Massachusetts Department of
ELEMENTARY & SECONDARY
EDUCATION

2010 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

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

The Massachusetts Comprehensive Assessment System (MCAS) is the Commonwealth's program for student assessment developed in accