reactions	30	• Chemical Bonding • Chemical Reactions and Stoichiometry • Standard 8.4 from subtopic Acids and Bases and Oxidation Reduction Rates
Properties of Matter and Thermochemistry	25	• Properties of Matter • States of Matter, Kinetic Molecular Theory, and Thermochemistry
Solutions, Equilibrium, and Acid-Base Theory	20	• Solutions, Rates of Reaction, and Equilibrium • Acids and Bases and Oxidation Reduction Rates
Total	100	

Table 3-14. 2012 MCAS: High School Introductory Physics Common Point Distribution by Strand

MCAS Reporting Category	*Percent of Raw Score Points*	*Related Framework Strand(s)*
Motion and Forces	40	- Motion and Forces - Conservation of Energy and Momentum
Heat and Heat Transfer	15	- Heat and Heat Transfer
Waves and Radiation	25	- Waves - Electromagnetic Radiation
Electromagnetism	20	- Electromagnetism
Total	100	

Table 3-15. 2012 MCAS: High School Technology/Engineering Common Point Distribution by Strand

MCAS Reporting Category	*Percent of Raw Score Points*	*Related Framework Strand(s)*
Engineering Design	20	- Engineering Design
Construction and Manufacturing	20	- Construction Technologies - Manufacturing Technologies
Fluid and Thermal Systems	30	- Energy and Power Technologies – Fluid Systems - Energy and Power Technologies – Thermal Systems
Electrical and Communication Systems	30	- Energy and Power Technologies – Electrical Systems - Communication Technologies
Total	100	

3.2.4.5 Cognitive and Quantitative Skills

Each item on the STE test is assigned a cognitive level according to the cognitive demand of the item. Cognitive levels are not synonymous with difficulty. The cognitive level describes each item based on the complexity of the mental processing a student must use to answer the item correctly. Only one cognitive skill is designated for a common item, although several different cognitive skills may apply to a single item. In addition to the identified cognitive skill, an item may also be identified as having a quantitative component.

Table 3-16. 2012 MCAS: STE Cognitive Levels

Cognitive Skill (from basic to more demanding)	*Description*
Foundational	• Declarative knowledge • Recall of facts • Definition/vocabulary
Conceptual	• Recognition of a concept • Description of a principle • Description of a process
Application	• Procedural knowledge • Application of conceptual knowledge to a novel situation • Use of predetermined models to devise a solution • Classification of diverse objects into unifying groups *Note: This cognitive level does not automatically include all practical contexts for a concept; the application/situation for the concept must be a new, different example for the concept, not the example used in most textbooks.*
Constructive/ Synthetic	• Synthesis of a novel response (by pulling several different pieces of knowledge together) • Application of multi-step problem solving • Application of experimental design and critique • Formulation of a hypothesis • Application of predictive reasoning • Interpretation of experimental data analysis • Application of scientific inquiry or engineering design process
Other	*Description*
Quantitative	• Analysis of data • Computation of numerical solution • Graphical interpretation and interpretation of data in tables • Predictive calculations

3.2.4.6 Use of Calculators, Formula Sheets, and Rulers

Formula sheets are provided to students taking the high school Chemistry, Introductory Physics, and Technology/Engineering tests. These sheets contain information that students may need to answer certain test items. Students taking the Chemistry test also receive a copy of the Periodic Table of the Elements to use for reference during the test. Students taking the Technology/Engineering test receive an MCAS ruler. The use of calculators is allowed for all four of the high school STE tests, but is not required for the Biology test.

3.2.5 Test Development Process

Table 3-17 details the test development process.

Table 3-17. 2012 MCAS: Overview of Test Development Process

Development Step	*Details of the Process*
Select reading passages	For ELA only, test developers find potential passages and present them to the ESE, then to the grade level ADC, and finally to the Bias and Sensitivity Review Committee for review and recommendations.
Develop items	Test developers develop items in ELA, mathematics, and STE aligned to Massachusetts standards.
Review items and passages	1. Test developers review items internally with lead developer. 2. ESE reviews items prior to sending to ADCs. 3. ADCs review items and make recommendations. 4. Bias Committee reviews items and makes recommendations. 5. ESE determines final disposition of recommendations.
Edit items	Test developers make ESE-approved edits.
Field-test items	ESE-approved new items are included in the matrix portion of the MCAS tests.
Expert review of items	Experts and practitioners review all field-tested items for content accuracy.
Benchmark OR items and compositions	ESE and MP staff determine appropriate benchmark papers for training of scorers of OR items and compositions.
Item statistics meeting	ADCs review field-test statistics and recommend items for the common-eligible pool.
Test construction	Test developers from MP and ESE meet to construct the common and matrix portions of each test. Psychometricians are available to provide test characteristic curves and statistical information.
Operational test items	Items become part of the common item set and are used to determine individual student scores.
Released items	Approximately 50% of the common items in grades 3–8 are released to the public, and the remaining items return to the common-eligible pools; 100% of high school common items are released.

3.2.5.1 Item Development and ELA Passage Selection

Item Development

All items used on the MCAS tests are developed specifically for Massachusetts and are directly linked to the Massachusetts curriculum frameworks. The content standards contained within the frameworks are the basis for the reporting categories developed for each content area and are used to guide the development of assessment items. See Section 3.2.1 for specific content standard alignment. Content not found in the curriculum frameworks is not subject to the statewide assessment.

English Language Arts Reading Passages

Passages used in the reading comprehension portion of the ELA tests are authentic passages selected for the MCAS. See Section 3.2.2.7 for a detailed description of passage types and lengths. Test developers review numerous texts in order to find passages that possess the characteristics required for use in the ELA tests. Passages must be of interest to students; have a clear beginning, middle, and end; support the development of unique assessment items; and be free of bias and sensitivity issues. All passages used for MCAS English Language Arts assessments are published passages and are considered to be authentic literature.

3.2.5.2 Item and ELA Passage Reviews

Before being used as a part of ELA tests, all proposed passages, items, and scoring guides undergo extensive reviews. Test developers are cognizant of the passage requirements and carefully evaluate texts before presenting them to the ESE for review.

Review by ESE

ESE Passage Review

ESE content specialists review potential passages before presenting the passages for ADC review. Passages are reviewed for

- grade-level appropriateness;
- content appropriateness;
- richness of content (i.e., will it yield the requisite number of items?);
- bias and sensitivity issues.

Passages that are approved by the ESE are presented to the ADCs as well as the Bias and Sensitivity Committee for review and approval. Development of items with corresponding passages does not begin until the ESE has approved the passages.

ESE Item Review

All items and scoring guides are reviewed by the ESE content staff before presentation to the ADCs for review. The ESE evaluates the new items for the following elements:

- Alignment: Are the items aligned to the standards? Is there a better standard to which to align the item?
- Content: Does the item show a depth of understanding of the subject?
- Contexts: Are contexts used when appropriate? Are they realistic?
- Grade-level appropriateness: Are the content, language, and contexts appropriate for the grade level?
- Creativity: Does the item demonstrate creativity with regard to approaches to items and contexts?
- Distractors: Have the distractors for multiple-choice items been chosen based on common sources of error? Are they plausible?
- Mechanics: How well are the items written? Do they follow the conventions of item writing?

- Missed opportunities (for reading comprehension only): Were there items that should have been written based on the passage but were not written?

ESE staff members, in consultation with MP test developers, discuss and revise the proposed item sets in preparation for ADC review.

Review by ADCs

Once the ESE has reviewed passages, items, and scoring guides, and any requested changes have been made, materials are submitted to ADCs for further review. Each grade and content area has its own ADC composed of educators from across the state. Committees review new items for the elements listed above and provide insight into how standards are interpreted across the state. Committees make the following recommendations regarding new items:

- accept
- accept with edits (may include suggested edits)
- reject

ELA ADCs have the additional task of reviewing all passages before any corresponding items are written. Committee members consider all the elements listed above for passages (i.e., grade-level and content appropriateness, richness of content, and bias and sensitivity issues) as well as familiarity to students. If a passage is well known to many students or if the passage comes from a book that is widely taught, there is likely to be an unfair advantage to those students who are familiar with the work. Committee members make the same recommendations for passages that they make for items:

- accept
- accept with edits (may include suggested edits)
- reject

The committee members provide suggestions for items that could be written for the passage. They also provide recommendations for formatting and presentation of the passage, including suggestions for the purpose-setting statement, recommendations for words to be footnoted, and recommendations for graphics, illustrations, and photographs to be included with the text. For a list of committee members, see Appendix A.

Review by Bias and Sensitivity Review Committee

The Bias and Sensitivity Review Committee is composed of educators and members of the educational community from across the state who assist the ESE in reviewing items for possible bias and sensitivity concerns. (For a list of committee members, see Appendix A.) The Bias and Sensitivity Review Committee does not make recommendations regarding the content, alignment, or grade-level appropriateness of items or passages. Committee members review materials strictly and solely for issues of bias and sensitivity that may cause differential performance of students for reasons that are not related to the content being assessed.

Bias and Sensitivity Passage Review

All passages undergo a review by the Bias and Sensitivity Review Committee before they are approved for development. Committee members evaluate the content of all passages in terms of gender, race, ethnicity, geography, religion, sexual orientation, culture, and social appropriateness and make recommendations to accept or reject passages. They review the passages to ensure that students taking the test are not at a disadvantage because of issues not related to the construct being tested. All recommendations to reject passages are accompanied by explanations of the bias or sensitivity issue and why the passage should not be accepted. The ESE makes the final decision to accept or reject a passage. Items are not developed for passages until the passages have been accepted by the Bias and Sensitivity Review Committee and approved by the ESE.

Bias and Sensitivity Item Review

All items also undergo scrutiny by the Bias and Sensitivity Review Committee. The committee reviews all items after they have been developed and reviewed by the ADCs. (If an ADC rejects an item, the item does not go to the Bias and Sensitivity Review Committee.) The Bias and Sensitivity Review Committee makes the following recommendations regarding items:

- accept
- accept with edits (the committee identifies the nature of the issue prompting this request)
- reject (the committee describes the problem with the item and why rejecting the item is recommended)

Once the Bias and Sensitivity Review Committee has made its recommendations and the ESE has determined whether to act on the recommendations, the items move to the next step in the development process.

Review by External Content Experts

When items are selected to be included on the field-test portion of the MCAS, they are submitted to expert reviewers for their feedback. The task of the expert reviewer is to consider the accuracy of the content of items. Each item is reviewed by two independent expert reviewers. All expert reviewers for MCAS hold a doctoral degree in either philosophy or education and are affiliated with institutions of higher education in either teaching or research positions. Each expert reviewer has been approved by the ESE. Expert reviewers' comments are included with the items when they are sent to ADC meetings for statistics reviews. Expert reviewers are not expected to comment on grade-level appropriateness, mechanics of items, or any other aspect of an item other than content.

3.2.5.3 Item Editing

ESE content specialists review the recommendations of the committees and edit items accordingly. The items are also reviewed and edited by MP editors to ensure adherence to style guidelines in *The Chicago Manual of Style*, to MCAS-specific style guidelines, and to sound testing principles. According to these principles, items should

- demonstrate correct grammar, punctuation, usage, and spelling;
- be written in a clear, concise style;
- contain unambiguous explanations that tell students what is required to attain a maximum score;

- be written at a reading level that allows students to demonstrate their knowledge of the subject matter being tested;
- exhibit high technical quality regarding psychometric characteristics.

3.2.5.4 Field-Testing of Items

Items that have made it through the reviews listed above are approved to be field-tested. Field-tested items appear in the matrix portion of the test. Each item is answered by a minimum of 1,800 students, resulting in enough responses to yield reliable performance data.

3.2.5.5 Scoring of Field-Tested Items

Each field-tested multiple-choice item is machine-scored. Each constructed-response item (short-answer, short-response, or open-response) is hand-scored. In order to train scorers, the ESE works closely with the scoring staff to refine the rubrics and to select benchmark papers that exemplify the score points and the variations within each score point. Approximately 1,800 student responses are scored per constructed-response item.

3.2.5.6 Data Review of Field-Tested Items

Data Review by ESE

The ESE reviews all item statistics prior to making them available to the ADCs for review. Items that display statistics that indicate the item did not perform as expected are closely reviewed to ensure that the item is not flawed.

Data Review by ADCs

The ADCs meet to review the items with their field-test statistics. The ADCs make one of the following recommendations regarding each field-test item:

- accept
- edit and re-field-test
- reject

If an item is edited after it has been field-tested, the item cannot be used in the common portion of the test until it has been field-tested again. If the ADC recommends editing an item based on the item statistics, that item returns to the field-test-eligible pool to be re-field-tested. ADCs consider the following when reviewing field-test item statistics:

- item difficulty (or mean score for polytomous items)
- item discrimination
- differential item functioning

Data Review by Bias and Sensitivity Review Committee

The Bias and Sensitivity Review Committee also reviews the field-tested items with their item statistics. The committee reviews only the items that the ADCs have accepted. The Bias and

Sensitivity Review Committee pays special attention to items that may show differential function when comparing the following subgroups of test-takers:

- female/male
- black/white
- Hispanic/white
- English language learners and former English language learners who have been transitioned out of ELL for fewer than two years/native English speakers and former English language learners who have been transitioned from ELL for two or more years.

The Bias and Sensitivity Review Committee makes recommendations to the ESE regarding the disposition of items based on their item statistics. The ESE makes the final decision regarding the Bias and Sensitivity Committee recommendations.

3.2.5.7 Item and ELA Passage Selection and Operational Test Assembly

MP test developers propose a set of items to be used in the common portion of the test. Test developers work closely with psychometricians to ensure that the proposed tests meet the statistical requirements set forth by the ESE. In preparation for meeting with the ESE content specialists, the test developers and psychometricians at MP consider the following criteria in selecting sets of items to propose for the common portion of the test:

- **Content coverage/match to test design and blueprints.** The test designs and blueprints stipulate a specific number of items per item type for each content area. Item selection for the embedded field test is based on the number of items in the existing pool of items that are eligible for the common portion of the test.
- **Item difficulty and complexity.** Item statistics drawn from the data analysis of previously field-tested items are used to ensure similar levels of difficulty and complexity from year to year as well as high-quality psychometric characteristics.
- **"Cueing" items.** Items are reviewed for any information that might "cue" or help the students answer another item.

The test developers then distribute the items into test forms. During assembly of the test forms, the following criteria are considered:

- **Key patterns.** The sequence of keys (correct answers) is reviewed to ensure that their order appears random.
- **Option balance.** Items are balanced across forms so that each form contains a roughly equivalent number of key options (As, Bs, Cs, and Ds).
- **Page fit.** Item placement is modified to ensure the best fit and arrangement of items on any given page.
- **Facing-page issues.** For multiple-choice items associated with a stimulus (reading passages and high school biology modules) and multiple-choice items with large graphics, consideration is given to whether those items need to begin on a left- or right-hand page and to the nature and amount of material that needs to be placed on facing pages. These considerations serve to minimize the amount of page flipping required of students.
- **Relationships among forms.** Although field-test items differ from form to form, these items must take up the same number of pages in all forms so that sessions begin on the same page

in every form. Therefore, the number of pages needed for the longest form often determines the layout of all other forms.

- **Visual appeal.** The visual accessibility of each page of the form is always taken into consideration, including such aspects as the amount of "white space," the density of the test, and the number of graphics.

3.2.5.8 Operational Test Draft Review

The proposed operational test is delivered to the ESE for review. The ESE content specialists consider the proposed items, make recommendations for changes, and then meet with MP test developers and psychometricians to construct the final versions of the tests.

3.2.5.9 Special Edition Test Forms

Students with Disabilities

All MCAS 2012 operational tests and retests were available in the following editions for students with disabilities (in order to be eligible to receive one of these editions, a student needed to have an IEP or a 504 plan, or have a 504 plan in development):

- Large-print – Form 1 of the operational test is translated into a large-print edition. The large-print edition contains all common and matrix items found in Form 1.
- Braille – This form includes only the common items found in the operational test.
- Electronic text reader CD – This CD, in Kurzweil format, contains only common items found in the operational test.

In addition, the grade 10 MCAS Mathematics test was available to students with disabilities in an American Sign Language DVD edition, which contains only the common items found in the operational test.

Spanish-Speaking Students

A Spanish/English edition of the spring Grade 10 Mathematics test and the March and November Mathematics retests was available for Spanish-speaking ELL students who had been enrolled in school in the continental United States for fewer than three years and could read and write in Spanish at or near grade level. The Spanish/English edition of the spring grade 10 Mathematics test contains all common and matrix items found in Form 1 of the operational test. Each item is presented twice, once in Spanish on the left-hand page and once in English on the right-hand page. The Spanish/English edition of the test is not available in any other special formats.

Quality-Control Procedures

The approved test form is submitted to a primary translation vendor. The test form and translation is then sent to a second vendor for review. A set of translation guidelines has been developed over the years, and is provided to both vendors. This document outlines ESE preferences to maintain consistency from administration to administration as well as from year to year. When the reviewed document is sent back to Measured Progress, any questions or issues reported from the secondary vendor are resolved with the primary vendor. When necessary, an in-house expert is used for consultation.

Schools ordered special editions in advance of testing.

3.3 Test Administration

3.3.1 Test Administration Schedule

The standard MCAS tests were administered during three periods in the spring of 2012:

- March–April
    - Grades 3–8 and 10 ELA
- May
    - Grades 3–8 and 10 Mathematics
    - Grades 5 and 8 STE
- June
    - High school (grades 9–12) end-of-course STE
        - Biology
        - Chemistry
        - Introductory Physics
        - Technology/Engineering

The 2012 MCAS administration also included retest opportunities in ELA and Mathematics for students in grades 11 and 12, and students who had exited high school, who had not previously passed one or both grade 10 tests. Retests were offered in November 2011 and March 2012.

An additional high school (grades 9–12) end-of-course STE test in Biology was administered in February 2012.

Table 3-18 shows the complete 2011–2012 MCAS test administration schedule.

Table 3-18. 2012 MCAS: Test Administration Schedule

Grade and Content Area	*Test Administration Date(s)*	*Deadline for Return of Materials to Contractor*
Retest Administration Windows		
November 9–16, 2011		
ELA Composition Retest	November 14	November 18
ELA Reading Comprehension Retest		
Sessions 1 and 2	November 15	
Session 3	November 16	
Mathematics Retest		
Session 1	November 9	
Session 2	November 10	

continued

February 29–March 7, 2012		
ELA Composition Retest	February 29	March 12
ELA Reading Comprehension Retest Sessions 1 and 2 Session 3	March 1 March 2	
Mathematics Retest Session 1 Session 2	March 5 March 7	March 12
March–April 2012 Test Administration Window		
Grades 3–8 ELA Reading Comprehension	March 20–April 2	April 4
Grades 4, 7, and 10 ELA Composition	March 20	
Grade 10 ELA Reading Comprehension Sessions 1 and 2 Session 3	March 21 March 22	
Grades 4, 7, and 10 ELA Composition Make-Up	March 29	
May 2012 Test Administration Window		
Grades 3–8 Mathematics	May 7–22	May 23
Grades 5 and 8 STE	May 8–22	
Grade 10 Mathematics Session 1 Session 2	May 15 May 16	
High School (Grades 9–12) End-of-Course STE Test Administration Windows		
February 1–2, 2012		
Biology	February 1–2	February 7
June 5–6, 2012		
Biology	June 5–6	June 12
Chemistry		
Introductory Physics		
Technology/Engineering		

3.3.2 Security Requirements

Principals are responsible for ensuring that all test administrators comply with the requirements and instructions contained in the *Test Administrator's Manuals*. In addition, other administrators, educators, and staff within the school are responsible for complying with the same requirements. Schools and school staff who violate the test security requirements are subject to numerous possible sanctions and penalties, including employment consequences, delays in reporting of test results, the invalidation of test results, the removal of school personnel from future MCAS administrations, and possible licensure consequences for licensed educators.

Full security requirements, including details about responsibilities of principals and test administrators, examples of testing irregularities, guidance for establishing and following a document tracking system, and lists of approved and unapproved resource materials, can be found in the *Spring 2012 Principal's Administration Manual*, the *Fall 2011/Winter 2012 Principal's Administration Manual*, and all *Test Administrator's Manuals*.

3.3.3 Participation Requirements

In spring 2012, students educated with Massachusetts public funds were required by state and federal laws to participate in MCAS testing. The 1993 Massachusetts Education Reform Act mandates that **all** students in the tested grades who are educated with Massachusetts public funds participate in the MCAS, including the following groups of students:

- students enrolled in public schools
- students enrolled in charter schools
- students enrolled in educational collaboratives
- students enrolled in private schools receiving special education that is publicly funded by the Commonwealth, including approved and unapproved private special education schools within and outside Massachusetts
- students enrolled in institutional settings receiving educational services
- students in mobile military families
- students in the custody of either the Department of Children and Families (DCF) or the Department of Youth Services (DYS)
- students with disabilities, including students with temporary disabilities such as a broken arm.
- ELL students

It is the responsibility of the principal to ensure that all enrolled students participate in testing as mandated by state and federal laws. To certify that **all** students participate in testing as required, principals are required to complete the online Principal's Certification of Proper Test Administration (PCPA) following each test administration. See Appendix B for a summary of participation rates.

3.3.3.1 Students Not Tested on Standard Tests

A very small number of students educated with Massachusetts public funds are not required to take the standard MCAS tests. These students are strictly limited to the following categories:

- ELL students in their first year of enrollment in U.S. schools, who are not required to participate in ELA testing
- students with significant disabilities who must instead participate in the MCAS-Alt (see Chapter 4 for more information)
- students with a medically documented absence who are unable to participate in make-up testing

More details about test administration policies and student participation requirements at all grade levels, including requirements for earning a CD, requirements for students with disabilities or students who are ELLs, and students educated in alternate settings, can be found in the *Spring 2012 Principal's Administration Manual* and the *Fall 2011/Winter 2012 Principal's Administration Manual*.

3.3.4 Administration Procedures

It is the principal's responsibility to coordinate the school's MCAS test administration. This coordination responsibility includes the following:

- understanding and enforcing test security requirements
- ensuring that all enrolled students participate in testing at their grade level and that all eligible high school students are given the opportunity to participate in testing
- coordinating the school's test administration schedule and ensuring that tests with prescribed dates are administered on those dates
- ensuring that accommodations are properly provided and that transcriptions, if required for any accommodation, are done appropriately (Accommodation frequencies can be found in Appendix C. For a list of test accommodations, see Appendix D.)
- completing and ensuring the accuracy of information provided on the PCPA
- monitoring the ESE's website (www.doe.mass.edu/mcas) throughout the school year for important updates
- providing the ESE with correct contact information to receive important notices via fax during test administration

More details about test administration procedures, including ordering test materials, scheduling test administration, designating and training qualified test administrators, identifying testing spaces, meeting with students, providing accurate student information, and accounting for and returning test materials, can be found in the *Spring 2012 Principal's Administration Manual* and the *Fall 2011/Winter 2012 Principal's Administration Manual*.

The MCAS program is supported by the MCAS Service Center, which includes a toll-free telephone line answered by staff members who provide telephone support to schools and districts. The MCAS Service Center operates weekdays from 7:00 a.m. to 5:00 p.m. (Eastern Standard Time), Monday through Friday.

3.4 Scoring

MP scanned each MCAS student answer booklet into an electronic imaging system called iScore—a secure server-to-server interface designed by MP.

Student identification information, demographic information, school contact information, and student answers to multiple-choice items were converted to alphanumeric format. This information was not visible to scorers. Digitized student responses to constructed-response items were sorted into specific content areas, grade levels, and items before being scored.

3.4.1 Machine-Scored Items

Student responses to multiple-choice items were machine-scored by applying a scoring key to the captured responses. Correct answers were assigned a score of one point; incorrect answers were assigned a score of zero points. Student responses with multiple marks and blank responses were also assigned zero points.

3.4.2 Hand-Scored Items

Item-specific groups of responses were scored one item at a time; readers within each group scored one response at a time. Each individual response was linked through iScore to its original booklet number, so scoring leadership had access, if necessary, to a student's entire answer booklet.

3.4.2.1 Scoring Location and Staff

While the iScore database, its operation, and its administrative controls were all based in Dover, New Hampshire, MCAS item responses were scored in various locations, as summarized in Table 3-19.

Table 3-19. 2012 MCAS: Summary of Scoring Locations and Scoring Shifts

Measured Progress Scoring Center, Content Area	*Grade(s)*	*Shift*	*Hours*
Dover, NH			
STE: Chemistry	HS (9–12)	Day	8:00 a.m.–4:00 p.m.
STE: Technology/Engineering	HS (9–12)	Day	8:00 a.m.–4:00 p.m.
Longmont, CO			
ELA reading comprehension	4, 7, 8, 10	Day	8:00 a.m.–4:00 p.m.
ELA reading comprehension	3, 5, 6	Night	5:30 p.m.–10:30 p.m.
Mathematics	3, 7, 8, 10	Day	8:00 a.m.–4:00 p.m.
Mathematics	4, 5, 6	Night	5:30 p.m.–10:30 p.m.
Louisville, KY			
ELA composition	4	Day	8:00 a.m.–4:00 p.m.
Menands, NY			
ELA composition	7	Day	8:00 a.m.–4:00 p.m.
ELA composition	10	Day	8:00 a.m.–4:00 p.m.
STE: Biology	HS (9–12)	Day	8:00 a.m.–4:00 p.m.
STE: Introductory Physics	HS (9–12)	Night	5:30 p.m.–10:30 p.m.
STE	5	Day	8:00 a.m.–4:00 p.m.
STE	8	Night	5:30 p.m.–10:30 p.m.

The following staff members were involved with scoring the 2012 MCAS responses:

- The **MCAS scoring project manager (SPM)** was located in Dover, New Hampshire, and oversaw communication and coordination of MCAS scoring across all scoring sites.
- The **iScore operations manager** was located in Dover, New Hampshire, and coordinated technical communication across all scoring sites.
- A **scoring center manager (SCM)** was located at each satellite scoring location and provided logistical coordination for his or her scoring site.
- A **chief reader (CR)** in mathematics, STE, ELA reading comprehension, or ELA composition ensured consistency of content area benchmarking and scoring across all grade levels at all scoring locations. Chief readers monitored and read behind onsite and offsite **quality assurance coordinators**.
- Several **quality assurance coordinators (QACs)**, selected from a pool of experienced senior readers, participated in benchmarking, training, scoring, and cleanup activities for specified content areas and grade levels. QACs monitored and read behind **senior readers**.

- **Senior readers (SRs)**, selected from a pool of skilled and experienced readers, monitored and read behind **readers** at their scoring tables. Each senior reader monitored 2 to 11 readers.

3.4.2.2 Benchmarking Meetings

Samples of student responses to field-test items were read, scored, and discussed by members of MP's Scoring Services division and Content, Design & Development division as well as ESE staff at content- and grade-specific benchmarking meetings. All decisions were recorded and considered final upon ESE signoff.

The primary goals of the field-test benchmarking meetings were to

- revise, if necessary, an item's scoring guide;
- revise, if necessary, an item's scoring notes, which are listed beneath the score point descriptions and provide additional information about the scoring of that item;
- assign official score points to as many of the sample responses as possible;
- approve various individual responses and sets of responses (e.g., anchor, training) to be used to train field-test scorers.

3.4.2.3 Scorer Recruitment and Qualifications

MCAS scorers, a diverse group of individuals with a wide range of backgrounds, ages, and experiences, were primarily obtained through the services of a temporary employment agency, Kelly Services. All MCAS scorers successfully completed at least two years of college; hiring preference was given to those with a four-year college degree. Scorers for all grades 9–12 common, equating, and field-test responses were required to have a four-year baccalaureate.

Teachers, tutors, and administrators (principals, guidance counselors, etc.) currently under contract or employed by or in Massachusetts schools, and people under 18 years of age, were not eligible to score MCAS responses. Potential scorers were required to submit an application and documentation such as résumés and transcripts, which were carefully reviewed. Regardless of their degree, if potential scorers did not clearly demonstrate content area knowledge or have at least two college courses with average or above-average grades in the content area they wished to score, they were eliminated from the applicant pool.

Table 3-20 is a summary of scorers' backgrounds across all scoring shifts at all scoring locations.

Table 3-20. 2012 MCAS: Summary of Scorers' Backgrounds Across Scoring Shifts and Scoring Locations

Education	*Number*	*Percent*
Less than 48 college credits	0	0.0
Associate's degree/more than 48 college credits	191	9.8
Bachelor's degree	1110	57.2
Master's degree/doctorate	641	33.0

continued

Teaching Experience		
No teaching certificate or experience	957	49.3
Teaching certificate or experience	842	43.3
College instructor	143	7.4
Scoring Experience		
No previous experience as scorer	699	36.0
1–3 years of experience	669	34.4
3+ years of experience	574	29.6

3.4.2.4 Methodology for Scoring Polytomous Items

The MCAS tests included polytomous items requiring students to generate a brief response. Polytomous items included short-answer items, with assigned scores of 0–1; short-response items (grade 3 ELA only), with assigned scores of 0–2; open-response items requiring a longer or more complex response, with assigned scores of 0–4; and the writing prompt for the ELA composition, with assigned scores of 1–4 and 1–6.

The sample below of a 4-point mathematics open-response scoring guide was one of the many different item-specific MCAS scoring guides used in 2012. The task associated with this scoring guide required students to design four different gardens, each with a different shape.

Table 3-21. 2012 MCAS: Four-Point OR Item Scoring Guide – Grade 10 Mathematics

Score	*Description*
4	The student response demonstrates an exemplary understanding of the Measurement concepts involved in using area formulas to determine dimensions of a rectangle, triangle, trapezoid, and circle of a given area.
3	The student response demonstrates a good understanding of the Measurement concepts involved in using area formulas to determine dimensions of a rectangle, triangle, trapezoid, and circle of a given area. Although there is significant evidence that the student was able to recognize and apply the concepts involved, some aspect of the response is flawed. As a result the response merits 3 points.
2	The student response demonstrates fair understanding of the Measurement concepts involved in using area formulas to determine dimensions of a rectangle, triangle, trapezoid, and circle of a given area. While some aspects of the task are completed correctly, others are not. The mixed evidence provided by the student merits 2 points.
1	The student response demonstrates only minimal understanding of the Measurement concepts involved in using area formulas to determine dimensions of a rectangle, triangle, trapezoid, and circle of a given area.
0	The student response contains insufficient evidence of an understanding of the Measurement concepts involved in using area formulas to determine dimensions of a rectangle, triangle, trapezoid, and circle of a given area to merit any points.

Readers could assign a score-point value to a response or designate the response as one of the following:

- **Blank:** The written response form is completely blank (no graphite).
- **Unreadable:** The text on the scorer’s computer screen is too faint to see accurately.
- **Wrong Location:** The response seems to be a legitimate answer to a different question.

Responses initially marked as “Unreadable” or “Wrong Location” were resolved by scorers and iScore staff by matching all responses with the correct item or by pulling the actual answer booklet to look at the student’s original work.

Scorers may have also flagged a response as a “Crisis” response, which would be sent to scoring leadership for immediate attention.

A response may have been flagged as a “Crisis” response if it indicated

- perceived, credible desire to harm self or others;
- perceived, credible, and unresolved instances of mental, physical, or sexual abuse;
- presence of dark thoughts or serious depression;
- sexual knowledge well beyond the student’s developmental age;
- ongoing, unresolved misuse of legal/illegal substances (including alcohol);
- knowledge of or participation in real, unresolved criminal activity;
- direct or indirect request for adult intervention/assistance (e.g., crisis pregnancy, doubt about how to handle a serious problem at home).

Student responses were either single-scored (each response was scored only once) or double-blind scored (each response was independently read and scored by two different scorers). In double-blind scoring, neither scorer knew whether the response had been scored before, and if it had been scored, what score it had been given. A double-blind response with discrepant scores between the two scorers (i.e., a difference greater than one point if there are three or more score points) was sent to the arbitration queue and read by an SR or QAC.

All polytomous items on all high school tests (ELA, mathematics, and STE), as well as the ELA composition at grades 4, 7, and 10, were double-blind scored. Ten percent of polytomous items on the ELA reading comprehension, mathematics, and STE tests at grades 3–8 were double-blind scored.

In addition to the 10% or 100% double-blind scoring, SRs, at random points throughout the scoring shift, engaged in read-behind scoring for each of the scorers at his or her table. This process involved SRs viewing responses recently scored by a particular scorer and, without knowing the scorer’s score, assigning his or her own score to that same response. The SR would then compare scores and advise or counsel the scorer as necessary.

Table 3-22 outlines the rules for instances when the two read-behind or two double-blind scores were not identical (i.e., adjacent or discrepant).

Table 3-22. 2012 MCAS: Read-Behind and Double-Blind Resolution Charts

Read-Behind Scoring*			
Scorer #1	*Scorer #2*	*QAC/SR Resolution*	*Final*
4	-	4	4
4	-	3	3
4	-	2	2
* In all cases, the QAC score is the final score of record.			

continued

Double-Blind Scoring*			
Scorer #1	*Scorer #2*	*QAC/SR Resolution*	*Final*
4	4	-	4
4	3	-	4
3	4	-	4
4	2	3	3
4	1	2	2
3	1	1	1

* If scorer scores are identical or adjacent, the highest score is used as the final score. If scorer scores are neither identical nor adjacent, the resolution score is used as the final score.

Writing Standard English Conventions Double-Blind Scoring*			
Scorer #1	*Scorer #2*	*QAC/SR Resolution*	*Final*
4	4	-	8
4	3	-	7
4	2	4	8
4	2	3	7
4	1	3	7
4	1	2	3

* Identical or adjacent reader scores are summed to obtain the final score. The resolution score, if needed, is summed with an identical reader score; or, if the resolution score is adjacent to reader #1 and/or #2 but not identical with either, then the two highest adjacent scores are summed for the final score.

Writing Topic Development Double-Blind Scoring*				
Scorer #1	*Scorer #2*	*QAC/SR Resolution*	*Chief Reader*	
6	6	-	-	
6	5	-	-	
6	4	4	-	
6	4	5	-	
6	2	4	4	
6	2	4	3	
6	2	3	-	

* Identical or adjacent scorer scores are summed to obtain the final score. The resolution score, if needed, is summed with an identical scorer score; or, if the resolution score is adjacent to reader #1 and/or #2 but not identical with either, then the two highest adjacent scores are summed for the final score. If the resolution score is still discrepant, the CR assigns a fourth score, which is doubled to obtain the final score.

3.4.2.5 Reader Training

CRs had overall responsibility for ensuring that readers scored responses consistently, fairly, and according to the approved scoring guidelines. Scoring materials were carefully compiled and checked for consistency and accuracy. The timing, order, and manner in which the materials were presented to readers were planned and carefully standardized to ensure that all scorers had the same training environment and scoring experience, regardless of scoring location, content, grade level, or item scored.

MCAS trainers often had an opportunity to choose between modes of delivery for the training. The trainer may have trained by physically standing in front of, and speaking directly to, an entire room of scorers. If the scoring room contained a number of different subgroups of readers scoring different items, grade levels, content areas, etc., trainers trained their select subgroup via computer software that allowed document sharing, electronic polling, texting via an instant messaging system, and back-and-forth communication through headphones with built-in microphones.

Due to technological advances and robust computer servers, scorers were trained on some items via computers connected to a remote location; that is, the CR or training QAC was sitting at his or her computer in one scoring center, and the readers were sitting at their computers at a different scoring center. Interaction between readers and trainers remained uninterrupted through instant messaging or two-way audio communication devices, or through the onsite training supervisors.

CRs started the training process with an overview of the MCAS; this general orientation included the purpose and goal of the testing program and any unique features of the test and the testing population. Reader training for a specific item to be scored always started with a thorough review and discussion of the scoring guide, which consisted of the task, the scoring rubric, and any specific scoring notes for that task. All scoring guides were previously approved by the ESE during field-test benchmarking meetings and used without any additions or deletions.

As part of training, prospective readers carefully reviewed up to four different sets of actual student responses, some of which had been used to train readers when the item was a field-test item:

- **Anchor sets** are ESE-approved sets consisting of two to three sample responses at each score point. Each response is typical, rather than unusual or uncommon; solid, rather than controversial; and true, meaning that these responses have scores that cannot be changed.
- **Training sets** include unusual, discussion-provoking responses, illustrating the range of responses encountered in operational scoring (e.g., responses with both very high and very low attributes, exceptionally creative approaches, extremely short or disorganized responses).
- **Ranking sets** include one clear, mid-range example for each score point, distributed to readers in mixed (scrambled) score-point order. Ranking sets are not always used, but if they are, scorers rank-order them according to their true score points.
- **Qualifying sets** consist of 10 responses that were clear, typical examples of each of the score points. Qualifying sets are used to determine if readers were able to score according to the ESE-approved scoring rubric.

Meeting or surpassing the minimum acceptable standard on an item's qualifying set was an absolute requirement for scoring student responses to that item. An individual scorer must have attained a scoring accuracy rate of 70% exact and 90% exact plus adjacent agreement (at least 7 out of the 10 were exact score matches and either zero or one discrepant) on either of two potential qualifying sets.

3.4.2.6 Leadership Training

CRs also had overall responsibility for ensuring that scoring leadership (QACs and SRs) scored consistently, fairly, and according to the approved scoring guidelines. Scoring leadership must have met or surpassed a higher qualification standard of at least 80% exact and 90% exact plus adjacent, or, for grade 10 leadership, at least 80% exact and 100% adjacent.

3.4.2.7 Monitoring of Scoring Quality Control

Once MCAS readers met or exceeded the minimum standard on a qualifying set and were allowed to begin scoring, they were constantly monitored throughout the entire scoring window to be sure they scored student responses as accurately and consistently as possible. If a reader fell below the minimum standard on any of the quality control tools, there was some form of reader intervention, ranging from counseling to retraining to dismissal. Readers were required to meet or exceed the minimum standard of 70% exact and 90% exact plus adjacent agreement on the following:

- recalibration assessments (RAs)
- embedded committee-reviewed responses (CRRs)
- read-behind readings (RBs)
- double-blind readings (DBs)
- compilation reports, an end-of-shift report combining RAs and RBs

RAs given to readers at the very beginning of a scoring shift consisted of a set of five responses representing the entire range of possible scores. If scorers had an exact score match on at least four of the five responses, and were at least adjacent on the fifth response, they were allowed to begin scoring operational responses. Readers who had discrepant scores, or only two or three exact score matches, were retrained and, if approved by the SR, given extra monitoring assignments such as additional RBs and allowed to begin scoring. Readers who had zero or one out of the five exact were typically reassigned to another item or sent home for the day.

CRRs were responses approved by the CR and loaded into iScore for blind distribution to readers at random points during the scoring of their first 200 operational responses. While the number of CRRs ranged from 5 to 30, depending on the item, for most items MCAS readers received 10 of these previously scored responses during the first day of scoring that particular item. Readers who fell below the 70% exact and 90% exact plus adjacent accuracy standard were counseled and, if approved by the SR, given extra monitoring assignments such as additional RBs and allowed to resume scoring.

RBs involved responses that were first read and scored by a reader, then read and scored by an SR. SRs would, at various points during the scoring shift, command iScore to forward the next one, two, or three responses to be scored by a particular reader. After the reader scored these responses, and without knowing the score given by the reader, the SR would give his or her own score to the response and then be allowed to compare his or her score to the reader's score. RBs were performed at least 10 times for each full-time day shift reader and at least five times for each evening shift and partial-day shift reader. Readers who fell below the 70% exact and 90% exact plus adjacent score match standard were counseled, given extra monitoring assignments such as additional RBs, and allowed to resume scoring.

DBs involved responses scored independently by two different readers. Readers knew some of the responses they scored were going to be scored by others, but they had no way of knowing if they were the first, second, or only scorer. Readers who fell below the 70% exact and 90% exact plus adjacent score match standard during the scoring shift were counseled, given extra monitoring assignments such as additional RBs, and likely allowed to resume scoring. Responses given discrepant scores by two independent readers were read and scored by an SR.

Compilation reports combined a reader's percentage of exact, adjacent, and discrepant scores on the RA with that reader's percentage of exact, adjacent, and discrepant scores on the reader/SR RBs. Once the SR completed the minimum number of required RBs for a reader, the reader's overall percentages on the compilation reports were automatically calculated. If the compilation report at the end of the scoring shift listed individuals who were still below the 70% exact/90% exact plus adjacent level, their scores for that day were voided. Responses with scores voided were returned to the scoring queue for other readers to score.

If a reader fell below standard on the end-of-shift compilation report, and therefore had his or her scores voided on three separate occasions, the reader was automatically dismissed from scoring that item. If a reader was dismissed from scoring two MCAS items within a grade and content area, the reader was not allowed to score any additional items within that grade and content area. If a reader was dismissed from two different grade/content areas, the reader was dismissed from the project.

3.5 Classical Item Analyses

As noted in Brown (1983), "A test is only as good as the items it contains." A complete evaluation of a test's quality must include an evaluation of each item. Both *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 1999) and the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004) include standards for identifying quality items. Items should assess only knowledge or skills that are identified as part of the domain being tested and should avoid assessing irrelevant factors. Items should also be unambiguous and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. In addition, items must not unfairly disadvantage students in particular racial, ethnic, or gender groups.

Both qualitative and quantitative analyses are conducted to ensure that MCAS items meet these standards. Qualitative analyses are described in earlier sections of this chapter; this section focuses on quantitative evaluations. Statistical evaluations are presented in four parts: (1) difficulty indices, (2) item-test correlations, (3) differential item functioning (DIF) statistics, and (4) dimensionality analyses. The item analyses presented here are based on the statewide administration of the MCAS in spring 2012. Note that the information presented in this section is based on the items common to all forms, since those are the items on which student scores are calculated. (Item analyses are also performed for field-test items, and the statistics are then used during the item review process and form assembly for future administrations.)

3.5.1 Classical Difficulty and Discrimination Indices

All multiple-choice and open-response items are evaluated in terms of item difficulty according to standard classical test theory practices. Difficulty is defined as the average proportion of points achieved on an item and is measured by obtaining the average score on an item and dividing it by the maximum possible score for the item. Multiple-choice items are scored dichotomously (correct vs. incorrect) so, for these items, the difficulty index is simply the proportion of students who correctly answered the item. Open-response items are scored polytomously, meaning that a student can achieve a score of 0, 1, 2, 3, or 4. By computing the difficulty index as the average proportion of points achieved, the indices for the different item types are placed on a similar scale, ranging from 0.0 to 1.0 regardless of the item type. Although this index is traditionally described as a measure of difficulty, it is properly interpreted as an easiness index, because larger values indicate easier items.

An index of 0.0 indicates that all students received no credit for the item, and an index of 1.0 indicates that all students received full credit for the item.

Items that are answered correctly by almost all students provide little information about differences in student abilities, but they do indicate knowledge or skills that have been mastered by most students. Similarly, items that are correctly answered by very few students provide little information about differences in student abilities, but they may indicate knowledge or skills that have not yet been mastered by most students. In general, to provide the best measurement, difficulty indices should range from near-chance performance (0.25 for four-option multiple-choice items or essentially zero for open-response items) to 0.90, with the majority of items generally falling between 0.4 and 0.7. However, on a standards-referenced assessment such as the MCAS, it may be appropriate to include some items with very low or very high item difficulty values to ensure sufficient content coverage.

A desirable characteristic of an item is for higher-ability students to perform better on the item than lower-ability students. The correlation between student performance on a single item and total test score is a commonly used measure of this characteristic of the item. Within classical test theory, the item-test correlation is referred to as the item's discrimination, because it indicates the extent to which successful performance on an item discriminates between high and low scores on the test. For open-response items, the item discrimination index used was the Pearson product-moment correlation; for multiple-choice items, the corresponding statistic is commonly referred to as a point-biserial correlation. The theoretical range of these statistics is -1.0 to 1.0, with a typical observed range from 0.2 to 0.6.

Discrimination indices can be thought of as measures of how closely an item assesses the same knowledge and skills assessed by other items contributing to the criterion total score. That is, the discrimination index can be thought of as a measure of construct consistency, where 1 represents a high level of construct consistency and -1 represents a negative relationship.

A summary of the item difficulty and item discrimination statistics for each grade and content area combination is presented in Table 3-23. Note that the statistics are presented for all items as well as by item type (multiple-choice and open-response). The mean difficulty (*p*-value) and discrimination values shown in the table are within generally acceptable and expected ranges and are consistent with results obtained in previous administrations.

Table 3-23. 2012 MCAS: Summary of Item Difficulty and Discrimination Statistics by Content Area and Grade

Content Area	*Grade*	*Item Type*	*Number of Items*	p-*Value*		*Discrimination*	
				Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
ELA	3	ALL	41	0.77	0.10	0.43	0.07
		MC	36	0.79	0.09	0.42	0.07
		OR	5	0.64	0.06	0.48	0.09
	4	ALL	42	0.76	0.12	0.41	0.09
		MC	36	0.79	0.08	0.39	0.06
		OR	6	0.55	0.14	0.59	0.05
	5	ALL	40	0.73	0.13	0.40	0.08
		MC	36	0.76	0.11	0.38	0.05
		OR	4	0.52	0.06	0.58	0.06

continued

Content Area	*Grade*	*Item Type*	*Number of Items*	p-*Value*		*Discrimination*	
				Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
ELA	6	ALL	40	0.74	0.11	0.41	0.10
		MC	36	0.76	0.10	0.39	0.07
		OR	4	0.57	0.04	0.61	0.02
	7	ALL	42	0.73	0.13	0.41	0.13
		MC	36	0.75	0.12	0.36	0.08
		OR	6	0.64	0.11	0.67	0.02
	8	ALL	40	0.78	0.09	0.43	0.10
		MC	36	0.80	0.08	0.41	0.09
		OR	4	0.63	0.04	0.61	0.02
	10	ALL	42	0.79	0.11	0.41	0.12
		MC	36	0.81	0.11	0.37	0.07
		OR	6	0.68	0.10	0.66	0.04
Mathematics	3	ALL	36	0.76	0.13	0.43	0.07
		MC	26	0.80	0.11	0.42	0.06
		OR	10	0.67	0.12	0.48	0.08
	4	ALL	42	0.70	0.11	0.41	0.10
		MC	32	0.70	0.12	0.38	0.07
		OR	10	0.69	0.10	0.50	0.11
	5	ALL	42	0.72	0.13	0.46	0.10
		MC	32	0.74	0.12	0.43	0.08
		OR	10	0.63	0.12	0.55	0.12
	6	ALL	42	0.72	0.12	0.46	0.11
		MC	32	0.74	0.11	0.42	0.08
		OR	10	0.64	0.12	0.60	0.09
	7	ALL	42	0.68	0.10	0.46	0.09
		MC	32	0.68	0.10	0.44	0.06
		OR	10	0.67	0.12	0.51	0.13
	8	ALL	42	0.67	0.11	0.49	0.11
		MC	32	0.67	0.11	0.47	0.08
		OR	10	0.66	0.13	0.57	0.14
	10	ALL	42	0.67	0.12	0.47	0.13
		MC	32	0.69	0.13	0.41	0.08
		OR	10	0.63	0.10	0.66	0.08
STE	5	ALL	42	0.70	0.16	0.38	0.09
		MC	38	0.73	0.14	0.36	0.07
		OR	4	0.46	0.07	0.57	0.02
	8	ALL	42	0.67	0.14	0.40	0.09
		MC	38	0.70	0.12	0.38	0.07
		OR	4	0.44	0.05	0.59	0.06
Biology	HS	ALL	45	0.71	0.14	0.42	0.11
		MC	40	0.74	0.12	0.40	0.08
		OR	5	0.47	0.09	0.63	0.07
Chemistry	HS	ALL	45	0.68	0.14	0.43	0.11
		MC	40	0.70	0.14	0.40	0.08
		OR	5	0.54	0.10	0.66	0.07

continued

Content Area	*Grade*	*Item Type*	*Number of Items*	p-*Value*		*Discrimination*	
				Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
Introductory Physics	HS	ALL	45	0.65	0.15	0.40	0.13
		MC	40	0.67	0.14	0.37	0.08
		OR	5	0.46	0.14	0.70	0.03
Technology/ Engineering	HS	ALL	45	0.62	0.17	0.37	0.12
		MC	40	0.63	0.17	0.35	0.09
		OR	5	0.51	0.17	0.60	0.07

A comparison of indices across grade levels is complicated because these indices are population dependent. Direct comparisons would require that either the items or students were common across groups. Since that is not the case, it cannot be determined whether differences in performance across grade levels are explained by differences in student abilities, differences in item difficulties, or both.

Difficulty indices for multiple-choice items tend to be higher (indicating that students performed better on these items) than the difficulty indices for open-response items because multiple-choice items can be answered correctly by guessing. Similarly, discrimination indices for the 4-point open-response items were larger than those for the dichotomous items because of the greater variability of the former (i.e., the partial credit these items allow) and the tendency for correlation coefficients to be higher, given greater variances of the correlates. Note that these patterns are an artifact of item type, so when interpreting classical item statistics, comparisons should be made only among items of the same type.

In addition to the item difficulty and discrimination summaries presented above, item-level classical statistics and item-level score point distributions were also calculated. Item-level classical statistics are provided in Appendix E; item difficulty and discrimination values are presented for each item. The item difficulty and discrimination indices are within generally acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that students who performed well on individual items tended to perform well overall. There were a small number of items with discrimination indices below 0.20, but none were negative. While it is not inappropriate to include items with low discrimination values or with very high or very low item difficulty values to ensure that content is appropriately covered, there were very few such cases on the MCAS. Item-level score point distributions are provided for open-response items in Appendix F; for each item, the percentage of students who received each score point is presented.

3.5.2 Differential Item Functioning

The *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004) explicitly states that subgroup differences in performance should be examined when sample sizes permit and that actions should be taken to ensure that differences in performance are attributable to construct-relevant, rather than irrelevant, factors. *Standards for Educational and Psychological Testing* (AERA et al., 1999) includes similar guidelines. As part of the effort to identify such problems, psychometricians evaluated MCAS items in terms of DIF statistics.

For the MCAS, the standardization DIF procedure (Dorans & Kulick, 1986) was employed to evaluate subgroup differences. The standardization DIF procedure is designed to identify items for which subgroups of interest perform differently, beyond the impact of differences in overall achievement. The DIF procedure calculates the difference in item performance for two groups of students (at a time) matched for achievement on the total test. Specifically, average item performance is calculated for students at every total score. Then an overall average is calculated, weighting the total score distribution so that it is the same for the two groups. For all grades and content areas except high school STE, DIF statistics are calculated for all subgroups that include at least 100 students; for high school STE, the minimum is 50 students. To enable calculation of DIF statistics for the limited English proficient/formerly limited English proficient (LEP/FLEP) comparison, the minimum was set at 50 for all grade levels.

When differential performance between two groups occurs on an item (i.e., a DIF index in the "low" or "high" categories explained below), it may or may not be indicative of item bias. Course-taking patterns or differences in school curricula can lead to low or high DIF, but for construct-relevant reasons. On the other hand, if subgroup differences in performance can be traced to differential experience (such as geographical living conditions or access to technology), the inclusion of such items should be reconsidered.

Computed DIF indices have a theoretical range from -1.0 to 1.0 for multiple-choice items, and the index is adjusted to the same scale for open-response items. Dorans and Holland (1993) suggested that index values between -0.05 and 0.05 should be considered negligible. The majority of MCAS items fell within this range. Dorans and Holland further stated that items with values between -0.10 and -0.05 and between 0.05 and 0.10 (i.e., "low" DIF) should be inspected to ensure that no possible effect is overlooked, and that items with values outside the -0.10 to 0.10 range (i.e., "high" DIF) are more unusual and should be examined very carefully.[2]

For the 2012 MCAS, DIF analyses were conducted for all subgroups (as defined in NCLB) for which the sample size was adequate. In all, six subgroup comparisons were evaluated for DIF:

- male/female
- white/black
- white/Hispanic
- no disability/disability
- not LEP-FLEP/LEP-FLEP
- not low-income/low-income

The tables in Appendix G present the number of items classified as either "low" or "high" DIF, in total and by group favored. Overall, a moderate number of items exhibited low DIF and several exhibited high DIF; the numbers were fairly consistent with results obtained for previous administrations of the test.

3.5.3 Dimensionality Analysis

Because tests are constructed with multiple content area subcategories and their associated knowledge and skills, the potential exists for a large number of dimensions being invoked beyond the common primary dimension. Generally, the subcategories are highly correlated with each other; therefore, the primary dimension they share typically explains an overwhelming majority of variance

[2] DIF for items is evaluated initially at the time of field testing. If an item displays high DIF, it is flagged for review by a Measured Progress content specialist. The content specialist consults with the ESE to determine whether to include the flagged item in a future operational test administration. All DIF statistics are reviewed by the Assessment Development Committees at their stat reviews.

in test scores. In fact, the presence of just such a dominant primary dimension is the psychometric assumption that provides the foundation for the unidimensional item response theory (IRT) models that are used for calibrating, linking, scaling, and equating the MCAS test forms for grades 3 through 8 and high school.

The purpose of dimensionality analysis is to investigate whether violation of the assumption of test unidimensionality is statistically detectable and, if so, (a) the degree to which unidimensionality is violated and (b) the nature of the multidimensionality. Dimensionality analyses were performed on common items for all MCAS tests administered during the spring 2011-12 administrations. A total of 20 tests were analyzed, and the results for these analyses are reported below, including a comparison with the results from 2010-11.

The dimensionality analyses were conducted using the nonparametric IRT-based methods DIMTEST (Stout, 1987; Stout, Froelich, & Gao, 2001) and DETECT (Zhang & Stout, 1999). Both of these methods use as their basic statistical building block the estimated average conditional covariances for item pairs. A conditional covariance is the covariance between two items conditioned on true score (expected value of observed score) for the rest of the test, and the average conditional covariance is obtained by averaging all possible conditioning scores. When a test is strictly unidimensional, all conditional covariances are expected to take on values within random noise of zero, indicating statistically independent item responses for examinees with equal expected scores. Non-zero conditional covariances are essentially violations of the principle of local independence, and such local *dependence* implies multidimensionality. Thus, nonrandom patterns of positive and negative conditional covariances are indicative of multidimensionality.

DIMTEST is a hypothesis-testing procedure for detecting violations of local independence. The data are first randomly divided into a training sample and a cross-validation sample. Then an exploratory analysis of the conditional covariances is conducted on the training sample data to find the cluster of items that displays the greatest evidence of local dependence. The cross-validation sample is then used to test whether the conditional covariances of the selected cluster of items display local dependence, conditioning on total score on the nonclustered items. The DIMTEST statistic follows a standard normal distribution under the null hypothesis of unidimensionality.

DETECT is an effect-size measure of multidimensionality. As with DIMTEST, the data are first randomly divided into a training sample and a cross-validation sample (these samples are drawn independently of those used with DIMTEST). The training sample is used to find a set of mutually exclusive and collectively exhaustive clusters of items that best fit a systematic pattern of positive conditional covariances for pairs of items from the same cluster and negative conditional covariances from different clusters. Next, the clusters from the training sample are used with the cross-validation sample data to average the conditional covariances: within-cluster conditional covariances are summed, from this sum the between-cluster conditional covariances are subtracted, this difference is divided by the total number of item pairs, and this average is multiplied by 100 to yield an index of the average violation of local independence for an item pair. DETECT values less than 0.2 indicate very weak multidimensionality (or near unidimensionality); values of 0.2 to 0.4, weak to moderate multidimensionality; values of 0.4 to 1.0, moderate to strong multidimensionality; and values greater than 1.0, very strong multidimensionality.

DIMTEST and DETECT were applied to the common items of the MCAS tests administered during spring 2012 (a total of 20 tests). The data for each grade were split into a training sample and a cross-validation sample. Each of the elementary and middle school grades had over 68,000 student examinees per test. For the high school tests, mathematics and ELA also had over 68,000 student examinees, biology had over 52,000, physics had over 17,000, chemistry had over 1,500, and technology/engineering had over 2,000. Because DIMTEST had an upper limit of 24,000 students, the training and cross-validation samples for the tests that had over 24,000 students were limited to 12,000 each, randomly sampled from the total sample. DETECT, on the other hand, had an upper limit of 500,000 students, so every training sample and cross-validation sample used all the available data. After randomly splitting the data into training and cross-validation samples, DIMTEST was applied to each dataset to see if the null hypothesis of unidimensionality would be rejected. DETECT was then applied to each dataset for which the DIMTEST null hypothesis was rejected in order to estimate the effect size of the multidimensionality.

3.5.3.1 DIMTEST Analyses

The results of the DIMTEST analyses indicated that the null hypothesis was rejected at a significance level of 0.01 for every dataset. Because strict unidimensionality is an idealization that almost never holds exactly for a given dataset, the statistical rejections in the DIMTEST results were not surprising. Indeed, because of the very large sample sizes involved in most of the datasets (over 50,000 in 17 of the 20 tests), DIMTEST would be expected to be sensitive to even quite small violations of unidimensionality.

3.5.3.2 DETECT Analyses

Next, DETECT was used to estimate the effect size for the violations of local independence for all the tests. Table 3-24 below displays the multidimensionality effect-size estimates from DETECT.

Table 3-24. 2012 MCAS: Multidimensionality Effect Sizes by Grade and Content Area

Content Area	*Grade*	*Multidimensionality Effect Size*	
		2012	***2011***
ELA	3	0.13	0.11
	4	0.15	0.21
	5	0.14	0.14
	6	0.13	0.17
	7	0.16	0.15
	8	0.17	0.17
	10	0.17	0.16
	Average	**0.15**	**0.16**
Mathematics	3	0.12	0.13
	4	0.14	0.15
	5	0.19	0.14
	6	0.14	0.18
	7	0.18	0.13
	8	0.14	0.17
	10	0.14	0.12
	Average	**0.15**	**0.15**
STE	5	0.14	0.08
	8	0.11	0.11
	(Biology) 9–12	0.08	0.07
	(Chemistry) 9–12	0.13	0.18

continued

Content Area	*Grade*	*Multidimensionality Effect Size*	
		2012	***2011***
STE	(Introductory Physics) 9–12	0.11	0.09
	(Technology/Engineering) 9–12	0.07	0.12
	Average	**0.11**	**0.11**

The DETECT values indicate very weak to weak multidimensionality for all the tests for 2012. The ELA test forms (average effect size of about 0.15) and the mathematics test forms (average of about 0.15) tended to show slightly greater multidimensionality than did the science test forms (average of about 0.11). Also shown in Table 3-24 are the values reported in last year's dimensionality analyses. The average DETECT values in 2011 for ELA and mathematics were 0.16 and 0.15, respectively, and the average for the science tests was 0.11. Thus, last year's results are very similar to those from this year.

The way in which DETECT divided the tests into clusters was also investigated to determine whether there were any discernable patterns with respect to the multiple-choice and constructed-response item types. Inspection of the DETECT clusters indicated that multiple-choice/constructed-response separation generally occurred much more strongly with ELA than with mathematics or science, a pattern that has been consistent across all five years of dimensionality analyses for the MCAS tests. Specifically, for ELA every grade had one set of clusters dominated by multiple-choice items and another set of clusters dominated by constructed-response items. This particular pattern within ELA has occurred in all six years of the MCAS dimensionality analyses, with the exception of grade 3, which has usually not shown this pattern. However, over the past two years ELA grade 3 separation of constructed-response and multiple-choice items has been clear. Of the seven mathematics tests, only grade 4 showed evidence of consistent separation of multiple-choice and constructed-response. Of the six science tests, only grade 5 showed strong multiple-choice/constructed-response separation. In comparison to past years, no single grade has had consistent multiple-choice/constructed-response separation every year within the mathematics or science content areas.

Thus, a tendency is suggested for multiple-choice and constructed-response items to sometimes measure statistically separable dimensions, especially in regard to the ELA tests. This has been consistent across all six years of MCAS analyses. However, the sizes of the violations of local independence have been small in all cases. The degree to which these small violations can be attributed to item type differences tends to be greater for ELA than for mathematics or science. More investigation by content experts would be required to better understand the violations of local independence that are due to sources other than item type.

In summary, for the 2011–12 analyses the violations of local independence, as evidenced by the DETECT effect sizes, were either weak or very weak in all cases. Thus, these effects do not seem to warrant any changes in test design or scoring. In addition, the magnitude of the violations of local independence have been consistently low over the years, and the patterns with respect to the multiple-choice and constructed-response items have also been consistent, with ELA tending to display more separation than the other two content areas.

3.6 MCAS IRT Scaling and Equating

This section describes the procedures used to calibrate, equate, and scale the MCAS tests. During the course of these psychometric analyses, a number of quality control procedures and checks on the processes were conducted. These procedures included

- evaluations of the calibration processes (e.g., checking the number of Newton cycles required for convergence for reasonableness);
- checking item parameters and their standard errors for reasonableness;
- examination of test characteristic curves (TCCs) and test information functions (TIFs) for reasonableness);
- evaluation of model fit;
- evaluation of equating items (e.g., delta analyses, rescore analyses);
- examination of *a*-plots and *b*-plots for reasonableness;
- evaluation of the scaling results (e.g., parallel processing by the Psychometrics and Research and Data and Reporting Services divisions, comparing lookup tables to the previous year's).

An equating report, which provided complete documentation of the quality-control procedures and results, was reviewed by the ESE and approved prior to production of the *Spring 2012 MCAS Tests Parent/Guardian Reports* (MP Department of Psychometrics and Research, *2011–2012 MCAS Equating Report*, unpublished manuscript).

Table 3-25 lists items that required intervention either during item calibration or as a result of the evaluations of the equating items. For each flagged item, the table shows the reason it was flagged (e.g., the *c* parameter could not be estimated, the delta analysis indicated that the item was flawed) and what action was taken. The number of items identified for evaluation was similar to the number identified in previous years and in other states across the grades and content areas. Descriptions of the evaluations and results are included in the Item Response Theory Results and Equating Results sections of this document.

Table 3-25. 2012 MCAS: Items That Required Intervention During IRT Calibration and Equating

Content Area	*Grade*	*Item Number*	*Reason*	*Action*
English Language Arts	03	257124	delta analysis	removed from equating
English Language Arts	03	270156	c-parameter	fix c=0
English Language Arts	03	276675	c-parameter	fix c=0
English Language Arts	03	284942	c-parameter	fix c=0
English Language Arts	04	255565	delta analysis	removed from equating
English Language Arts	04	255585	b/b analysis	removed from equating
English Language Arts	04	276990	c-parameter	fix c=0
English Language Arts	04	277002	c-parameter	fix c=0
English Language Arts	04	279990	c-parameter	fix c=0
English Language Arts	04	285358	c-parameter	fix c=0
English Language Arts	05	276124	c-parameter	fix c=0
English Language Arts	05	276125	c-parameter	fix c=0
English Language Arts	05	276127	c-parameter	fix c=0
English Language Arts	05	283505	c-parameter	fix c=0
English Language Arts	05	283510	c-parameter	fix c=0
English Language Arts	05	284538	c-parameter	fix c=0
English Language Arts	05	284540	c-parameter	fix c=0

continued

Content Area	Grade	Item Number	Reason	Action
English Language Arts	05	286700	c-parameter	fix c=0
English Language Arts	05	286801	b/b analysis	removed from equating
English Language Arts	05	287624	c-parameter	fix c=0
English Language Arts	06	257665	delta analysis	removed from equating
English Language Arts	06	285580	c-parameter	fix c=0
English Language Arts	06	285596	c-parameter	fix c=0
English Language Arts	06	292309	c-parameter	fix c=0
English Language Arts	06	292328	c-parameter	fix c=0
English Language Arts	06	292334	c-parameter	fix c=0
English Language Arts	06	292346	c-parameter	fix c=0
English Language Arts	06	292351	c-parameter	fix c=0
English Language Arts	07	266612	c-parameter	fix c=0
English Language Arts	07	277042	c-parameter	fix c=0
English Language Arts	07	279225	c-parameter	fix c=0
English Language Arts	07	279227	c-parameter	fix c=0
English Language Arts	07	280268	c-parameter	fix c=0
English Language Arts	07	280270	c-parameter	fix c=0
English Language Arts	07	280279	c-parameter	fix c=0
English Language Arts	07	292823	c-parameter	fix c=0
English Language Arts	08	227774	c-parameter	fix c=0
English Language Arts	08	275968	c-parameter	fix c=0
English Language Arts	08	275981	c-parameter	fix c=0
English Language Arts	08	276026	c-parameter	fix c=0
English Language Arts	08	291936	c-parameter	fix c=0
English Language Arts	08	291952	c-parameter	fix c=0
English Language Arts	08	293338	c-parameter	fix c=0
English Language Arts	08	294082	c-parameter	fix c=0
English Language Arts	10	279457	c-parameter	fix c=0
English Language Arts	10	279460	c-parameter	fix c=0
English Language Arts	10	280613	c-parameter	fix c=0
English Language Arts	10	292762C	b/b analysis	removed from equating
English Language Arts	10	292762T	b/b analysis	removed from equating
English Language Arts	10	293560	c-parameter	fix c=0
English Language Arts	10	293565	c-parameter	fix c=0
English Language Arts	10	293569	c-parameter	fix c=0
English Language Arts	10	293864	c-parameter	fix c=0
Mathematics	03	207697	c-parameter	fix c=0
Mathematics	03	207704	c-parameter	fix c=0
Mathematics	04	250326	c-parameter	fix c=0
Mathematics	04	258243	c-parameter	fix c=0
Mathematics	04	280093	b/b analysis	removed from equating
Mathematics	04	286762	c-parameter	fix c=0
Mathematics	04	299678	c-parameter	fix c=0
Mathematics	05	206922	b/b analysis	removed from equating
Mathematics	05	248875	delta analysis	removed from equating
Mathematics	05	261216	c-parameter	fix c=0
Mathematics	05	287253	c-parameter	fix c=0
Mathematics	05	287261	c-parameter	fix c=0
Mathematics	06	282272	b/b analysis	removed from equating
Mathematics	06	282272	delta analysis	removed from equating

continued

Content Area	Grade	Item Number	Reason	Action
Mathematics	06	299675	c-parameter	fix c=0
Mathematics	07	250345	delta analysis	removed from equating
Mathematics	07	261082	c-parameter	fix c=0
Mathematics	08	264730	c-parameter	fix c=0
Mathematics	08	264757	c-parameter	fix c=0
Science	05	245522	c-parameter	fix c=0
Science	05	264868	delta analysis	removed from equating
Science	05	273729	c-parameter	fix c=0
Science	05	273784	c-parameter	fix c=0
Science	05	281799	c-parameter	fix c=0
Science	08	265249	c-parameter	fix c=0
Science	08	282002	c-parameter	fix c=0
Science	08	288350	c-parameter	fix c=0
Science	08	291915	c-parameter	fix c=0
Biology	10	273543	delta analysis	retained for equating
Biology	10	283136	c-parameter	fix c=0
Biology	10	283215	delta analysis	retained for equating
Biology	10	290993	c-parameter	fix c=0
Chemistry	10	258965	b/b analysis	retained for equating
Chemistry	10	260762	b/b analysis	retained for equating
Physics	10	272652	delta analysis	retained for equating
Physics	10	287226	c-parameter	fix c=0
Physics	10	287234	b/b analysis	retained for equating
Technology and Engineering Sciences	10	206611	delta analysis	retained for equating

3.6.1 Item Response Theory

All MCAS items were calibrated using IRT. IRT uses mathematical models to define a relationship between an unobserved measure of student performance, usually referred to as theta (θ), and the probability (p) of getting a dichotomous item correct or of getting a particular score on a polytomous item (Hambleton, Swaminathan, & Rogers, 1991; Hambleton & Swaminathan, 1985). In IRT, it is assumed that all items are independent measures of the same construct (i.e., of the same θ). Another way to think of θ is as a mathematical representation of the latent trait of interest. Several common IRT models are used to specify the relationship between θ and p (Hambleton & van der Linden, 1997; Hambleton & Swaminathan, 1985). The process of determining the mathematical relationship between θ and p is called item calibration. After items are calibrated, they are defined by a set of parameters that specify a nonlinear, monotonically increasing relationship between θ and p. Once the item parameters are known, an estimate of θ for each student can be calculated. This estimate, $\hat{\theta}$, is considered to be an estimate of the student's true score or a general representation of student performance. It has characteristics that may be preferable to those of raw scores for equating purposes.

For the 2012 MCAS, the graded-response model (GRM) was used for polytomous items (Nering & Ostini, 2010) for all grade and content area combinations. The three-parameter logistic (3PL) model was used for dichotomous items for all grade and content-area combinations except high school STE, which used the one-parameter logistic (1PL) model (Hambleton & van der Linden, 1997; Hambleton, Swaminathan, & Rogers, 1991). The 1PL model was chosen for high school STE because there was concern that the tests might have too few examinees to support the 3PL model in future administrations.

The 3PL model for dichotomous items can be defined as:

$$P_i\left(1\middle|\theta_j,\xi_i\right) = c_i + (1 - c_i)\frac{\exp\left[Da_i\left(\theta_j - b_i\right)\right]}{1 + \exp\left[Da_i\left(\theta_j - b_i\right)\right]}$$

where
i indexes the items,
j indexes students,
a represents item discrimination,
b represents item difficulty,
c is the pseudo guessing parameter,
θ is the student ability,
ξ represents the set of item parameters (*a*, *b*, and *c*), and
D is a normalizing constant equal to 1.701.

For high school STE, this reduces to the following:

$$P_j(\theta_i) = \frac{\exp\left[D\left(\theta_i - b_j\right)\right]}{1 + \exp\left[D\left(\theta_i - b_j\right)\right]}$$

In the GRM for polytomous items, an item is scored in *k* + 1 graded categories that can be viewed as a set of *k* dichotomies. At each point of dichotomization (i.e., at each threshold), a two-parameter model can be used. This implies that a polytomous item with *k* + 1 categories can be characterized by *k* item category threshold curves (ICTCs) of the two-parameter logistic form:

$$P_{ik}^*\left(1\middle|\theta_j,a_i,b_i,d_{ik}\right) = \frac{\exp\left[Da_i\left(\theta_j - b_i + d_{ik}\right)\right]}{1 + \exp\left[Da_i\left(\theta_j - b_i + d_{ik}\right)\right]}$$

where
i indexes the items,
j indexes students,
k indexes threshold,
θ is the student ability,
a represents item discrimination,
b represents item difficulty,
d represents threshold, and
D is a normalizing constant equal to 1.701.

After computing *k* ICTCs in the GRM, *k* + 1 item category characteristic curves (ICCCs) are derived by subtracting adjacent ICTCs:

$$P_{ik}\left(1\middle|\theta_j\right) = P_{i(k-1)}^*\left(1\middle|\theta_j\right) - P_{ik}^*\left(1\middle|\theta_j\right)$$

where
i indexes the items,
j indexes students,
k indexes threshold,
θ is the student ability,
P_{ik} represents the probability that the score on item i falls in category *k*, and
P_{ik}^* represents the probability that the score on item i falls above the threshold *k*.
($P_{i0}^* = 1$ and $P_{i(m+1)}^* = 0$).

The GRM is also commonly expressed as:

$$P_{ik}\left(k\middle|\theta_j,\xi_i\right)=\frac{\exp\left[Da_i\left(\theta_j-b_i+d_k\right)\right]}{1+\exp\left[Da_i\left(\theta_j-b_i+d_k\right)\right]}-\frac{\exp\left[Da_i\left(\theta_j-b_i+d_{k+1}\right)\right]}{1+\exp\left[Da_i\left(\theta_j-b_i+d_{k+1}\right)\right]}$$

where
i indexes the items,
j indexes students,
k indexes threshold,
θ is the student ability,
a represents item discrimination,
b represents item difficulty,
d represents threshold, and
D is a normalizing constant equal to 1.701.
ξ_i represents the set of item parameters for item *i*.

Finally, the item characteristic curve (ICC) for polytomous items is computed as a weighted sum of ICCCs, where each ICCC is weighted by a score assigned to a corresponding category. The expected score for a student with a given theta is expressed as:

$$E_i\left(\theta_j\right)=\sum_{k}^{m+1} w_{ik}P_{ik}\left(k\middle|\theta_j\right)$$

where
i indexes the items,
j indexes students,
k indexes score category,
θ is the student ability,
w is the weighting constant, and is equal to the number of score points for the score category, and
P is the probability of a student with ability θ achieving score category *k*.

For more information about item calibration and determination, see Lord and Novick (1968), Hambleton and Swaminathan (1985), or Baker and Kim (2004).

3.6.2 Item Response Theory Results

The tables in Appendix H give the IRT item parameters and associated standard errors of all common items on the 2012 MCAS tests by grade and content area. Note that the standard errors for some parameters are equal to zero. In these cases, the parameter or parameters were not estimated, either because the item was an equating item or because the parameter's value was fixed (see explanation below). In addition, Appendix I contains graphs of the TCCs and TIFs, which are defined below. Because of the use of the 1PL model, a TIF is not provided for high school STE.

TCCs display the expected (average) raw score associated with each θ_j value between -4.0 and 4.0. Mathematically, the TCC is computed by summing the ICCs of all items that contribute to the raw score. Using the notation introduced in Section 3.6.1, the expected raw score at a given value of θ_j is

$$E(X|\theta_j) = \sum_{i=1}^{n} P_i(1|\theta_j)$$

where
i indexes the items (and n is the number of items contributing to the raw score),
j indexes students (here, θ_j runs from -4 to 4), and
$E(X|\theta_j)$ is the expected raw score for a student of ability θ_j.

The expected raw score monotonically increases with θ_j, consistent with the notion that students of high ability tend to earn higher raw scores than students of low ability. Most TCCs are "S-shaped": they are flatter at the ends of the distribution and steeper in the middle.

The TIF displays the amount of statistical information that the test provides at each value of θ_j. Information functions depict test precision across the entire latent trait continuum. There is an inverse relationship between the information of a test and its standard error of measurement (SEM). For long tests, the SEM at a given θ_j is approximately equal to the inverse of the square root of the statistical information at θ_j (Hambleton, Swaminathan, & Rogers, 1991), as follows:

$$SEM(\theta_j) = \frac{1}{\sqrt{I(\theta_j)}}$$

Compared to the tails, TIFs are often higher near the middle of the θ distribution where most students are located and where most items are sensitive by design.

Table 3-25 above lists items that were flagged based on the quality control checks implemented during the calibration process. (Note that some items were flagged as a result of the evaluations of the equating items; those results are described below.) In all cases, items flagged during this step were identified because of the guessing parameter (*c* parameter) being poorly estimated. Difficulty in estimating the *c* parameter is not at all unusual and is well documented in psychometric literature (see, for example, Nering & Ostini, 2010), especially when the item's discrimination is below 0.50. In all cases, fixing the *c* parameter resulted in reasonable and stable item parameter estimates and improved model fit.

The number of Newton cycles required for convergence for each grade and content area during the IRT analysis can be found in Table 3-26. The number of cycles required fell within acceptable ranges (less than 150) for all tests.

Table 3-26. 2012 MCAS: Number of Newton Cycles Required for Convergence

Content Area	*Grade*	*Cycles*	
		Initial	*Equating*
ELA	3	45	77
	4	50	153
	5	39	111
	6	40	96
	7	43	94
	8	48	45
	10	61	22
Mathematics	3	26	88
	4	40	91
	5	19	110
	6	35	123
	7	33	84
	8	69	127
	10	66	1
STE	5	29	96
	8	35	114
Biology	9–12	40	1
Chemistry	9–12	44	1
Introductory Physics	9–12	53	1
Technology/Engineering	9–12	34	1

3.6.3 Equating

The purpose of equating is to ensure that scores obtained from different forms of a test are equivalent to each other. Equating may be used if multiple test forms are administered in the same year, as well as to equate one year's forms to those used in the previous year. Equating ensures that students are not given an unfair advantage or disadvantage because the test form they took is easier or harder than those taken by other students. See Section 3.2 for more information about how the test development process supports successful equating.

The 2012 administration of the MCAS used a raw score-to-theta equating procedure in which test forms were equated to the theta scale established on the reference form (i.e., the form used in the most recent standard setting) for ELA grades 3–8, mathematics grades 3–10, STE, biology, chemistry, introductory physics, and technology/engineering. The procedure for ELA grade 10 is different, and is described at the end of this section. This is accomplished through the chained linking design, in which every new form is equated back to the theta scale of the previous year's test form. It can therefore be assumed that the theta scale of every new test form is the same as the theta scale of the reference form, since this is where the chain originated.

The groups of students who took the equating items on the 2012 MCAS ELA reading comprehension tests are not equivalent to the groups who took them in the reference years. IRT is particularly useful for equating scenarios that involve nonequivalent groups (Allen & Yen, 1979). Equating for the MCAS uses the anchor test-nonequivalent groups design described by Petersen, Kolen, and Hoover (1989). In this equating design, no assumption is made about the equivalence of the examinee groups taking different test forms (i.e., naturally occurring groups are assumed).

Comparability is instead evaluated by using a set of anchor items (also called equating items). The equating items are designed to mirror the common test in terms of item types and content coverage. Subsets of the equating items are matrix sampled across forms.

Item parameter estimates for 2012 were placed on the 2011 scale by using the Fixed Common Item Parameter method (FCIP2; Kim, 2006), which is based on the IRT principle of item parameter invariance. According to this principle, the equating items for both the 2011 and 2012 MCAS tests should have the same item parameters. After the item parameters for each 2012 test were estimated using PARSCALE (Muraki & Bock, 2003) to check for parameter drift of the equating items, the FCIP2 method was employed to place the nonequating items onto the operational scale. This method is performed by fixing the parameters of the equating items to their previously obtained on-scale values, and then calibrating using PARSCALE to place the remaining items on scale.

As described above, MCAS uses post-equating for most tests. In grade 10 all tests are pre-equated. Pre-equating is a procedure where the item parameters are estimated in a previous administration, and then fixed to these values rather than re-estimating them after the operational administration. These known item parameters are then used for estimating student performance and can also be used to bring new items onto scale. Since student performance and reported scores are based on the pre-equated item parameters, it is critical that the pre-equated item parameters are accurate and meet strict guidelines for fit and quality. In ELA grade 10 this year, the Topic Development and Conventions scores for the Composition were shown to exhibit poor model fit based on the previously estimated parameters (all the thresholds were three or more standard deviations away from the line of best fit for the other equating items). As a result, the Composition was removed from the equating process and the parameters for the Reading Comprehension sections were re-estimated using the FCIP procedures employed in the other MCAS post-equated tests.

3.6.4 Equating Results

Prior to equating the 2012 tests, various evaluations of the equating items were conducted. Items that were flagged as a result of these evaluations are listed in Table 3-25 at the beginning of this section. Each of these items was scrutinized, and a decision was made whether to include each item as an equating item or to discard it. The procedures used to evaluate the equating items are described below.

Appendix J presents the results from the delta analysis. This procedure was used to evaluate the adequacy of equating items; the discard status presented in the appendix indicates whether or not the item was flagged as potentially inappropriate for use in equating.

Also presented in Appendix J are the results from the rescore analysis. In this analysis, 200 random papers from the previous year were interspersed with this year's papers to evaluate scorer consistency from one year to the next. An effect size, comparing the difference between last year's score and this year's score using the same set of student responses with a new set of raters was calculated. All effect sizes were well below 0.80, the criterion value for excluding an item as an equating item.

Finally, *a*-plots and *b*-plots, which show IRT parameters for 2011 plotted against the values for 2012, are presented in Appendix K. Any items that appeared as outliers in the plots were evaluated in terms of suitability for use as equating items.

Once all flagged items had been evaluated and appropriate action taken, the FCIP2 method of equating was used to place the item parameters onto the previous year's scale, as described above. The next administration of the MCAS (2013) will be scaled to the 2012 administration using the same equating method described above.

3.6.5 Achievement Standards

Cutpoints for all MCAS tests were set via standard setting in previous years, establishing the theta cuts used for reporting each year. These theta cuts are presented in Table 3-27. These operational θ -metric cut scores will remain fixed throughout the assessment program unless standards are reset. Also shown in the table are the cutpoints on the reporting score scale (*2007 Standard Setting Report).*

Table 3-27. 2012 MCAS: Cut Scores on the Theta Metric and Reporting Scale by Content Area and Grade

Content Area	*Grade*	*Theta*				*Scaled Score*			
		Cut 1	*Cut 2*	*Cut 3*	*Min*	*Cut 1*	*Cut 2*	*Cut 3*	*Max*
ELA	3	-1.692	-0.238	1.128	200	220	240	260	280
	4	-1.126	0.067	1.572	200	220	240	260	280
	5	-1.535	-0.248	1.152	200	220	240	260	280
	6	-1.380	-0.279	1.392	200	220	240	260	280
	7	-1.529	-0.390	1.460	200	220	240	260	280
	8	-1.666	-0.637	1.189	200	220	240	260	280
	10	-0.414	0.384	1.430	200	220	240	260	280
Mathematics	3	-1.011	-0.087	1.031	200	220	240	260	280
	4	-0.859	0.449	1.308	200	220	240	260	280
	5	-0.714	0.170	1.049	200	220	240	260	280
	6	-0.510	0.232	1.112	200	220	240	260	280
	7	-0.485	0.264	1.190	200	220	240	260	280
	8	-0.318	0.418	1.298	200	220	240	260	280
	10	-0.189	0.420	1.038	200	220	240	260	280
STE	5	-1.130	0.090	1.090	200	220	240	260	280
	8	-0.500	0.540	1.880	200	220	240	260	280
Biology	9–12	-0.962	-0.129	1.043	200	220	240	260	280
Chemistry	9–12	-0.134	0.425	1.150	200	220	240	260	280
Introductory Physics	9–12	-0.714	0.108	1.133	200	220	240	260	280
Technology/ Engineering	9–12	-0.366	0.201	1.300	200	220	240	260	280

Appendix L shows achievement-level distributions by content area and grade. Results are shown for each of the last three years.

3.6.6 Reported Scaled Scores

Because the θ scale used in IRT calibrations is not understood by most stakeholders, reporting scales were developed for the MCAS. The reporting scales are linear transformations of the underlying θ scale within each performance level. Student scores on the MCAS tests are reported in even-integer values from 200 to 280. Because there are four separate transformations (one for each achievement level, shown in Table 3-28), a 2-point difference between scaled scores in the Warning/Failing level does not mean the same thing as a 2-point difference in the Needs Improvement level. Because the scales differ across achievement levels, it is not appropriate to calculate means and standard deviations with scaled scores.

By providing information that is more specific about the position of a student's results, scaled scores supplement achievement-level scores. Students' raw scores (i.e., total number of points) on the 2012 MCAS tests were translated to scaled scores using a data analysis process called scaling. Scaling simply converts from one scale to another. In the same way that a given temperature can be expressed on either the Fahrenheit or Celsius scale, or the same distance can be expressed in either miles or kilometers, student scores on the 2012 MCAS tests can be expressed in raw or scaled scores.

It is important to note that converting from raw scores to scaled scores does not change students' achievement-level classifications. Given the relative simplicity of raw scores, it is fair to question why scaled scores for the MCAS are reported instead of raw scores. The answer is that scaled scores make the reporting of results consistent. To illustrate, standard setting typically results in different raw cut scores across content areas. The raw cut score between Needs Improvement and Proficient could be, for example, 35 in grade 3 mathematics but 33 in grade 4 mathematics, yet both of these raw scores would be transformed to scaled scores of 240. It is this uniformity across scaled scores that facilitates the understanding of student performance. The psychometric advantage of scaled scores over raw scores comes from their being linear transformations of θ. Since the θ scale is used for equating, scaled scores are comparable from one year to the next. Raw scores are not.

The scaled scores are obtained by a simple translation of ability estimates ($\hat{\theta}$) using the linear relationship between threshold values on the θ metric and their equivalent values on the scaled score metric. Students' ability estimates are based on their raw scores and are found by mapping through the TCC. Scaled scores are calculated using the linear equation

$$SS = m\hat{\theta} + b$$

where
m is the slope, and
b is the intercept.

A separate linear transformation is used for each grade and content area combination and for each achievement level. Table 3-28 shows the slope and intercept terms used to calculate the scaled scores for each grade, content area, and achievement level. Note that the values in Table 3-28 will not change unless the standards are reset.

Appendix M contains raw-score-to-scaled-score lookup tables. The tables show the scaled score equivalent of each raw score for this year and last year.

Appendix N contains scaled score distribution graphs for each grade and content area. These distributions were calculated using the sparse data matrix files that were used in the IRT calibrations.

Table 3-28. 2012 MCAS: Scaled Score Slopes and Intercepts by Content Area and Grade

Content Area	*Grade*	*Line 1*		*Line 2*		*Line 3*		*Line 4*	
		Slope	*Intercept*	*Slope*	*Intercept*	*Slope*	*Intercept*	*Slope*	*Intercept*
ELA	3	4.8133	228.1442	13.7552	243.2737	14.6413	243.4846	10.6838	247.9487
	4	4.8025	225.4077	16.7645	238.8768	13.2890	239.1096	14.0056	237.9832
	5	6.0233	229.2458	15.5400	243.8539	14.2857	243.5429	10.8225	247.5325
	6	5.6670	227.8204	18.1653	245.0681	11.9689	243.3393	12.4378	242.6866
	7	6.1550	229.4110	17.5593	246.8481	10.8108	244.2162	12.9870	241.0390
	8	6.4299	230.7122	19.4363	252.3810	10.9529	246.9770	11.0436	246.8691
	10	4.2290	221.7508	25.0627	230.3759	19.1205	232.6577	12.7389	241.7834
Mathematics	3	6.1264	226.1938	21.6450	241.8831	17.8891	241.5564	10.1574	249.5277
	4	5.7890	224.9728	15.2905	233.1346	23.2829	229.5460	11.8203	244.5390
	5	5.9309	224.2347	22.6244	236.1538	22.7531	236.1320	10.2512	249.2465
	6	5.4533	222.7812	26.9542	233.7466	22.7273	234.7273	10.5932	248.2203
	7	4.9964	222.4233	26.7023	232.9506	21.5983	234.2981	11.0497	246.8508
	8	5.2593	221.6725	27.1739	228.6413	22.7273	230.5000	11.7509	244.7474
	10	4.2313	220.7997	32.8407	226.2069	32.3625	226.4078	10.1937	249.4190
STE	5	5.2597	225.9434	16.3934	238.5246	20.0000	238.2000	10.4712	248.5864
	8	4.8965	222.4483	19.2308	229.6154	14.9254	231.9403	17.8571	226.4286
Biology	9–12	4.9400	224.7523	24.0096	243.0972	17.0648	242.2014	10.2197	249.3408
Chemistry	9–12	4.3287	220.5800	35.7782	224.7943	27.5862	228.2759	10.8108	247.5676
Introductory Physics	9–12	4.4428	223.1722	24.3309	237.3723	19.5122	237.8927	10.7124	247.8629
Technology/ Engineering	9–12	7.1510	222.6173	35.2734	232.9101	18.1984	236.3421	11.7647	244.7059

3.7 MCAS Reliability

Although an individual item's performance is an important factor in evaluating an assessment, a complete evaluation must also address the way an overall set of items functions together and complements one another. Tests that function well provide a dependable assessment of a student's level of ability. A variety of factors can contribute to a given student's score being higher or lower than his or her true ability. For example, a student may misread an item or mistakenly fill in the wrong bubble when he or she knows the correct answer. Collectively, extraneous factors that affect a student's score are referred to as measurement error. Any assessment includes some amount of measurement error because no measurement is perfect.

There are a number of ways to estimate an assessment's reliability. One approach, called "test-retest reliability," is to give the same test to the same students at two different points in time. If students receive the same scores on each test, then the extraneous factors affecting performance are small and the test is reliable. A problem with this approach is that students may remember items from the first administration or may have gained (or lost) knowledge or skills in the interim between the two administrations. Another approach, "alternate forms reliability," is to give a different, but parallel, test at the second administration. If student scores on each test correlate highly, the test is considered reliable. This approach, however, does not address the possibility that students may have gained (or lost) knowledge or skills in the interim between the two administrations. In addition, the practical challenges of developing and administering parallel forms are substantial. A third approach, "split-half estimate of reliability," addresses the problems associated with the first two approaches. A test is split in half, and students' scores on the two half-tests are correlated; this in effect treats each half-test as a complete test. If the two half-test scores correlate highly, items on the two half-tests must be measuring very similar knowledge or skills. This is evidence that the items complement one another and function well as a group, suggesting that measurement error is minimal.

The split-half method requires psychometricians to select items that contribute to each half-test score. This decision may have an impact on the resulting correlation, since each different possible split of the test into halves will result in a different correlation. In addition, the split-half method underestimates reliability, because a shorter test is less reliable than a longer test. Cronbach (1951) provided a statistic, α (alpha), which eliminates the item selection and shorter test drawbacks of the split-half method by comparing individual item variances to total test variance. Cronbach's α was used to assess the reliability of the 2012 MCAS tests:

$$a \equiv \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} \sigma_{(Y_i)}^2}{\sigma_x^2}\right]$$

where
i indexes the item,
n is the total number of items,
$\sigma_{(Y_i)}^2$ represents individual item variance, and
σ_x^2 represents the total test variance.

3.7.1 Reliability and Standard Errors of Measurement

Table 3-29 presents descriptive statistics, Cronbach's *α* coefficient, and raw score SEMs for each content area and grade. (Statistics are based on common items only.) Generally, reliability estimates are in acceptable ranges, greater than 0.8, and are consistent with results obtained for previous administrations of the tests.

Table 3-29. 2012 MCAS: Raw Score Descriptive Statistics, Cronbach's Alpha, and SEMs by Content Area and Grade

Content Area	*Grade*	*Number of Students*	*Raw Score*			*Alpha*	*SEM*
			Maximum	*Mean*	*Standard Deviation*		
ELA	3	69,439	48	35.85	8.46	0.91	2.58
	4	68,873	72	49.32	10.12	0.89	3.28
	5	69,989	52	35.54	8.56	0.89	2.81
	6	70,239	52	36.41	8.98	0.90	2.87
	7	70,554	72	50.27	10.54	0.90	3.38
	8	71,657	52	38.88	8.78	0.90	2.74
	10	68,106	72	53.77	9.98	0.90	3.17
Mathematics	3	69,510	40	29.98	7.62	0.90	2.43
	4	69,032	54	37.24	10.31	0.90	3.30
	5	70,035	54	37.06	11.64	0.91	3.43
	6	70,258	54	37.92	11.44	0.92	3.26
	7	70,694	54	36.44	11.72	0.91	3.42
	8	71,536	54	35.05	12.58	0.93	3.38
	10	68,054	60	40.06	12.91	0.92	3.59
STE	5	70,107	54	34.97	9.84	0.88	3.44
	8	71,504	54	33.44	10.29	0.89	3.43
Biology	9–12	48,423	60	39.02	11.11	0.91	3.32
Chemistry	9–12	1,355	60	38.90	12.27	0.91	3.67
Introductory Physics	9–12	17,375	60	35.90	11.68	0.91	3.59
Technology/ Engineering	9–12	2,033	60	35.43	10.74	0.89	3.57

Because different grades and content areas have different test designs (e.g., the number of items varies by test), it is inappropriate to make inferences about the quality of one test by comparing its reliability to that of another test from a different grade or content area.

3.7.2 Interrater Consistency

Section 3.4.2 of this report describes the processes that were implemented to monitor the quality of the hand-scoring of student responses for constructed-response items. One of these processes was double-blind scoring: either 100% (for compositions and all high school tests) or 10% (all other open-response items) of student responses were randomly selected and scored independently by two different scorers. Results of the double-blind scoring were used during the scoring process to identify scorers who required retraining or other intervention, and are presented here as evidence of the reliability of the MCAS tests. A summary of the interrater consistency results are presented in Table 3-30 below. Results in the table are organized across the hand-scored items by content area and grade. The table shows the number of score categories, the number of included scores, the percent exact agreement, percent adjacent agreement, correlation between the first two sets of scores, and the percent of responses that required a third score. This same information is provided at the item level in Appendix O. These interrater consistency statistics are the result of the processes implemented to ensure valid and reliable hand-scoring of items as described in Section 3.4.2.

Table 3-30. 2012 MCAS: Summary of Interrater Consistency Statistics Organized Across Items by Content Area and Grade

Content Area	*Grade*	*Number of Items*	*Number of Score Categories*	*Number of Included Scores*	*Percent Exact*	*Percent Adjacent*	*Correlation*	*Percent of Third Scores*
ELA	3	4	3	27,596	81.29	18.21	0.77	0.50
		1	5	6,894	63.75	34.57	0.71	1.52
	4	1	4	67,966	66.79	32.44	0.62	1.29
		4	5	27,202	66.19	32.21	0.76	1.43
		1	6	67,966	68.06	31.08	0.75	1.29
	5	4	5	27,607	65.27	33.06	0.77	1.47
	6	4	5	27,445	57.76	38.93	0.72	2.93
	7	1	4	67,590	71.54	28.08	0.64	0.84
		4	5	27,275	65.03	33.42	0.79	1.35
		1	6	67,590	67.46	31.83	0.71	0.84
	8	4	5	27,657	62.39	35.94	0.76	1.44
	10	1	4	63,941	73.11	26.40	0.60	1.06
		4	5	261,898	65.55	33.35	0.77	0.90
		1	6	63,941	64.41	34.58	0.63	1.06
Mathematics	3	6	2	41,614	99.01	0.99	0.98	0.00
		4	3	27,807	92.34	7.41	0.93	0.26
	4	6	2	41,158	98.62	1.38	0.97	0.00
		4	5	27,456	81.06	16.99	0.91	1.93
	5	6	2	41,690	98.62	1.38	0.97	0.00
		4	5	27,809	84.15	13.79	0.94	2.02
	6	6	2	41,896	98.71	1.29	0.97	0.00
		4	5	27,891	80.49	17.62	0.92	1.84
	7	6	2	41,942	99.09	0.91	0.98	0.00
		4	5	28,024	84.86	12.96	0.94	2.12
	8	6	2	42,614	98.96	1.04	0.97	0.00
		4	5	28,143	85.54	13.19	0.95	1.22
	10	4	2	262,707	98.78	1.22	0.97	0.00
		6	5	397,404	85.23	14.07	0.95	0.64
STE	5	4	5	29,307	61.30	32.10	0.83	6.29
	8	4	5	27,880	75.29	22.77	0.90	1.84
Biology	9–12	5	5	229,927	71.27	26.25	0.86	2.19
Chemistry	9–12	5	5	6,542	65.77	29.38	0.86	4.49
Introductory Physics	9–12	5	5	82,643	71.42	25.62	0.89	2.80
Technology/ Engineering	9–12	5	5	9,455	66.37	29.32	0.85	4.00

3.7.3 Subgroup Reliability

The reliability coefficients discussed in the previous section were based on the overall population of students who took the 2012 MCAS tests. Appendix P presents reliabilities for various subgroups of interest. Cronbach's α coefficients were calculated using the formula defined above based only on the members of the subgroup in question in the computations; values are only calculated for subgroups with 10 or more students.

For several reasons, the subgroup reliability results should be interpreted with caution. First, inherent differences between grades and content areas preclude valid inferences about the reliability of a test based on statistical comparisons with other tests. Second, reliabilities are dependent not only on the measurement properties of a test but also on the statistical distribution of the studied subgroup. For example, Appendix P shows that subgroup sample sizes may vary considerably, which results in natural variation in reliability coefficients. Alternatively α, which is a type of correlation coefficient, may be artificially depressed for subgroups with little variability (Draper & Smith, 1998). Third, there is no industry standard to interpret the strength of a reliability coefficient, and this is particularly true when the population of interest is a single subgroup.

3.7.4 Reporting Subcategory Reliability

Reliabilities for the reporting subcategories within MCAS content areas are described in Section 3.2. Cronbach's α coefficients for subcategories were calculated via the same formula defined previously using just the items of a given subcategory in the computations. Results are presented in Appendix P. Once again, because they are based on a subset of items rather than the full test, subcategory reliabilities were lower (sometimes substantially so) than were overall test reliabilities, as expected, and interpretations should take this into account. The subcategory reliabilities were lower than those based on the total test and approximately to the degree one would expect, based on classical test theory. Qualitative differences between grades and content areas once again preclude valid inferences about the reliability of the full test based on statistical comparisons among subtests.

3.7.5 Reliability of Achievement Level Categorization

The accuracy and consistency of classifying students into achievement levels are critical components of a standards-based reporting framework (Livingston & Lewis, 1995). For the MCAS tests, students are classified into one of four achievement levels: *Warning* (*Failing* at high school), *Needs Improvement*, *Proficient*, or *Advanced*. MP conducted decision accuracy and consistency (DAC) analyses to determine the statistical accuracy and consistency of the classifications. This section explains the methodologies used to assess the reliability of classification decisions, and gives the results of these analyses. Section 3.2 describes the reporting categories in greater detail.

Accuracy refers to the extent to which achievement classifications based on test scores match the classifications that would have been assigned if the scores did not contain any measurement error. Accuracy must be estimated, because errorless test scores do not exist. Consistency measures the extent to which classifications based on test scores match the classifications based on scores from a second, parallel form of the same test. Consistency can be evaluated directly from actual responses to test items if two complete and parallel forms of the test are administered to the same group of students. In operational testing programs, however, such a design is usually impractical. Instead, techniques have been developed to estimate both the accuracy and consistency of classifications based on a single administration of a test. The Livingston and Lewis (1995) technique was used for the 2012 MCAS tests because it is easily adaptable to all types of testing formats, including mixed formats.

The DAC estimates reported in Appendix P make use of “true scores” in the classical test theory sense. A true score is the score that would be obtained if a test had no measurement error. True scores cannot be observed and so must be estimated. In the Livingston and Lewis method, estimated true scores are used to categorize students into their “true” classifications.

For the 2012 MCAS tests, after various technical adjustments (described in Livingston & Lewis, 1995), a four-by-four contingency table of accuracy was created for each content area and grade, where cell $[i, j]$ represented the estimated proportion of students whose true score fell into classification i (where $i = 1$ to 4) and observed score into classification j (where $j = 1$ to 4). The sum of the diagonal entries (i.e., the proportion of students whose true and observed classifications matched) signified overall accuracy.

To calculate consistency, true scores were used to estimate the joint distribution of classifications on two independent, parallel test forms. Following statistical adjustments (per Livingston and Lewis, 1995), a new four-by-four contingency table was created for each content area and grade and populated by the proportion of students who would be categorized into each combination of classifications according to the two (hypothetical) parallel test forms. Cell $[i, j]$ of this table represented the estimated proportion of students whose observed score on the first form would fall into classification i (where $i = 1$ to 4) and whose observed score on the second form would fall into classification j (where $j = 1$ to 4). The sum of the diagonal entries (i.e., the proportion of students categorized by the two forms into exactly the same classification) signified overall consistency.

MP also measured consistency on the 2012 MCAS tests using Cohen’s (1960) coefficient κ (kappa), which assesses the proportion of consistent classifications after removing the proportion of consistent classifications that would be expected by chance. It is calculated using the following formula:

$$\kappa = \frac{(\text{Observed agreement}) - (\text{Chance agreement})}{1 - (\text{Chance agreement})} = \frac{\sum_i C_{ii} - \sum_i C_{i.}C_{.i}}{1 - \sum_i C_{i.}C_{.i}}$$

where
$C_{i.}$ is the proportion of students whose observed achievement level would be level *i* (where *i* = 1–4) on the first hypothetical parallel form of the test;
$C_{.i}$ is the proportion of students whose observed achievement level would be level *i* (where *i* = 1–4) on the second hypothetical parallel form of the test;
C_{ii} is the proportion of students whose observed achievement level would be level *i* (where *i* = 1–4) on both hypothetical parallel forms of the test.

Because κ is corrected for chance, its values are lower than other consistency estimates.

3.7.6 Decision Accuracy and Consistency Results

Results of the DAC analyses described above are provided in Table 3-31. The table includes overall accuracy and consistency indices, including kappa. Accuracy and consistency values conditional upon achievement level are also given. For these calculations, the denominator is the proportion of students associated with a given achievement level. For example, the conditional accuracy value is 0.72 for *Needs Improvement* for grade 3 mathematics. This figure indicates that among the students whose true scores placed them in this classification, 72% would be expected to be in this classification when categorized according to their observed scores. Similarly, a consistency value of 0.64 indicates that 64% of students with observed scores in the *Needs Improvement* level would be expected to score in this classification again if a second, parallel test form were taken.

For some testing situations, the greatest concern may be decisions around achievement level thresholds. For example, for tests associated with NCLB, the primary concern is distinguishing between students who are proficient and those who are not yet proficient. In this case, the accuracy of the *Needs Improvement*/*Proficient* threshold is critically important. Table 3-32 provides accuracy and consistency estimates for the 2012 MCAS tests at each cutpoint, as well as false positive and false negative decision rates. (A false positive is the proportion of students whose observed scores were above the cut and whose true scores were below the cut. A false negative is the proportion of students whose observed scores were below the cut and whose true scores were above the cut.)

Table 3-31. 2012 MCAS: Summary of Decision Accuracy (and Consistency) Results by Content Area and Grade—Overall and Conditional on Achievement Level

Content Area	*Grade*	*Overall*	*Kappa*	*Conditional on Achievement Level*			
				*Warning**	*Needs Improvement*	*Proficient*	*Advanced*
ELA	3	0.75 (0.67)	0.51	0.82 (0.74)	0.82 (0.76)	0.72 (0.67)	0.68 (0.53)
	4	0.77 (0.69)	0.55	0.80 (0.69)	0.76 (0.70)	0.74 (0.66)	0.85 (0.71)
	5	0.78 (0.69)	0.56	0.78 (0.66)	0.77 (0.71)	0.75 (0.67)	0.86 (0.73)
	6	0.78 (0.70)	0.55	0.81 (0.72)	0.74 (0.66)	0.78 (0.72)	0.81 (0.67)
	7	0.82 (0.75)	0.6	0.77 (0.63)	0.78 (0.71)	0.83 (0.78)	0.86 (0.73)
	8	0.83 (0.77)	0.61	0.76 (0.61)	0.75 (0.66)	0.84 (0.82)	0.88 (0.75)
	10	0.85 (0.80)	0.65	0.72 (0.49)	0.78 (0.68)	0.84 (0.80)	0.90 (0.82)
Mathematics	3	0.76 (0.68)	0.56	0.81 (0.73)	0.72 (0.64)	0.66 (0.58)	0.90 (0.79)
	4	0.77 (0.68)	0.55	0.81 (0.72)	0.80 (0.74)	0.71 (0.63)	0.80 (0.65)
	5	0.77 (0.68)	0.57	0.82 (0.76)	0.73 (0.65)	0.70 (0.61)	0.88 (0.77)
	6	0.79 (0.71)	0.6	0.83 (0.77)	0.75 (0.67)	0.72 (0.63)	0.89 (0.80)
	7	0.77 (0.69)	0.58	0.82 (0.75)	0.75 (0.68)	0.71 (0.63)	0.88 (0.76)
	8	0.80 (0.73)	0.63	0.84 (0.78)	0.77 (0.69)	0.74 (0.66)	0.90 (0.80)
	10	0.83 (0.77)	0.63	0.76 (0.64)	0.71 (0.61)	0.73 (0.64)	0.93 (0.89)
STE	5	0.75 (0.66)	0.53	0.79 (0.69)	0.76 (0.68)	0.65 (0.56)	0.85 (0.73)
	8	0.79 (0.70)	0.56	0.82 (0.74)	0.76 (0.69)	0.80 (0.74)	0.71 (0.44)
Biology	9–12	0.80 (0.72)	0.6	0.77 (0.68)	0.75 (0.66)	0.79 (0.73)	0.87 (0.77)
Chemistry	9–12	0.77 (0.69)	0.58	0.83 (0.78)	0.70 (0.59)	0.69 (0.59)	0.88 (0.79)
Introductory Physics	9–12	0.79 (0.71)	0.59	0.79 (0.70)	0.75 (0.67)	0.78 (0.72)	0.87 (0.76)
Technology/Engineering	9–12	0.79 (0.70)	0.56	0.81 (0.73)	0.73 (0.65)	0.82 (0.76)	0.79 (0.57)

**Failing* on all high school tests

Table 3-32. 2012 MCAS: Summary of Decision Accuracy (and Consistency) Results by Content Area and Grade—Conditional on Cutpoint

Content Area	*Grade*	*Warning*/Needs Improvement*			*Needs Improvement/Proficient*			*Proficient/Advanced*		
		Accuracy (consistency)	*False Positive*	*False Negative*	*Accuracy (consistency)*	*False Positive*	*False Negative*	*Accuracy (consistency)*	*False Positive*	*False Negative*
ELA	3	0.98 (0.97)	0.01	0.01	0.92 (0.89)	0.04	0.03	0.85 (0.82)	0.11	0.04
	4	0.95 (0.93)	0.02	0.03	0.89 (0.85)	0.06	0.05	0.93 (0.90)	0.05	0.02
	5	0.96 (0.95)	0.02	0.02	0.90 (0.86)	0.06	0.04	0.92 (0.89)	0.06	0.02
	6	0.97 (0.95)	0.02	0.02	0.92 (0.89)	0.04	0.04	0.89 (0.86)	0.07	0.03
	7	0.98 (0.97)	0.01	0.01	0.91 (0.88)	0.05	0.04	0.93 (0.90)	0.05	0.02
	8	0.98 (0.98)	0.01	0.01	0.94 (0.91)	0.03	0.03	0.91 (0.88)	0.06	0.02
	10	1.00 (1.00)	0.00	0.00	0.96 (0.94)	0.02	0.02	0.90 (0.86)	0.06	0.04
Mathematics	3	0.95 (0.93)	0.02	0.02	0.91 (0.87)	0.06	0.04	0.90 (0.87)	0.07	0.03
	4	0.96 (0.94)	0.02	0.02	0.91 (0.87)	0.06	0.04	0.90 (0.87)	0.07	0.03
	5	0.95 (0.93)	0.03	0.02	0.91 (0.88)	0.05	0.03	0.90 (0.87)	0.07	0.03
	6	0.96 (0.94)	0.02	0.02	0.92 (0.89)	0.04	0.03	0.91 (0.87)	0.06	0.03
	7	0.94 (0.92)	0.03	0.03	0.91 (0.88)	0.05	0.03	0.92 (0.89)	0.06	0.02
	8	0.95 (0.93)	0.03	0.02	0.93 (0.90)	0.04	0.03	0.93 (0.90)	0.05	0.02
	10	0.98 (0.96)	0.01	0.01	0.94 (0.92)	0.03	0.03	0.92 (0.88)	0.05	0.03
STE	5	0.95 (0.93)	0.02	0.03	0.90 (0.86)	0.06	0.04	0.90 (0.87)	0.07	0.03
	8	0.93 (0.91)	0.03	0.03	0.90 (0.86)	0.06	0.04	0.95 (0.93)	0.04	0.01
Biology	9–12	0.96 (0.95)	0.02	0.02	0.93 (0.90)	0.04	0.03	0.91 (0.88)	0.06	0.03
Chemistry	9–12	0.93 (0.90)	0.04	0.03	0.93 (0.90)	0.04	0.03	0.92 (0.88)	0.05	0.03
Introductory Physics	9–12	0.95 (0.94)	0.02	0.02	0.92 (0.88)	0.05	0.04	0.92 (0.89)	0.05	0.03
Technology/Engineering	9–12	0.93 (0.90)	0.03	0.03	0.90 (0.86)	0.06	0.04	0.95 (0.94)	0.04	0.01

**Failing* on all high school tests.

The above indices are derived from Livingston and Lewis's (1995) method of estimating DAC. Livingston and Lewis discuss two versions of the accuracy and consistency tables. A standard version performs calculations for forms parallel to the form taken. An "adjusted" version adjusts the results of one form to match the observed score distribution obtained in the data. The tables use the standard version for two reasons: (1) this "unadjusted" version can be considered a smoothing of the data, thereby decreasing the variability of the results; and (2) for results dealing with the consistency of two parallel forms, the unadjusted tables are symmetrical, indicating that the two parallel forms have the same statistical properties. This second reason is consistent with the notion of forms that are parallel (i.e., it is more intuitive and interpretable for two parallel forms to have the same statistical distribution).

As with other methods of evaluating reliability, DAC statistics that are calculated based on small groups can be expected to be lower than those calculated based on larger groups. For this reason, the values presented in Tables 3-31 and 3-32 should be interpreted with caution. In addition, it is important to remember that it is inappropriate to compare DAC statistics across grades and content areas.

3.8 Reporting of Results

The MCAS tests are designed to measure student achievement in the Massachusetts content standards. Consistent with this purpose, results on the MCAS were reported in terms of achievement levels, which describe student achievement in relation to these established state standards. There are four achievement levels: *Warning* (at grades 3–8) or *Failing* (at high school), *Needs Improvement*, *Proficient*, and *Advanced*. Students receive a separate achievement level classification in each content area. Reports are generated at the student level. *Parent/Guardian Reports* and student results labels are printed and mailed to districts for distribution to schools. The details of the reports are presented in the sections that follow. See Appendix Q for a sample *Parent/Guardian Report* and sample student labels.

3.8.1 Unique Reporting Notes

Beginning in 2012, what was once a performance level is now referred to as an achievement level.

3.8.2 Parent/Guardian Report

The *Parent/Guardian Report* is a standalone single page (11 inches by 17 inches) with a centerfold, and it is generated for all students eligible to take the MCAS tests. The front cover provides student identification information, a commissioner's letter to parents, general information about the test, and website information for parent resources. The inside portion contains the achievement level, scaled score, and standard error of the scaled score for each content area tested. If the student does not receive a scaled score, the reason is displayed under the heading "Achievement Level." Historical scaled scores are reported where appropriate and available. An achievement-level summary of school, district, and state results for each content area is reported. The student's growth percentiles in ELA and mathematics are reported if sufficient data exist to calculate growth percentiles. The median growth percentiles for the school and district are also reported. On the back cover, the student's performance on individual test questions is reported, along with a sub-content area summary for each tested content area.

A note is printed on the report if the student took the ELA or Mathematics test with one of the following nonstandard accommodations:

- The ELA reading comprehension test was read aloud to the student.
- The ELA composition was scribed for the student.
- The student used a calculator during the non-calculator session of the Mathematics test.

At the high school level, there is an additional note stating whether a student has met the graduation requirement for each content area, as well as whether the student is required to fulfill an Educational Proficiency Plan (EPP) in order to meet the graduation requirement. EPPs are applicable to ELA and mathematics only. The nonstandard accommodation note and additional high school note appear where the scaled score and achievement level are reported. The growth percentiles for ELA and mathematics (if applicable) are reported along with an explanation of the growth percentile. There are two black-and-white printed copies of the reports: one for the parent and one for the school. Sample reports are provided in Appendix Q.

The front page of the report provides the following student identification information:

- student name
- grade
- birth date
- student ID (SASID)
- school
- district

A student results label is produced for each student receiving a *Parent/Guardian Report*. The following information appears on the label.

- student name
- grade
- birth date
- test date
- student ID (SASID)
- school code
- school name
- district name
- student's scaled score and achievement level (or the reason the student did not receive a score)

One copy of the student labels is shipped with *Parent/Guardian Reports*.

3.8.3 Decision Rules

To ensure that MCAS results are processed and reported accurately, a document delineating decision rules is prepared before each reporting cycle. The decision rules are observed in the analyses of the MCAS test data and in reporting results. These rules also guide data analysts in identifying students to be excluded from school-, district-, and state-level summary computations. Copies of the decision rules are included in Appendix R.

3.8.4 Quality Assurance

Quality assurance measures are implemented throughout the process of analysis and reporting at MP. The data processors and data analysts perform routine quality control checks of their computer programs. When data are handed off to different units within the Data and Reporting Services division (DRS), the sending unit verifies that the data are accurate before handoff. Additionally, when a unit receives a data set, the first step is to verify the accuracy of the data. Once report designs have been approved by the ESE, reports are run using demonstration data to test the application of the decision rules. These reports are then approved by the ESE.

Another type of quality assurance measure used at MP is parallel processing. One data analyst is responsible for writing all programs required to populate the student-level and aggregate reporting tables for the administration. Each reporting table is assigned to a second data analyst who uses the decision rules to independently program the reporting table. The production and quality assurance tables are compared; when there is 100 percent agreement, the tables are released for report generation.

The third aspect of quality control involves procedures to check the accuracy of reported data. Using a sample of schools and districts, the quality assurance group verifies that the reported information is correct. The selection of sample schools and districts for this purpose is very specific because it can affect the success of the quality control efforts. There are two sets of samples selected that may not be mutually exclusive. The first set includes samples that satisfy the following criteria:

- one-school district
- two-school district
- multi-school district
- private school
- special school (e.g., a charter school)
- small school that does not have enough students to report aggregations
- school with excluded (not tested) students

The second set of samples includes districts or schools that have unique reporting situations that require the implementation of a decision rule. This set is necessary in order to ensure that each rule is applied correctly.

The quality assurance group uses a checklist to implement its procedures. Once the checklist is completed, sample reports are circulated for review by psychometric and program management staff. The appropriate sample reports are then sent to the ESE for review and signoff.

3.9 MCAS Validity

One purpose of this report is to describe the technical aspects of the MCAS program that support valid score interpretations. According to the *Standards for Educational and Psychological Testing* (AERA et al., 1999), the sources of evidence that should be considered when constructing a validity argument include: test content, response processes, internal structure, relationship to other variables, and consequences of testing. Thus, as described below, each section of the report (test development and design, test administration, scoring, scaling and equating, item analyses, reliability, and score reporting) contributes to a comprehensive evaluation of validity.

3.9.1 Test Content Validity Evidence

Test content validity demonstrates how well the assessment tasks represent the curriculum and standards for each content area and grade level. Content validation is informed by the item development process, including how the test blueprints and test items align to the curriculum and standards. Viewed through the lens provided by the standards, evidence based on test content is extensively described in Sections 3.2 and 3.3. The following are all components of validity evidence based on test content: item alignment with Massachusetts curriculum framework content standards; item bias, sensitivity, and content appropriateness review processes; adherence to the test blueprint; use of multiple item types; use of standardized administration procedures, with accommodated options for participation; and appropriate test administration training. As discussed earlier, all MCAS items are aligned by Massachusetts educators to specific Massachusetts curriculum framework content standards, and undergo several rounds of review for content fidelity and appropriateness.

3.9.2 Response Process Validity Evidence

Items are presented to students in multiple formats (multiple choice open response, short answer, short response, and writing prompt). The scoring information in Section 3.4 describes the steps taken to train and monitor hand-scorers, as well as quality-control procedures related to scanning and machine scoring. Finally, tests are administered according to state-mandated standardized procedures, and all test administrators are required to attend annual training sessions. Additional studies that might include an investigation of students' cognitive methods using think-aloud protocols could enable stakeholders to develop a more comprehensive understanding of student-response processes.

3.9.3 Internal Structure Validity Evidence

Evidence based on internal structure is presented in great detail in the discussions of item analyses, reliability, and scaling and equating in Sections 3.5 through 3.7. Technical characteristics of the internal structure of the assessments are presented in terms of classical item statistics (item difficulty, item-test correlation), DIF analyses, dimensionality analyses, reliability, SEM, and IRT parameters and procedures. Each test is equated to the previous year's test in that grade and content area in order to preserve the meaning of scores over time. In general, item difficulty and discrimination indices were within acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that most items were assessing consistent constructs, and students who performed well on individual items tended to perform well overall. See the individual sections for more complete results of the different analyses.

In addition to the routine procedures MP provides for evaluating an assessment's internal structure, a set of special studies conducted by the Center for Educational Assessment at the University of Massachusetts Amherst was commissioned by the ESE to provide a multi-year analysis of specific items exhibiting DIF (Clauser & Hambleton, 2011a; Clauser & Hambleton, 2011b). The first study explored items administered on the 2008, 2009, and 2010 grade 8 STE assessments. A similar study was conducted on the 2008, 2009, and 2010 grade 10 ELA assessments. Both studies concluded that any advantages in favor of one subgroup over another were small or nonexistent, thus furthering the validity evidence.

3.9.4 Validity Evidence in Relationships to Other Variables

Massachusetts has accumulated a substantial amount of evidence of the criterion-related validity of the MCAS tests. This evidence shows that MCAS test results are correlated strongly with relevant measures of academic achievement. Specific examples may be found in the *2007 MCAS Technical Report*.

3.9.5 Validity Evidence Based on Consequences of Testing

Evidence based on the consequences of testing is addressed in the scaled score information in Section 3.6.6 and the reporting information in Section 3.8. Each of these sections speaks to the efforts undertaken to provide accurate and clear information to the public regarding test scores. Scaled scores offer the advantage of simplifying the reporting of results across content areas, grade levels, and subsequent years. Achievement levels provide users with reference points for mastery at each grade level. Several different standard reports are provided to stakeholders. In addition, a data analysis tool is provided to each school system to allow educators the flexibility to customize reports for local needs. Additional evidence of the consequences of testing could be supplemented with broader investigation of the impact of testing on student learning.

In summary, the evidence presented in this chapter supports inferences made about student achievement on the content represented in the Massachusetts content standards for ELA, mathematics, and STE. As such, the evidence provided also supports the use of MCAS results for the purposes of program and instructional improvement and as a component of school accountability.

Chapter 4. MCAS-Alt

4.1 Overview

4.1.1 Background

This chapter presents evidence in support of the technical quality of the MCAS Alternate Assessment (MCAS-Alt) and documents the procedures used to administer, score, and report student results on the MCAS-Alt student portfolio. These procedures have been implemented to ensure, to the extent possible, the validity of score interpretations based on the MCAS-Alt. While flexibility is built into the MCAS-Alt to allow teachers to customize academic goals at an appropriate level of challenge for each student, the procedures described in this report are also intended to constrain unwanted variability wherever possible.

For each phase of the alternate assessment process, this chapter includes a separate section that documents how the assessment evaluates the knowledge and skills of students with significant disabilities in the context of grade-level content standards. Together, these sections provide a basis for the validity of the results.

This chapter is intended primarily for a technical audience and requires highly specialized knowledge and a solid understanding of measurement concepts. However, teachers, parents, and the public will also be interested in how the portfolio products both inform and emerge from daily classroom instruction.

4.1.2 Purposes of the Assessment System

The MCAS is the state's program of student academic assessment, implemented in response to the Massachusetts Education Reform Act of 1993. Statewide assessments, along with other components of education reform, are designed to strengthen public education in Massachusetts and to ensure that all students receive challenging instruction based on the standards in the Massachusetts curriculum frameworks. The law requires that the curriculum of all students, including those with disabilities, be aligned with state standards. The MCAS is designed to improve teaching and learning by reporting detailed results to districts, schools, and parents; to serve as the basis, with other indicators, for school and district accountability; and to certify that students have met the Competency Determination (CD) standard in order to graduate from high school. Students with significant disabilities who are unable to take the standard MCAS tests, even if accommodations are provided, are designated by their IEP and 504 teams to take the MCAS-Alt.

The purposes of the MCAS-Alt are to

- determine whether students with significant disabilities are receiving a program of instruction based on the state's academic learning standards;
- determine how much of the academic curriculum has been taught, and what the student has learned;
- include difficult-to-assess students in statewide assessment and accountability systems;

- help teachers provide challenging academic instruction;
- provide an alternative pathway for some students with disabilities to earn a CD and become eligible to receive a diploma.

The MCAS-Alt was developed between 1998 and 2000, and has been refined and enhanced each year since its implementation in 2001.

4.1.3 Format

The MCAS-Alt consists of a structured portfolio of "evidence" collected in each strand and subject required for assessment during the school year. This portfolio documents the student's performance and progress in the skills, knowledge, and concepts outlined in the state's curriculum frameworks. The student portfolio also includes the student's demographic information and weekly schedule, parent verification and sign-off, and a school calendar, which together with the student's "evidence" is submitted to the state each spring. Preliminary results are reported to parents, schools, and the public in June, with final results provided in August.

The ESE's *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities* (2006) describes the content to be assessed and provides strategies for adapting and using the state's learning standards to instruct and assess students taking the MCAS-Alt.

4.2 Test Design and Development

4.2.1 Test Content

MCAS-Alt assessments are required for all grades and content areas in which standard MCAS tests are administered, although the range and level of complexity of the standards being assessed is somewhat diminished. Specific MCAS-Alt requirements for students in each grade level are listed in Table 4-1.

Table 4-1. 2012 MCAS-Alt: Requirements

Grade	*ELA Strands Required*	*Mathematics Strands Required*	*STE Strands Required*
3	• Language (General Standard 4) • Reading and Literature (General Standard 8)	• Number Sense and Operations • Patterns, Relations, and Algebra	
4	• Language (General Standard 4) • Reading and Literature (General Standard 8) • Composition	• Number Sense and Operations • Data Analysis, Statistics, and Probability	
5	• Language (General Standard 4) • Reading and Literature (General Standard 8)	• Number Sense and Operations • Measurement	Any three of the four STE strands*
6	• Language (General Standard 4) • Reading and Literature (General Standard 8)	• Number Sense and Operations • Patterns, Relations, and Algebra	

continued

Grade	*ELA Strands Required*	*Mathematics Strands Required*	*STE Strands Required*
7	• Language (General Standard 4) • Reading and Literature (General Standard 8) • Composition	• Number Sense and Operations • Data Analysis, Statistics, and Probability	
8	• Language (General Standard 4) • Reading and Literature (General Standard 8)	• Number Sense and Operations • Geometry	Any three of the four STE strands*
10	• Language (General Standard 4) • Reading and Literature (General Standard 8) • Composition	• Any three of the five mathematics strands	Any three learning standards in either • Biology • Chemistry • Introductory Physics or • Technology/ Engineering

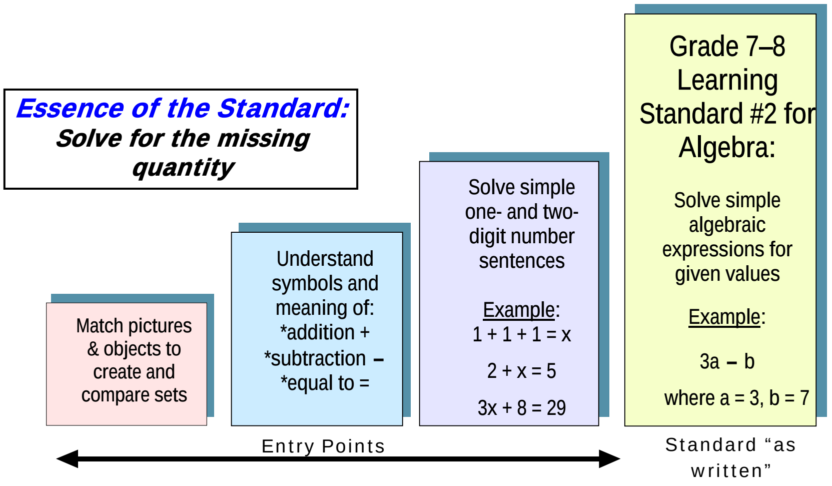
* Earth and Space Science, Life Science, Physical Science, Technology/Engineering

4.2.1.2 Access to the Grade-Level Curriculum

The *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities* is used to determine appropriate curriculum goals based on the curriculum frameworks at each grade level that engage and challenge each student, as shown in Figure 4-1.

Most students with significant disabilities can access the "essence" of each learning standard by addressing one of several entry points listed in the Resource Guide. Entry points are outcomes based on grade-level content for which the level of complexity has been modified below grade-level expectations. A small number of students with the most complex and significant disabilities may not yet be ready to address academic content through entry points, even at the lowest levels of complexity. These students will instead focus on targeted social, communication, or motor skills (access skills) practiced during academic activities that expose them to the tools, materials, and academic content. For example, a student may practice operating an electronic switch on cue to indicate whose turn is next during a mathematics activity; or reach, grasp, and release the materials being used during a physical science activity; or focus on a story read aloud for increasing periods of time during English language arts (ELA). Figure 4-1 shows a mathematics example of access to the general curriculum through entry points that address the essence of the standard.

Figure 4-1. 2012 MCAS-Alt: Access to the General Curriculum (Mathematics Example) through Entry Points that Address the Essence of the Standard


Essence of the Standard:
Solve for the missing quantity
Match pictures & objects to create and compare sets
Understand symbols and meaning of: *addition + *subtraction – *equal to =
Solve simple one- and two-digit number sentences
Example: 1 + 1 + 1 = x
2 + x = 5
3x + 8 = 29
Grade 7–8 Learning Standard #2 for Algebra:
Solve simple algebraic expressions for given values
Example:
3a – b
where a = 3, b = 7
Entry Points
Standard "as written"

4.2.1.3 Assessment Design

The MCAS-Alt portfolio consists of primary evidence, supporting documentation, and other required information.

Primary Evidence

Portfolios must include three or more pieces of primary evidence in each strand being assessed.

One of the three must be a data chart (e.g., field data chart, line graph, bar graph) that includes the following information:

- the targeted skill based on the learning standard being assessed
- tasks performed by the student on at least eight distinct dates, with a brief description of each activity
- percentage of accuracy for each performance
- percentage of independence for each performance
- progress over time, indicating that the student has attempted a new skill

Two or more additional pieces of primary evidence must document the student's performance of the same skill or outcome identified on the data chart, and may include either

- work samples,
- photographs, or
- audio or video clips.

Each piece of primary evidence must be labeled with

- the student's name,
- the date of the activity,
- the percentage of accuracy for the performance, and
- the percentage of independence for the performance.

The data chart and at least two additional pieces of primary evidence comprise the “core set of evidence” required in each portfolio strand.

Supporting Documentation

In addition to the required pieces of primary evidence, supporting documentation (described in Section 4.2.1.4) may be included at the discretion of the teacher to indicate the context in which the activity was conducted. Supporting documentation may include any of the following:

- **narrative descriptions** by the teacher or parent describing how the task or activity was conducted or what the student was asked to do
- **photographs** of the student that show how the student engaged in the instructional activity (i.e., the context of the activity)
- **tools, templates, or examples** used by the student
- **reflection sheet or other self-evaluation** documenting the student’s awareness, perceptions, choice, decision-making, and self-assessment of work he or she created, and the learning that occurred as a result. For example, a student may respond to questions such as:
    - What did we do? What did I learn?
    - What did I do well? What am I good at?
    - Did I correct my inaccurate response?
    - How could I do better? Where do I need help?
    - What should I work on next? What would I like to learn?
- **letters of support** or notes from employers, counselors, after-school program supervisors, community service providers, peers, or parents
- **work description labels** providing a brief description of the activity or work sample

4.2.1.4 Assessment Dimensions (Scoring Rubric Areas)

The Rubric for Scoring Portfolio Strands is used to generate a score in each portfolio strand based in each rubric area: Level of Complexity (score range of 1–5); Demonstration of Skills and Concepts (M, 1–4); Independence (M, 1–4); Self-Evaluation (M, 1, 2); and Generalized Performance (1, 2). A score of “M” means there was insufficient evidence or information to generate a numerical score in a rubric area.

Trained and qualified scorers examine each piece of evidence in the strand and apply criteria described in the Guidelines for Scoring Student Portfolios (available at http://www.doe.mass.edu/mcas/alt/results.html) to produce a score in each rubric area. Scores are based on the following:

- **completeness** of portfolio materials
- **level of complexity** at which the student addressed learning standards in the Massachusetts curriculum frameworks in the content area being assessed
- **accuracy** of the student’s responses or performance of specific tasks
- **independence** demonstrated by the student in responding to questions or performing tasks
- **self-evaluation** during or after each task or activity (e.g., reflection, self-correction, goal-setting)

- **generalized performance** of the skill in different instructional contexts, or using different materials or methods of presentation or response

4.2.1.5 MCAS-Alt Grade-Level and Competency Portfolios

All students, including students with disabilities, are required to meet the CD standard to be eligible to earn a high school diploma. Students must attain a score of *Proficient* or higher on the English Language Arts and Mathematics MCAS tests (or *Needs Improvement*, plus fulfilling the requirements of an Educational Proficiency Plan) and a minimum score of *Needs Improvement* on a high school Science and Technology/Engineering (STE) test. Massachusetts allows students with disabilities who take alternate assessments to meet the graduation requirement, provided they can demonstrate in their MCAS-Alt portfolio a level of performance equivalent to a student who has achieved these scores on the MCAS tests. Since students with significant cognitive disabilities comprise the majority of students taking alternate assessments, the proportion of students who will achieve a score of *Needs Improvement* will likely remain low in comparison to the number of students who meet the CD requirement by taking standard MCAS tests.

A small number of MCAS-Alt grade-level portfolios (for students in grades 3–8) and competency portfolios (for high school students) are submitted each year for students who address learning standards at or near grade-level expectations but are unable to participate in standard MCAS testing, even with accommodations. The Participation Guidelines section of the *2012 Educator's Manual for MCAS-Alt* (available at http://www.doe.mass.edu/mcas/alt/edmanual.pdf) describes and profiles those students who may be considered for the MCAS-Alt, and for whom it is appropriate to submit grade-level and competency portfolios.

MCAS-Alt competency portfolios in ELA, mathematics, and STE include a larger, broader collection of work samples than the typical MCAS-Alt portfolio and are evaluated by panels of content experts to ensure that they meet the appropriate standard of performance in that subject.

For additional information on how grade-level and competency portfolios were evaluated, see Section 4.4.4 of this report.

4.2.2 Test Development

4.2.2.1 Rationale

IEP and 504 teams are directed to consider *how*, not *whether*, students with disabilities will participate in MCAS. Students with disabilities may either take MCAS tests, with or without accommodations, or participate in an alternate assessment if they are unable to take the standard tests because of the severity of their disabilities. Alternate assessment is the component of the state's assessment system that measures the academic performance of students with the most significant disabilities. Students with disabilities are required by federal and state laws to participate in the MCAS so that their performance of skills and knowledge of content described in the state's curriculum frameworks can be assessed, and so they can be visible and accountable in reports of results for each school and district.

The requirement that students with significant disabilities participate in alternate assessments ensures that these students have an opportunity to "show what they know" and to receive instruction at a level that is challenging but attainable. Alternate assessment results provide accurate and detailed feedback that can be used to identify challenging instructional goals for each student. When schools are held accountable for the performance of students with disabilities, these students are more likely to receive consideration when school resources are allocated.

Through use of the curriculum resources provided by the ESE, teachers have become adept at providing standards-based instruction at a level that challenges and engages each student, and report unanticipated gains in student performance.

4.2.2.2 Role of Advisory Committee

An MCAS-Alt Advisory Committee meets twice annually to discuss policy issues related to the alternate assessment. This diverse group of stakeholders—including teachers, parents, advocates, principals, private school and educational collaborative directors, special education directors and supervisors, and representatives of institutions of higher education—has been critical in the development, implementation, and continued enhancement of the MCAS-Alt. A list of advisory committee members is provided in Appendix A.

4.3 Test Administration

4.3.1 Instructional Data Collection

Each portfolio strand must include a data chart documenting the student's performance and progress in learning a new academic skill. Data must be collected on at least eight different dates in order to determine whether progress has been made and the degree to which the skill has been mastered. On each date, the data point must indicate how often a correct response was given (an overall percentage of accuracy on that date) and how often the student required cues or prompts (overall percentage of independence). Data is collected either during routine classroom instruction or during tasks and activities set up specifically for the purpose of assessing the student. All data charts must include a brief description of the activity (or activities) conducted on each date, describing how the task relates to the measurable outcome being assessed. Data charts may include performance data from a collection of work samples or from responses to specific tasks.

A Collection of Work Samples

The percentage of accuracy and independence of the student's responses on a given date can be charted for individual work samples or summarized for several work samples on each date, provided that all work is based on the same measurable outcome.

Responses to Specific Tasks

The percentage of accuracy and independence of the student's responses on each date can be charted for each activity, task, or trial, provided these are based on the same measurable outcome. All data recorded on a single date must be summarized and averaged for overall percentage of accuracy and independence for each date.

4.3.2 Construction of Portfolios

The student's MCAS-Alt portfolio must include all elements listed below. Required forms may either be photocopied from those found in the *2012 Educator's Manual for MCAS-Alt* or completed electronically using an online MCAS-Alt Forms and Graphs program available at www.doe.mass.edu/mcas/alt/resources.html.

- **artistic cover** designed and produced by the student and inserted in the front window of the three-ring portfolio binder (recommended but not required)
- **portfolio cover sheet** containing important information about the student
- **student's introduction to the portfolio** produced as independently as possible by the student using his or her primary mode of communication (e.g., written, dictated, or recorded on video or audio) describing "What I want others to know about me as a learner and about my portfolio"
- **verification form** signed by a parent, guardian, or primary care provider signifying that he or she has reviewed the student's portfolio or, at minimum, was invited to do so (In the event no signature was obtained, the school must include a record of attempts to invite a parent, guardian, or primary care provider to view the portfolio.)
- **signed consent form to photograph or audio/videotape a student** (kept on file at the school) if images or recordings of the student are included in the portfolio
- **weekly schedule** documenting the student's program of instruction, including participation in the general academic curriculum
- **school calendar** indicating dates in the current academic year on which the school was in session
- **strand cover sheet** describing the accompanying set of evidence addressing a particular outcome
- **product description** attached to each piece of primary evidence providing required labeling information (If product description labels are not used, this information must be written directly on each piece.)

The contents listed above, plus all evidence and other documentation, comprise the student's portfolio and are placed inside a white, three-ring plastic binder provided by the ESE for each student.

4.3.3 Participation Requirements

4.3.3.1 Identification of Students

All students educated with public funds, including students with disabilities educated inside or outside their home districts, must be engaged in an instructional program guided by the standards in the Massachusetts curriculum frameworks and must participate in assessments that correspond with the grades in which they are reported in the ESE's Student Information Management System (SIMS). Students with significant disabilities who are unable to take the standard MCAS tests, even with accommodations, must take the MCAS-Alt, as determined by the student's IEP Team.

A student with a disability may participate in the MCAS-Alt regardless of whether he or she has an IEP provided under the Individuals with Disabilities Education Act or a plan provided under Section 504 of the Rehabilitation Act of 1973.

4.3.3.2 Participation Guidelines

A student's IEP Team or 504 team determines how the student will participate in the MCAS for each content area scheduled for assessment, either by taking the test routinely or with accommodations, or by taking the alternate assessment. This information is documented in the student's IEP or 504 plan and must be revisited on an annual basis. A student may take the general assessment, with or without accommodations, in one subject, and the alternate assessment in another subject.

The student's team must consider the following questions each year for each content area scheduled for assessment:

- Can the student take the standard MCAS test under routine conditions?
- Can the student take the standard MCAS test with accommodations? If so, which accommodations are necessary for the student to participate?
- Does the student require an alternate assessment? (Alternate assessments are intended for a very small number of students with significant disabilities who are unable to take standard MCAS tests, even with accommodations.)

A student's team must review the options provided below.

Figure 4-2. 2012 MCAS-Alt: Participation Guidelines

Characteristics of Student's Instructional Program and Local Assessment	Recommended Participation in MCAS

OPTION 1

If the student is	*Then*
a) generally able to demonstrate knowledge and skills on a paper-and-pencil test, either with or without test accommodations; ***and*** is b) working on learning standards at or near grade-level expectations; ***or*** is c) working on learning standards that have been modified and are somewhat below grade-level expectations due to the nature of the student's disability,	the student should take the **standard MCAS test**, either under routine conditions or with accommodations that are generally consistent with the instructional accommodation(s) used in the student's educational program (according to the ESE's accommodations policy available at www.doe.mass.edu/mcas/participation/sped.pdf) and that are documented in an approved IEP or 504 plan prior to testing.

Characteristics of Student’s Instructional Program and Local Assessment	Recommended Participation in MCAS

OPTION 2

If the student is	***Then***
a) **generally unable** to demonstrate knowledge and skills on a paper-and-pencil test, even with accommodations; ***and*** is b) working on learning standards that have been **substantially modified** due to the nature and severity of his or her disability; ***and*** is c) receiving **intensive, individualized instruction** in order to acquire, generalize, and demonstrate knowledge and skills,	the student should take the **MCAS Alternate Assessment** (MCAS-Alt) in this content area.

Characteristics of Student’s Instructional Program and Local Assessment	Recommended Participation in MCAS

OPTION 3

If the student is	***Then***
a) working on learning standards at or near grade-level expectations; ***and*** is b) **sometimes able** to take a paper-and-pencil test, either without accommodations or with one or more accommodation(s); ***but*** c) has a complex and significant disability that does not allow the student to fully demonstrate knowledge and skills on a test of this format and duration, (Examples of complex and significant disabilities for which the student may require an alternate assessment are provided below.)	the student should take the **standard MCAS test**, if possible, with necessary accommodations that are consistent with the instructional accommodation(s) used in the student’s instructional program (according to the ESE’s accommodations policy) and that are documented in an approved IEP or 504 plan prior to testing. ***However,*** the team may recommend the MCAS-Alt when the nature and complexity of the disability prevent the student from fully demonstrating knowledge and skills on the standard test, even with the use of accommodations. In this case, the MCAS-Alt “grade-level” portfolio should be compiled and submitted.

While the majority of students who take alternate assessments have significant *cognitive* disabilities, participation in the MCAS-Alt is not limited to these students. When the nature and complexity of a student's disability present significant barriers or challenges to standardized testing, even with the use of accommodations, although the student may be working at or near grade-level expectations, the student's IEP or 504 team may determine the student should take the MCAS-Alt.

In addition to the criteria outlined in Options 2 and 3, the following examples of unique circumstances are provided to expand the team's understanding of the appropriate use of alternate assessments. An alternate assessment may be administered, for example, in each of the following situations:

- A student with a severe emotional, behavioral, or other disability is unable to maintain sufficient concentration to participate in standard testing, even with test accommodations.
- A student with a severe health-related disability, neurological disorder, or other complex disability is unable to meet the demands of a prolonged test administration.
- A student with a significant motor, communication, or other disability requires more time than is reasonable or available for testing, even with the allowance of extended time (i.e., the student cannot complete one full test session in a school day).

4.3.3.3 MCAS-Alt Participation Rates

Across all content areas, a total of 9,368 students, or 1.7% of the assessed population, participated in the 2012 MCAS-Alt in grades 3–10. A slightly higher relative proportion of students in grades 3–8 took the MCAS-Alt compared with students in grade 10, and slightly more students were alternately assessed in mathematics than in ELA. Additional information about MCAS-Alt participation rates by content area is provided in Appendix B, including the comparative rate of participation in each MCAS assessment format (i.e., routinely tested, tested with accommodations, or alternately assessed).

4.3.4 Educator Training

During the month of October 2011, a total of 2,600 educators received training for the 2012 MCAS-Alt. Educators attending the training had the option of attending one of two sessions: an overview for educators new to the MCAS-Alt process or an update for those with previous MCAS-Alt experience. Topics for the overview session included the following:

- decision-making for which students should take the MCAS-Alt
- portfolio requirements in each grade and content area
- collecting data on student performance and progress on measurable outcomes
- developing measurable outcomes using the *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities* (fall 2006)

Topics for the update session included the following:

- statewide 2011 MCAS-Alt results

- changes to the MCAS-Alt requirements for 2012
- where to find information in the *2012 Educator's Manual for MCAS-Alt*
- avoiding mistakes that lead to scores of Incomplete
- changes in reporting results and determining Adequate Yearly Progress (AYP)
- data collection process (step-by-step)
- using data charts to improve teaching and learning
- competency and grade-level portfolio requirements

During January 2012, a total of 1,194 educators received MCAS-Alt training: some were new to the process and did not attend the overview training in the fall; others wished to ask MCAS-Alt training specialists (i.e., expert teachers) specific questions about their portfolios-in-progress.

During March 2012, an additional 892 educators attended training, where they were able to review and discuss their students' portfolios and have their questions answered by expert teachers.

4.3.5 Support for Educators: the MCAS Service Center

ESE staff provided assistance throughout the year via email and telephone to educators with specific questions about their portfolios. Additionally, the MCAS Service Center provided toll-free telephone support to district and school staff regarding test administration, reporting, training, materials, and other relevant operations and logistics.

The Measured Progress (MP) project management team provided extensive training to the MCAS Service Center staff on the logistical, programmatic, and content-specific aspects of the MCAS-Alt. Training materials included screen shots of all Web-based applications used by the districts and schools, principal and test administrator manuals, and memoranda sent to the field. Informative scripts were written by the Service Center coordinator and approved by the ESE for all communications with the field. These scripts covered all activities handled by the Service Center such as Web support, enrollment inquiries, and discrepancy follow-up and resolution procedures.

4.4 Scoring

Portfolios were scored in Dover, New Hampshire, during April and May 2012. The ESE and MP closely monitored scorers to ensure that portfolio scores were accurate.

Evidence of the student's performance was evaluated and scored using research-based criteria for how students with significant disabilities learn and demonstrate knowledge and skills. The criteria included the application of a universal scoring rubric; verification that measurable outcomes were aligned with the standards required for assessment in the *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities* (fall 2006); and rigorous training and qualification of scorers based on the *2012 Guidelines for Scoring MCAS-Alt Portfolios*. The *MCAS-Alt Rubric for Scoring Portfolio Strands* was developed with assistance from teachers and the statewide advisory committee. The criteria for scoring portfolios are listed and described in detail on the following pages.

MCAS-Alt portfolios reflect the degree to which a student has learned and applied the knowledge and skills outlined in the Massachusetts curriculum frameworks. The portfolio measures progress over time, as well as the highest level of achievement attained by the student on the assessed skills, and takes into account the degree to which cues, prompts, and other assistance were required by the student.

4.4.1 Scoring Logistics

MCAS-Alt portfolios were reviewed and scored by trained scorers according to the procedures described in this section. Scores were entered into a computer based scoring system designed by MP and the ESE; scores were monitored for accuracy and completeness.

Security was maintained at the scoring site; access to unscored portfolios and completed score forms was restricted to ESE and MP staff. MCAS-Alt scoring leadership staff included several floor managers (FMs) and table leaders (TLs). Each TL managed a table with four to five scorers. The FM managed a group of tables at either elementary, middle, or secondary levels.

Communication and coordination among scorers were maintained through daily meetings with TLs to ensure that critical information and scoring rules were implemented across all grade clusters.

4.4.2 Selection, Training, and Qualification of Scorers

Selection of Training Materials

The MCAS-Alt Project Leadership Team (PLT) included ESE and MP staff, plus four teacher consultants. The PLT met for two days in July 2011 to accomplish the following:

- select sample portfolio strands to use for training, calibration, and qualification of scorers
- discuss issues to be addressed in the 2012 Guidelines for Scoring Student Portfolios

On the first day, the group reviewed and scored approximately 200 portfolios using the draft of the 2012 guidelines, noting any scoring problems that arose during the review. All concerns were resolved by using the *2012 Educator's Manual for MCAS-Alt* or by following additional scoring rules agreed upon by the PLT and subsequently addressed in the final 2012 guidelines.

Of the 200 portfolios reviewed, 96 sample strands were set aside as possible exemplars to train and calibrate scorers. These strands consisted of solid examples of each score point on the scoring rubric.

Each of these samples was triple-scored. Of the 96 triple-scores, 65 were in exact agreement in all five scoring dimensions: Level of Complexity, Demonstration of Skills and Concepts, Independence, Self-Evaluation, and Generalized Performance.

Of these 65 sample strands, the PLT decided to use 20, including several complete content areas, for scorer training and calibration. These 20 portfolio samples became the scorers' "sample set."

Recruitment and Training of Scorers

Recruitment

Through Kelly Services, MP recruited 148 prospective scorers and TLs for the MCAS-Alt Scoring Center. All TLs and many scorers had worked previously on scoring projects for other states' test or alternate assessment administrations, and all had four-year college degrees. Additionally, the PLT recruited 10 Massachusetts educators who had previously served as TLs or scorers to assist the ESE and MP.

Training

Scorers were rigorously trained in all rubric areas and score points by reviewing scoring rules and "mock scoring" of numerous sample portfolio strands selected to illustrate examples of each rubric score point. Scorers were given detailed instructions on how to review data charts and other primary evidence in order to tally the rubric area scores using a strand organizer. Scorers were taught to apply the scoring rubric to the information tallied on the strand organizer in order to arrive at overall scores for Level of Complexity, Demonstration of Skills and Concepts, Independence, Self-Evaluation, and Generalized Performance (see Section 4.4.3). Trainers facilitated discussions and review among scorers to clarify the rationale for each score point and describe special scoring scenarios and exceptions to the general scoring rules.

Scorer Qualification

Before scoring actual student portfolios, each scorer was required to take a qualifying assessment consisting of 24 questions and score a sample portfolio consisting of four strands (i.e., 20 scoring dimensions). The threshold score to qualify as a scorer on the 24 questions was 85% (21 correct out of 24 total questions); the threshold score to qualify as a scorer on the portfolio strands was 85% exact agreement overall for the five scoring dimensions (i.e., exact agreement on 17 out of 20 scorable dimensions for the four strands).

Scorers who did not achieve the required percentage of correct responses on the qualifying assessment were retrained using another qualifying assessment. Those that achieved an accurate response rate of at least 85% exact agreement were authorized to begin scoring student portfolios. If a scorer did not meet the required accuracy rate on the second qualifying assessment, he or she was released from scoring.

Recruitment, Training, and Qualification of TLs and FMs

TLs were recruited, trained, and qualified by the ESE using the same methods and criteria used to qualify scorers, except they were required to achieve a score of 90% correct or higher on the qualifying test. TLs and FMs also received training in logistical, managerial, and security procedures.

Ten licensed Massachusetts educators who had led a table during the previous year's scoring institute were designated as M-resolvers. M-resolvers assisted in the training of new TLs and performed resolution scores on portfolios with scores of M (indicating that evidence was missing or insufficient to determine a score).

The scoring room was monitored by two FMs, who were licensed Massachusetts educators, as well as MCAS-Alt teacher consultants who had served as FMs the previous year.

4.4.3 Scoring Methodology

Guided by a TL, scorers worked at tables with four or five other scorers, all scoring portfolios at the same grade. TLs were experienced scorers who qualified at a higher threshold and who had received additional training on logistics at the scoring center. Scorers were permitted to ask TLs questions as they reviewed portfolios. In the event a TL could not answer a question, the FM provided assistance. In the event the FM was unable to answer a question, ESE staff were available to provide clarification.

Scorers were randomly assigned a portfolio by their TL. Scorers first ensured that the required strands for each grade were submitted. Then, each strand was scored individually. A strand was considered complete if it included a data chart with at least eight different dates related to the same measurable outcome, and two additional pieces of evidence based on the same outcome.

Once the completeness of the portfolio was verified, each strand was scored in the following dimensions:

A. Level of Complexity
B. Demonstration of Skills and Concepts
C. Independence
D. Self-Evaluation
E. Generalized Performance

To assist in scoring, scorers used a worksheet called the strand organizer to record information and keep track of each piece of evidence. By completing the strand organizer, the scorer was able to perform the necessary calculations and determine the final scores in each rubric area without having to review the portfolio a second or third time.

The MCAS-Alt 2012 score distributions for all scoring dimensions are provided in Appendix F.

A. Level of Complexity

The score for Level of Complexity reflects at what level of difficulty (i.e., complexity) the student addressed curriculum framework learning standards (i.e., at grade level, through entry points, or using access skills). Using the *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities* (fall 2006), scorers confirmed that the student’s measurable outcomes were aligned with the intended learning standard and, if so, whether the evidence was addressed at grade-level performance expectations, was modified below grade-level expectations (“entry points”), or was addressed through skills in the context of an academic instructional activity (“access skills”).

Each strand was given a Level of Complexity score based on the scoring rubric for Level of Complexity (Table 4-2) that incorporates the criteria listed above.

Table 4-2. 2012 MCAS-Alt: Scoring Rubric for Level of Complexity

Score Point				
1	*2*	*3*	*4*	*5*
Portfolio strand reflects little or no basis in, or is unmatched to, curriculum framework learning standard(s) required for assessment.	Student primarily addresses social, motor, and communication "access skills" during instruction based on curriculum framework learning standards in this strand.	Student addresses curriculum framework learning standards that have been modified below grade-level expectations in this strand.	Student addresses a narrow sample of curriculum framework learning standards (1 or 2) at grade-level expectations in this strand.	Student addresses a broad range of curriculum framework learning standards (3 or more) at grade-level expectations in this strand.

B. Demonstration of Skills and Concepts

Each strand is given a score for Demonstration of Skills and Concepts based on the degree to which a student gave a correct (accurate) response in demonstrating the targeted skill.

Scorers confirmed that a "core set of evidence" was submitted and that all portfolio evidence was correctly labeled with the following information:

- the student's name
- the date of performance
- the percentage of accuracy
- the percentage of independence

If evidence was not labeled correctly, or if the minimum required pieces of evidence did not address the measurable outcome stated on the Strand Cover Sheet or work description, that piece was not scorable.

Brief descriptions of each activity on the data chart were also considered in determining the completeness of a data chart. Educators had been instructed during educator training workshops and in the *2012 Educator's Manual for MCAS-Alt* that "each data chart must include a brief description beneath each data point that clearly illustrates how the task or activity relates to the measurable outcome being assessed." One- or two-word descriptions were likely considered insufficient to document the relationship between the activity and the measurable outcome and therefore excluded those data points from being scored.

A score of M (i.e., evidence was missing or was insufficient to determine a score) was given in both Demonstration of Skills and Concepts and Independence if at least two pieces of scorable primary evidence and a completed data chart documenting the student's performance of the same skill were not submitted (see section C).

A score of M was also given if

- the data chart listed the percentages of both accuracy and independence at or above 80% for the duration of the data collection period, indicating that the student did not learn a challenging new skill in the strand;
- the data chart did not document a single measurable outcome based on the required learning standard or strand on at least eight different dates, and did not indicate the student's accuracy and independence on each task or trial;
- two additional pieces of primary evidence did not address the same measurable outcome as the data chart, or were not labeled with all required information.

If a "core set of evidence" was submitted in a strand, it was scored for Demonstration of Skills and Concepts by first identifying the "final 1/3 time frame" during which data was collected on the data chart (or the final three data points on the chart, if fewer than 12 points were listed).

Then, an average percentage was calculated based on the percentage of accuracy for

- all data points in the final 1/3 time frame of the data chart, and
- all other primary evidence in the strand produced during or after the final 1/3 time frame.

Based on the average percentage of the data points and evidence, the overall score in the strand was determined using the rubric shown in Table 4-3.

Table 4-3. 2012 MCAS-Alt: Scoring Rubric for Demonstration of Skills and Concepts

Score Point				
M	*1*	*2*	*3*	*4*
The portfolio strand contains insufficient information to determine a score.	Student's performance is primarily inaccurate and demonstrates minimal understanding in this strand (0–25% accurate).	Student's performance is limited and inconsistent with regard to accuracy and demonstrates limited understanding in this strand (26–50% accurate).	Student's performance is mostly accurate and demonstrates some understanding in this strand (51–75% accurate).	Student's performance is accurate and is of consistently high quality in this strand (76–100% accurate).

C. Independence

The score for Independence shows the degree to which the student responded without cues or prompts during tasks or activities based on the measurable outcome being assessed.

For strands that included a "core set of evidence," Independence was scored first by identifying the final 1/3 time frame on the data chart (or the final three data points, if fewer than 12 points were listed).

Then an average percentage was calculated based on the percent of independence for

- all data points during the final 1/3 time frame of the data chart, and
- all other primary evidence in the strand produced during or after the final 1/3 time frame.

Based on the average of the data points and evidence, the overall score in the strand was then determined using the rubric shown in Table 4-4 below.

A score of M (i.e., evidence was missing or was insufficient to determine a score) was given in both Demonstration of Skills and Concepts and Independence if at least two pieces of scorable primary evidence and a completed data chart documenting the student's performance of the same skill were not submitted.

A score of M was also given if

- the data chart listed the percentages of both accuracy and independence at or above 80% for the duration of the data collection period, indicating that the student did not learn a challenging new skill in the strand;
- the data chart did not document a single measurable outcome based on the required learning standard or strand on at least eight different dates, and did not indicate the student's accuracy and independence on each task or trial;
- two additional pieces of primary evidence did not address the same measurable outcome as the data chart, or were not labeled with all required information.

Table 4-4. 2012 MCAS-Alt: Scoring Rubric for Independence

Score Point				
M	*1*	*2*	*3*	*4*
The portfolio strand contains insufficient information to determine a score.	Student requires extensive verbal, visual, and physical assistance to demonstrate skills and concepts in this strand (0–25% independent).	Student requires frequent verbal, visual, and physical assistance to demonstrate skills and concepts in this strand (26–50% independent).	Student requires some verbal, visual, and physical assistance to demonstrate skills and concepts in this strand (51–75% independent).	Student requires minimal verbal, visual, and physical assistance to demonstrate skills and concepts in this strand (76–100% independent).

D. Self-Evaluation

The score for Self-Evaluation indicates the frequency of activities that involve self-correction, task-monitoring, goal-setting, reflection, and overall awareness by the student of his or her own learning. The 2012 MCAS-Alt score distributions for Self-Evaluation are provided in Appendix F.

Each strand was given a score of M, 1, or 2+ based on the scoring rubric shown in Table 4-5.

Table 4-5. 2012 MCAS-Alt: Scoring Rubric for Self-Evaluation, Individual Strand Score

Score Point		
M	*1*	*2+*
Evidence of self-correction, task-monitoring, goal-setting, and reflection was **not found** in the student's portfolio in this content area.	Student infrequently self-corrects, monitors, sets goals, and reflects in this content area—only **one example of self-evaluation** was found in this strand.	Student frequently self-corrects, monitors, sets goals, and reflects in this content area—**multiple examples** of self-evaluation were found in this strand.

Beginning in 2012, scores for Self-Evaluation were no longer reported as a combined score for the content area, but were reported instead for each strand.

E. Generalized Performance

The score for Generalized Performance reflected the number of contexts and instructional approaches used by the student to demonstrate knowledge and skills in the portfolio strand.

Each strand was given a score of either 1 or 2+ based on the rubric shown in Table 4-6.

Table 4-6. 2012 MCAS-Alt: Scoring Rubric for Generalized Performance

Score Point	
1	*2+*
Student demonstrates knowledge and skills in **one** context, or uses **one** approach and/or method of response and participation **in this strand.**	Student demonstrates knowledge and skills in **multiple** contexts, or uses **multiple** approaches and/or methods of response and participation **in this strand.**

As with Self-Evaluation, scores for Generalized Performance were not reported in 2012 as a combined score for the content area (as they had been previously) but were instead reported for each strand.

4.4.4 Monitoring the Scoring Quality

The FM monitored scoring consistency and the general flow of work in the scoring room. The TL ensured that scorers at his or her table were consistent and accurate in their scoring.

Scoring consistency and accuracy were maintained using the following methods:

- double-scoring
- read-behind scoring
- scorer tracking forms

Double-Scoring

Double-scoring means that a portfolio was scored by two scorers at different tables, without knowledge by either scorer of the score assigned by the other.

All portfolios for students in grades 9–12 were double-scored. At least one of the portfolios of each scorer in grades 3–8 was double-scored each morning and afternoon, and at least every fifth portfolio each scorer scored thereafter was double-scored. At least 20% of portfolios for students in grades 3–8 were double-scored.

The required rate of scoring accuracy for double-scored portfolios was 80% exact agreement. When there was a discrepancy between scores, the TL scored the portfolio a third time and that score became the score of record. The TL retrained the scorer if interrater consistency fell below 80% agreement with the TL's resolution score. The TL discussed discrepant scores with the responsible scorers and determined when they could resume scoring.

Table 4-10 in Section 4.6.3 shows the percentages of interrater agreement for the 2012 MCAS-Alt.

Read-Behind Scoring

Read-behind scoring refers to a TL rescoring a portfolio and comparing his or her score with the one assigned by the previous scorer. If there was exact score agreement, the first score was retained as the score of record. If the scores differed, the TL's score became the score of record.

Read-behinds were performed on every scorer's first three portfolios. If those scores were consistent with the TL's resolution scores, the scorer was allowed to continue scoring. A read-behind was performed at least once each morning, once each afternoon, and on every fifth subsequent portfolio per scorer.

If a scorer's first three portfolio scores were inconsistent with the TL's resolution scores, the scorer was retrained. The TL determined when a retrained scorer could resume scoring. Additionally, a read-behind was performed on each subsequent portfolio for any scorer permitted to resume scoring, until consistency with the TL's scores was established.

The required rate of agreement for read-behinds (after the first three portfolios) was 80% exact agreement.

Scorer Tracking Forms

The TL maintained both a daily and a cumulative Scorer Tracking Form for each scorer. The daily form showed the number of portfolios scored by that scorer each day, along with the scorer's percentage of accuracy on read-behinds and double-scores.

In addition to maintaining a record of scorers' accuracy and consistency over time, leadership also monitored scorers for output, with slower scorers remediated to increase their production. The scores were entered into a daily report, which showed the daily as well as the cumulative accuracy and productivity for each scorer.

4.4.5 Scoring of Grade-Level Portfolios in Grades 3 through 8 and Competency Portfolios in High School

Specific requirements for submission of grade-level and competency portfolios are described in the *2012 Educator's Manual for MCAS-Alt.*

Grade-Level Portfolios in Grades 3 through 8

Each grade-level portfolio (i.e., a portfolio for a student who requires an alternate assessment but who is working at or close to grade level expectations) was evaluated by a panel of content area experts to determine whether it met *Needs Improvement* (or higher) achievement-level requirements. To receive a achievement level of *Needs Improvement* or higher, the portfolio must have demonstrated

- that the student had independently and accurately addressed all required learning standards and strands described in the portfolio requirements, and
- that the student provided evidence of knowledge and skills at a level comparable with a student who received an achievement level of *Needs Improvement* or higher on the standard MCAS test in that content area.

Competency Portfolios in High School

Each 2012 competency portfolio was evaluated by a panel of content area experts to determine whether it met *Needs Improvement* (or higher) achievement-level requirements. To receive a achievement level of *Needs Improvement* or higher, the portfolio must have demonstrated

- that the student had independently and accurately addressed all required learning standards and strands described in the portfolio requirements, and
- that the student provided evidence of knowledge and skills at a level comparable with a student who received an achievement level of *Needs Improvement* or higher on the standard MCAS test in either ELA, mathematics, or STE.

If the student's competency portfolio met these requirements, the student was awarded a CD in that content area.

4.5 MCAS-Alt Classical Item Analyses

As noted in Brown (1983), “A test is only as good as the items it contains.” A complete evaluation of a test’s quality must therefore include an evaluation of each item. Both *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 1999) and the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004) include standards for identifying high-quality items. While the specific statistical criteria identified in these publications were developed primarily for general—not alternate—assessments, the principles and some of the techniques apply to the alternate assessment framework as well.

Both qualitative and quantitative analyses are conducted to ensure that the MCAS-Alt meets these standards. Qualitative analyses are described in earlier sections of this chapter; this section focuses on quantitative evaluations. The statistical evaluations include difficulty indices and discrimination (item-test correlations), structural relationships (correlations among the dimensions), and bias and fairness. The item analyses presented here are based on the statewide administration of the 2012 MCAS-Alt.

4.5.1 Item Difficulty and Discrimination

For purposes of calculating item statistics, three of the five dimension scores on each task (Level of Complexity, Demonstration of Skills and Concepts, and Independence) are included in the calculations. Although the other two dimension scores (Self-Evaluation and Generalized Performance) are reported and summarized, they do not contribute to a student’s overall achievement level. For this reason, they are not included in the calculation of item statistics. In calculating the item statistics, the dimension scores are considered to be similar to traditional test items. Using this definition, all items are evaluated in terms of item difficulty according to standard classical test theory practices. Difficulty is defined as the average proportion of points achieved on an item and is measured by obtaining the average score on an item and dividing by the maximum possible score for the item. MCAS-Alt tasks are scored polytomously, meaning that a student can achieve a score of 1, 2, 3, 4, or 5 for Level of Complexity and a score of M, 1, 2, 3, or 4 for Demonstration of Skills and Concepts and Independence. By computing the difficulty index as the average proportion of points achieved, the items are placed on a scale that ranges from 0.0 to 1.0. Although the *p*-value is traditionally described as a measure of difficulty (as it is described here), it is properly interpreted as an easiness index, because larger values indicate easier items. An index of 0.0 indicates that all students received no credit for the item, and an index of 1.0 indicates that all students received full credit for the item.

Items that have either a very high or very low difficulty index are considered to be potentially problematic, because they are either so difficult that few students get them right or so easy that nearly all students get them right. In either case, such items should be reviewed for appropriateness for inclusion on the assessment. If an assessment were composed entirely of very easy or very hard items, all students would receive nearly the same scores, and the assessment would not be able to differentiate high-ability students from low-ability students.

It is worth mentioning that using norm-referenced criteria such as *p*-values to evaluate test items is somewhat contradictory to the purpose of a criterion-referenced assessment like the MCAS-Alt. Criterion-referenced assessments are primarily intended to provide evidence of student progress relative to a standard rather than provide a comparison with other students. In addition, the MCAS-

Alt makes use of teacher-designed items to measure performance. For these reasons, the generally accepted criteria regarding classical item statistics should be cautiously applied to the MCAS-Alt.

A desirable feature of an item is that the higher-ability students perform better on the item than lower-ability students. The correlation between student performance on a single item and total test score is a commonly used measure of this characteristic of an item. Within classical test theory, this item-test correlation is referred to as the item's discrimination, because it indicates the extent to which successful performance on an item discriminates between high and low scores on the test. The discrimination index used to evaluate MCAS-Alt items was the Pearson product-moment correlation. The theoretical range of this statistic is −1.0 to 1.0.

Discrimination indices can be thought of as measures of how closely an item assesses the same knowledge and skills assessed by other items contributing to the criterion total score. That is, the discrimination index can be thought of as a measure of construct consistency. In light of this interpretation, the selection of an appropriate criterion total score is crucial to the interpretation of the discrimination index. For the MCAS-Alt, the sum of the three dimension scores, excluding the item being evaluated, was used as the criterion score.

A summary of the item difficulty and item discrimination statistics for each grade and content area is presented in Table 4-7. The mean difficulty values shown in the table indicate that, overall, students performed well on the items on the MCAS-Alt. In contrast to alternate assessments, the difficulty values for assessments designed for the general population tend to be in the 0.4 to 0.7 range for the majority of items. Because the nature of alternate assessments is different from that of general assessments, and because very few guidelines exist as to criteria for interpreting these values for alternate assessments, the values presented in Table 4-6 should not be interpreted to mean that the students performed better on the MCAS-Alt than the students who took general assessments performed on those tests.

Also shown in Table 4-7 are the mean discrimination values. Because the nature of the MCAS-Alt is different from that of a general assessment, and because very few guidelines exist as to criteria for interpreting these values for alternate assessments, the statistics presented in Table 4-7 should be interpreted with caution.

Table 4-7. 2012 MCAS-Alt: Summary of Item Difficulty and Discrimination Statistics by Content Area and Grade

Content Area	*Grade*	*Number of Items*	p-*Value*		*Discrimination*	
			Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
ELA	3	9	0.85	0.20	0.60	0.07
	4	9	0.84	0.19	0.43	0.07
	5	6	0.85	0.20	0.61	0.10
	6	6	0.85	0.20	0.67	0.09
	7	9	0.84	0.19	0.37	0.04
	8	6	0.85	0.20	0.63	0.08
	HS	9	0.84	0.18	0.40	0.07
Mathematics	3	12	0.85	0.19	0.58	0.10
	4	9	0.85	0.20	0.63	0.04
	5	9	0.85	0.19	0.63	0.09
	6	9	0.85	0.19	0.66	0.07
	7	9	0.85	0.20	0.62	0.06

continued

Content Area	*Grade*	*Number of Items*	p-*Value*		*Discrimination*	
			Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
Mathematics	8	6	0.85	0.19	0.63	0.10
	HS	15	0.84	0.17	0.38	0.09
STE	5	12	0.85	0.19	0.39	0.07
	8	12	0.85	0.19	0.34	0.08
Biology	HS	12	0.85	0.18	0.31	0.06
Chemistry	HS	9	0.84	0.20	0.37	0.13
Introductory Physics	HS	9	0.84	0.18	0.54	0.07
Technology/ Engineering	HS	9	0.82	0.19	0.48	0.17

In addition to the item difficulty and discrimination summaries presented above, item-level classical statistics and item-level score distributions were also calculated. Item-level classical statistics—item difficulty and discrimination values—are provided in Appendix E. Item-level score distributions (i.e., the percentage of students who received each score point) are provided in Appendix F for each item. Note that the Self-Evaluation and Generalized Performance dimension scores are included in Appendix F.

4.5.2 Structural Relationships between Dimensions

By design, the achievement-level classification of the MCAS-Alt is based on three of the five scoring dimensions (Level of Complexity, Demonstration of Skills and Concepts, and Independence). As with any assessment, it is important that these dimensions be carefully examined. This was achieved by exploring the relationships among student dimension scores with Pearson correlation coefficients. A very low correlation (near zero) would indicate that the dimensions are not related, a low negative correlation (approaching -1.00) indicates that they are inversely related (i.e., that a student with a high score on one dimension had a low score on the other), and a high positive correlation (approaching 1.00) indicates that the information provided by one dimension is similar to that provided by the other dimension.

The average correlations among the three dimensions by content area and grade level are shown in Table 4-8.

Table 4-8. 2012 MCAS-Alt: Average Correlations Among the Three Dimensions by Content Area and Grade

Content Area	*Grade*	*Number of Items*	*Average Correlation Between:**			*Correlation Standard Deviation**		
			Comp/Ind	*Comp/Sk*	*Ind/Sk*	*Comp/Ind*	*Comp/Sk*	*Ind/Sk*
ELA	3	2	0.11	0.19	0.20	0.03	0.02	0.03
	4	3	0.13	0.25	0.22	0.02	0.06	0.07
	5	2	0.16	0.21	0.23	0.01	0.11	0.01
	6	2	0.20	0.33	0.31	0.03	0.05	0.02
	7	3	0.12	0.26	0.19	0.04	0.03	0.03
	8	2	0.07	0.38	0.23	0.01	0.10	0.00
	HS	3	0.19	0.22	0.21	0.09	0.02	0.05

continued

Content Area	*Grade*	*Number of Items*	*Average Correlation Between:**			*Correlation Standard Deviation**		
			Comp/Ind	*Comp/Sk*	*Ind/Sk*	*Comp/Ind*	*Comp/Sk*	*Ind/Sk*
Mathematics	3	2	0.07	0.17	0.12	0.00	0.03	0.02
	4	2	0.22	0.24	0.18	0.01	0.01	0.03
	5	2	0.19	0.17	0.29	0.04	0.03	0.04
	6	2	0.18	0.25	0.30	0.00	0.01	0.07
	7	2	0.16	0.19	0.21	0.02	0.01	0.00
	8	2	0.23	0.17	0.28	0.02	0.06	0.04
	HS	5	0.23	0.15	0.22	0.07	0.05	0.06
STE	5	4	0.24	0.21	0.26	0.07	0.03	0.05
	8	4	0.13	0.30	0.21	0.03	0.08	0.03
Biology	HS	4	0.16	0.09	0.12	0.02	0.09	0.06
Chemistry	HS	3	0.16	0.56	0.08	0.06	0.34	0.16
Introductory Physics	HS	3	0.33	0.48	0.42	0.04	0.09	0.15
Technology/ Engineering	HS	3	0.29	0.39	0.28	0.18	0.01	0.23

*Comp = Level of Complexity; Sk = Demonstration of Skills and Concepts; Ind = Independence

The average correlations range from very weak (0.00 to 0.20) to weak (0.20 to 0.40) among the Level of Complexity and Independence dimensions, range from very weak to moderate (0.40 to 0.60) among the Level of Complexity and Demonstrations of Skills and Concepts dimensions, and range from very weak to moderate among the Independence and Demonstration of Skills and Concepts dimensions. Note that a weak relationship in some cases may be expected. For example, a weak correlation between Level of Complexity and Independence may not be surprising, whereas a stronger positive correlation is to be expected between Independence and Demonstration of Skills and Concepts. However, it is important to remember in interpreting the information in Table 4-8 that the correlations are based on small numbers of item scores and small numbers of students and should, therefore, be used with caution.

4.5.3 Bias/Fairness

The *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004) explicitly states that subgroup differences in performance should be examined when sample sizes permit and that actions should be taken to ensure that differences in performance are because of construct-relevant, rather than irrelevant, factors. *Standards for Educational and Psychological Testing* (AERA et al., 1999) includes similar guidelines.

When appropriate, the standardization differential item functioning (DIF) procedure (Dorans & Kulick, 1986) was employed to evaluate subgroup differences. The standardized DIF procedure is designed to identify items for which subgroups of interest perform differently, beyond the impact of differences in overall achievement. However, because of the small number of students who take the MCAS-Alt, and because those students take different combinations of tasks, it was not possible to conduct DIF analyses. This is because conducting DIF analyses using groups of fewer than 200 students would result in inflated type I error rates.

Although it is not possible to run quantitative analyses of item bias for MCAS-Alt, fairness is addressed through the portfolio development and assembly processes, and in the development of the standards themselves, which have been thoroughly vetted for bias and sensitivity. The *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities* provides instructional and assessment strategies for teaching students with disabilities the same learning standards (by grade level) as general education students. The Resource Guide is intended to promote access to the general curriculum, as required by law, and to assist educators in planning instruction and assessment for students with significant cognitive disabilities. It was developed by panels of education experts in each content area, including ESE staff, testing contractor staff, higher education faculty, MCAS Assessment Development Committee members, curriculum framework writers, and regular and special educators. Each section was written, reviewed, and validated by these panels to ensure that each modified standard (entry point) embodied the essence of the grade-level learning standard on which it was based, and that entry points at varying levels of complexity were aligned with grade-level content standards.

Specific guidelines direct educators to assemble MCAS-Alt portfolios based on academic outcomes in the content area and strand being assessed, while maintaining the flexibility necessary to meet the needs of diverse learners. The requirements for constructing student portfolios necessitate that challenging skills based on grade-level content standards be taught in order to produce the required evidence. Thus, students are taught academic skills based on the standards at an appropriate level of complexity.

Issues of fairness are also addressed in the portfolio scoring procedures. Rigorous scoring procedures hold scorers to high standards of accuracy and consistency using monitoring methods that include frequent double-scoring, monitoring, and recalibrating to verify and validate portfolio scores. These procedures, along with the ESE's review of each year's MCAS-Alt results, indicate that the MCAS-Alt is being successfully used for the purposes for which it was intended. Section 4.4 describes in greater detail the scoring rubrics used, selection and training of scorers, and scoring quality-control procedures. These processes ensure that bias due to differences in how individual scorers award scores is minimized.

4.6 Characterizing Errors Associated with Test Scores

As with the classical item statistics presented in the previous section, three of the five dimension scores on each task (Level of Complexity, Demonstration of Skills and Concepts, and Independence) were used as the item scores for purposes of calculating reliability estimates. Note that, due to the way in which student scores are awarded—that is, using an overall achievement level rather than a total raw score—it was not possible to run decision accuracy and consistency (DAC) analyses.

4.6.1 MCAS-Alt Reliability

In the previous section, individual item characteristics of the 2012 MCAS-Alt were presented. Although individual item performance is an important focus for evaluation, a complete evaluation of an assessment must also address the way in which items function together and complement one another. Any assessment includes some amount of measurement error; that is, no measurement is perfect. This is true of all academic assessments—some students will receive scores that underestimate their true ability, and others will receive scores that overestimate their true ability. When tests have a high amount of measurement error, student scores are very unstable. Students with high ability may get low scores and vice versa. Consequently, one cannot reliably measure a student's true level of ability with such a test. Assessments that have less measurement error (i.e., errors are small on average, and therefore students' scores on such tests will consistently represent their ability) are described as reliable.

There are several methods of estimating an assessment's reliability. One approach is to split the test in half and then correlate students' scores on the two half-tests; this in effect treats each half-test as a complete test. This is known as a "split-half estimate of reliability." If the two half-test scores correlate highly, items on the two half-tests must be measuring very similar knowledge or skills. This is evidence that the items complement one another and function well as a group. This also suggests that measurement error will be minimal.

The split-half method requires psychometricians to select items that contribute to each half-test score. This decision may have an impact on the resulting correlation, since each different possible split of the test into halves will result in a different correlation. Another problem with the split-half method of calculating reliability is that it underestimates reliability, because test length is cut in half. All else being equal, a shorter test is less reliable than a longer test. Cronbach (1951) provided a statistic, alpha (α), which eliminates the problem of the split-half method by comparing individual item variances to total test variance. Cronbach's α was used to assess the reliability of the 2012 MCAS-Alt. The formula is as follows:

$$\alpha \equiv \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} \sigma_{(Y_i)}^2}{\sigma_x^2}\right]$$

where
i indexes the item,
n is the number of items,
$\sigma_{(Y_i)}^2$ represents individual item variance, and
σ_x^2 represents the total test variance.

Table 4-9 presents raw score descriptive statistics (maximum possible score, average, and standard deviation), Cronbach's α coefficient, and raw score standard errors of measurement (SEMs) for each content area and grade.

Table 4-9. 2012 MCAS-Alt: Raw Score Descriptive Statistics, Cronbach's Alpha, and SEM by Content Area and Grade

Content Area	*Grade*	*Number of Students*	*Alpha*	*SEM*
ELA	3	1,162	0.64	0.64
	4	1,326	0.75	1.30
	5	1,298	0.66	0.64
	6	1,205	0.74	0.61
	7	1,091	0.76	1.53
	8	945	0.68	0.66
	HS	891	0.76	1.84
Mathematics	3	1,131	0.61	0.69
	4	1,261	0.67	0.70
	5	1,317	0.69	0.64
	6	1,240	0.74	0.64
	7	1,030	0.67	0.69

continued

Content Area	*Grade*	*Number of Students*	*Alpha*	*SEM*
Mathematics	8	1,012	0.68	0.66
	HS	874	0.87	1.36
STE	5	1,190	0.83	1.18
	8	913	0.83	1.33
Biology	HS	597	0.73	1.81
Chemistry	HS	43	0.76	1.53
Introductory Physics	HS	75	0.91	1.38
Technology/ Engineering	HS	82	0.82	1.74

An alpha coefficient toward the high end is taken to mean that the items are likely measuring very similar knowledge or skills; that is, they complement one another and suggest a reliable assessment.

4.6.2 Subgroup Reliability

The reliability coefficients discussed in the previous section were based on the overall population of students who participated in the 2012 MCAS-Alt. Appendix P presents reliabilities for various subgroups of interest. Subgroup Cronbach's α coefficients were calculated using the formula defined above based only on the members of the subgroup in question in the computations; values are only calculated for subgroups with 10 or more students.

For several reasons, the results documented in this section should be interpreted with caution. First, inherent differences between grades and content areas preclude making valid inferences about the quality of a test based on statistical comparisons with other tests. Second, reliabilities are dependent not only on the measurement properties of a test but also on the statistical distribution of the studied subgroup. For example, it can be readily seen in Appendix P that subgroup sample sizes may vary considerably, which results in natural variation in reliability coefficients. Or α, which is a type of correlation coefficient, may be artificially depressed for subgroups with little variability (Draper & Smith, 1998). Third, there is no industry standard to interpret the strength of a reliability coefficient, and this is particularly true when the population of interest is a single subgroup.

4.6.3 Interrater Consistency

Section 4.4 of this chapter describes the processes that were implemented to monitor the quality of the hand-scoring of student responses. One of these processes was double-blind scoring of at least 20% of student responses in grades 3–8 and 100% in high school. Results of the double-blind scoring, used during the scoring process to identify scorers who required retraining or other intervention, are presented here as evidence of the reliability of the MCAS-Alt. A summary of the interrater consistency results is presented in Table 4-10. Results in the table are aggregated across the tasks by content area, grade, and number of score categories (five for Level of Complexity and four for Demonstration of Skills and Concepts and Independence). The table shows the number of items, number of included scores, percent exact agreement, percent adjacent agreement, correlation between the first two sets of scores, and the percent of responses that required a third score. This information is also provided at the item level in Appendix O.

Table 4-10. 2012 MCAS-Alt: Summary of Interrater Consistency Statistics Aggregated Across Items by Content Area and Grade

Content Area	*Grade*	*Number of Items*	*Number of Score Categories*	*Number of Included Scores*	*Percent Exact*	*Percent Adjacent*	*Correlation*	*Percent of Third Scores*
ELA	3	4	4	604	99.34	0.66	0.98	0.66
		2	5	328	98.78	0.91	0.76	1.52
	4	6	4	1,108	99.19	0.81	0.96	1.26
		3	5	587	98.64	0.68	0.69	1.70
	5	4	4	1,122	99.73	0.18	0.98	2.05
		2	5	605	98.84	0.50	0.72	3.31
	6	4	4	802	98.75	1.12	0.95	1.25
		2	5	452	98.23	1.55	0.81	3.10
	7	6	4	1,392	99.21	0.79	0.97	1.65
		3	5	829	98.67	0.72	0.75	4.46
	8	4	4	918	98.91	1.09	0.96	1.20
		2	5	548	97.63	1.09	0.54	3.83
	HS	6	4	4,864	98.81	1.19	0.97	5.10
		3	5	3,063	96.80	2.06	0.53	6.92
Mathematics	3	4	4	594	99.16	0.84	0.96	0.84
		2	5	330	98.18	1.52	0.51	1.82
	4	4	4	724	99.31	0.69	0.97	0.69
		2	5	392	98.98	0.77	0.85	1.53
	5	4	4	1,134	99.29	0.71	0.96	2.47
		2	5	616	99.35	0.16	0.68	2.27
	6	4	4	820	99.39	0.61	0.98	1.10
		2	5	459	99.13	0.65	0.86	1.96
	7	4	4	922	99.24	0.76	0.98	1.19
		2	5	581	98.45	0.86	0.79	4.65
	8	4	4	984	99.80	0.20	0.99	0.20
		2	5	569	98.77	0.53	0.60	2.64
	HS	10	4	4,752	98.93	1.07	0.97	4.63
		5	5	3,091	97.31	1.65	0.59	5.95
STE	5	8	4	1,444	99.38	0.62	0.97	1.73
		4	5	783	98.85	1.02	0.77	2.30
	8	8	4	1,276	99.14	0.86	0.96	0.94
		4	5	751	96.67	1.60	0.42	4.26
Biology	HS	6	4	3,236	98.95	1.05	0.96	5.53
		3	5	2,324	96.04	1.76	0.40	7.27
Chemistry	HS	6	4	250	98.40	1.60	0.96	2.40
		3	5	164	93.90	3.66	0.64	9.15
Physics	HS	6	4	330	100.00	0.00	1.00	3.64
		3	5	227	94.71	4.41	0.64	8.37
Technology/ Engineering	HS	6	4	422	97.87	1.90	0.93	3.55
		3	5	283	93.99	4.24	0.70	8.13

4.7 MCAS-Alt Comparability across Years

The issue of comparability across years is addressed in the progression of learning outlined in the *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities*, which provides instructional and assessment strategies for teaching students with disabilities the same learning standards taught to general education students.

Comparability is also addressed in the portfolio scoring procedures. Consistent scoring rubrics are used each year along with rigorous quality control procedures that hold scorers to high standards of accuracy and consistency, as described in Section 4.4. Scorers are trained using the same procedures, models, examples, and methods each year.

Finally, comparability across years is encouraged through the classification of students into performance-level categories, using a lookup table that remains consistent each year (see Table 4-11). The description of each achievement level remains consistent, which ensures that the meaning of students' scores is comparable from one year to the next. Table 4-12 shows the achievement-level lookup table (i.e., the achievement level corresponding to each possible combination of dimension scores), which is used each year to combine and tally the overall achievement level from individual strand scores. In addition, achievement-level distributions are provided in Appendix L. The distributions include results for each of the last three years.

Table 4-11. 2012 MCAS-Alt Performance-Level Descriptions

Performance Level	*Description*
Incomplete (1)	**Insufficient evidence and information** was included in the portfolio to allow a performance level to be determined in the content area.
Awareness (2)	Students at this level demonstrate **very little understanding** of learning standards and core knowledge topics contained in the Massachusetts curriculum framework for the content area. Students require extensive prompting and assistance, and their performance is mostly inaccurate.
g (3)	Students at this level demonstrate a **simple understanding below-grade-level expectations** of a limited number of learning standards and core knowledge topics contained in the Massachusetts curriculum framework for the content area. Students require frequent prompting and assistance, and their performance is limited and inconsistent.
Progressing (4)	Students at this level demonstrate a **partial understanding below-grade-level expectations** of selected learning standards and core knowledge topics contained in the Massachusetts curriculum framework for the content area. Students are steadily learning new knowledge, skills, and concepts. Students require minimal prompting and assistance, and their performance is basically accurate.
Needs Improvement (5)	Students at this level demonstrate a **partial understanding of grade-level subject matter** and solve some simple problems.
Proficient (6)	Students at this level demonstrate a **solid understanding of challenging grade-level subject matter** and solve a wide variety of problems.
Advanced (7)	Students at this level demonstrate a **comprehensive understanding of challenging grade-level subject matter** and provide sophisticated solutions to complex problems.

Table 4-12. MCAS-Alt: Strand Performance-Level Lookup

Level of complexity	*Demonstration of skills*	*Independence*	*Performance Level*
2	1	1	1
2	1	2	1
2	1	3	1
2	1	4	1
2	2	1	1
2	2	2	1
2	2	3	1
2	2	4	1
2	3	1	1
2	3	2	1
2	3	3	2
2	3	4	2
2	4	1	1
2	4	2	1
2	4	3	2
2	4	4	2
3	1	1	1
3	1	2	1
3	1	3	1
3	1	4	1
3	2	1	1
3	2	2	1
3	2	3	2
3	2	4	2
3	3	1	1
3	3	2	2
3	3	3	3
3	3	4	3
3	4	1	1
3	4	2	2
3	4	3	3
3	4	4	3
4	1	1	1
4	1	2	1
4	1	3	1
4	1	4	1
4	2	1	1
4	2	2	1
4	2	3	2
4	2	4	2
4	3	1	1
4	3	2	2
4	3	3	3
4	3	4	3
4	4	1	1

continued

Level of complexity	*Demonstration of skills*	*Independence*	*Performance Level*
4	4	2	2
4	4	3	3
4	4	4	3
5	1	1	1
5	1	2	1
5	1	3	2
5	1	4	2
5	2	1	1
5	2	2	2
5	2	3	3
5	2	4	3
5	3	1	1
5	3	2	2
5	3	3	3
5	3	4	4
5	4	1	1
5	4	2	2
5	4	3	3
5	4	4	4

4.8 Reporting of Results

4.8.1 Primary Reports

MP created the following primary reports for the MCAS-Alt:

- *Portfolio Feedback Form*
- *Parent/Guardian Report*

4.8.1.1 Portfolio Feedback Forms

One *Portfolio Feedback Form* is produced for each student who submitted an MCAS-Alt portfolio. Content-area performance level(s), strand dimension scores, and comments relating to those scores are printed on the form. The *Portfolio Feedback Form* is a preliminary score report intended for the educator who submitted the portfolio.

4.8.1.2 Parent/Guardian Report

The *Parent/Guardian Report* provides the final scores (overall score and rubric dimension scores) for each student who submitted an MCAS-Alt portfolio. It provides background information on the MCAS-Alt, participation requirements, the purpose of the assessment, an explanation of the scores, and contact information for further information. Achievement levels are displayed for each content area relative to all possible achievement levels. The student's dimension scores are displayed in relation to all possible dimension scores for the assessed strands.

Two printed copies of the reports are provided for each student: one for the parent and one to be kept in the student's temporary record. Sample reports are provided in Appendix S.

4.8.2 Interpretive Materials

The *2012 Parent/Guardian Report* was redesigned to incorporate information that previously was published in a separate interpretive guide, which was not produced in 2012. Two parent focus groups provided feedback to the ESE on the report redesign.

4.8.3 Decision Rules

To ensure that reported results for the MCAS-Alt are accurate relative to the collected portfolio evidence, a document delineating decision rules is prepared before each reporting cycle. The decision rules are observed in the analyses of the MCAS-Alt data and in reporting results. Copies of the decision rules are included in Appendix T.

4.8.4 Quality Assurance

Quality assurance measures are implemented throughout the entire process of analysis and reporting at MP. The data processors and data analysts working on the MCAS-Alt data perform quality control checks of their respective computer programs. Moreover, when data are handed off to different units within the Data and Reporting Services division (DRS), the sending unit verifies that the data are accurate before handoff. Additionally, when a unit receives a data set, the first step is to verify the accuracy of the data.

Quality assurance is also practiced through parallel processing. One data analyst is responsible for writing all programs required to populate the student and aggregate reporting tables for the administration. Each reporting table is assigned to another data analyst who uses the decision rules to independently program the reporting table. The production and quality assurance tables are compared; if there is 100% agreement, the tables are released for report generation.

A third aspect of quality control involves the procedures implemented by the quality assurance group to check the accuracy of reported data. Using a sample of students, the quality assurance group verifies that the reported information is correct. The selection of specific sampled students for this purpose may affect the success of the quality control efforts.

The quality assurance group uses a checklist to implement its procedures. Once the checklist is completed, sample reports are circulated for psychometric checks and review by program management. The appropriate sample reports are then sent to the ESE for review and signoff.

4.9 MCAS-Alt Validity

One purpose of the *2012 MCAS and MCAS-Alt Technical Report* is to describe the technical aspects of the MCAS-Alt that contribute validity evidence in support of MCAS-Alt score interpretations. A framework for organizing this validity evidence is provided by the *Standards for Educational and Psychological Testing* (AERA et al., 1999). According to the *Standards*, the sources of evidence that should be considered when constructing a validity argument include: test content, response processes, internal structure, relationship to other variables, and consequences of testing.

Recall that the score interpretations for the MCAS-Alt include using the results to make inferences about student achievement on the ELA, mathematics, and STE content standards; to inform program and instructional improvement; and as a component of school accountability. Thus, as described below, each section of the report (development, administration, scoring, item analyses, reliability, performance levels, and reporting) contributes to one of the *Standards'* sources of validity evidence and, taken together, they form a comprehensive validity argument in support of MCAS-Alt score interpretations.

4.9.1 Test Content Validity Evidence

As described earlier, evidence for test content validity is determined by how well the assessment tasks, i.e., the primary evidence contained in the portfolios, represent the curriculum and standards for each content area and grade level. This evidence is described in detail in Section 4.2.1.

4.9.2 Internal Structure Validity Evidence

Evidence based on internal structure is presented in detail in the discussions of item analyses and reliability in Sections 4.5 and 4.6. Technical characteristics of the internal structure of the assessment are presented in terms of classical item statistics (item difficulty and item-test correlation), correlations among the dimensions (Level of Complexity; Demonstration of Skills and Concepts; and Independence), fairness/bias, and reliability, including alpha coefficients, interrater consistency, and decision accuracy and consistency.

4.9.3 Response Process Validity Evidence

The training and administration information in Section 4.3 describes the steps taken to train educators on procedures for assembling the MCAS-Alt. Portfolios are constructed and administered according to state-mandated procedures, as described in the *2012 Educator's Manual for MCAS-Alt*. Efforts by the ESE to provide educators with training, resources, and ongoing support serve to maximize consistency and enhance the quality and reliability of the inferences made based on results, and contribute to the validity of the assessment.

Procedures for training and monitoring the scoring of the MCAS-Alt (described in Section 4.4) also maximize consistency and contribute to overall validity.

4.9.4 Validity Evidence Based on Consequences of Testing

Information provided in Section 4.7 indicates how the reporting of results ensures comparability of scores across years, which in turn, contributes to validity.

Efforts were undertaken to provide the public with accurate and clear information about scores (described in Section 4.8), including reporting of achievement levels that provide reference points for mastery at each grade level and achievement level descriptors that provide a useful and consistent way to interpret scores.

4.9.5 Summary

The evidence for validity and reliability presented in this chapter supports the use of the assessment to make inferences about student achievement of the skills and content described in the Massachusetts curriculum frameworks for ELA, mathematics, and STE. As such, this evidence supports the use of MCAS-Alt results for the purposes of programmatic and instructional improvement and as a component of school accountability.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Belmont, CA: Wadsworth, Inc.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baker, F. B. (1992). *Item response theory: parameter estimation techniques.* New York: Marcel Dekker, Inc.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: parameter estimation techniques* (2nd ed.). New York: Marcel Dekker, Inc.

Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.). Fort Worth, TX: Holt, Rinehart and Winston.

Chicago Manual of Style (15th ed.). (2003). Chicago: University of Chicago Press.

Clauser, J. C., & Hambleton, R. K. (2011a). *Improving curriculum, instruction, and testing practices with findings from differential item functioning analyses: Grade 8, Science and Technology/Engineering* (Research Report No. 777). Amherst, MA: University of Massachusetts Amherst, Center for Educational Assessment.

Clauser, J. C., & Hambleton, R. K. (2011b). *Improving curriculum, instruction, and testing practices with findings from differential item functioning analyses: Grade 10, English language arts* (Research Report No. 796). Amherst, MA: University of Massachusetts Amherst, Center for Educational Assessment.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

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

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: John Wiley and Sons, Inc.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston: Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications, Inc.

Hambleton, R. K., & van der Linden, W. J. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: Author. Retrieved from http://www.apa.org/science/programs/testing/fair-code.aspx

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement 43*(*4*), 355–81.

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.

Measured Progress Department of Psychometrics and Research. (2011). *2010–2011 MCAS Equating Report*. Unpublished manuscript.

Muraki, E., & Bock, R. D. (2003). PARSCALE 4.1 [Computer software]. Lincolnwood, IL: Scientific Software International.

Nering, M., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York: Routledge.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan Publishing Company.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52,* 589–617.

Stout, W. F., Froelich, A. G., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 357–375). New York: Springer-Verlag.

Zhang, J., & Stout, W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 213–249.

Appendices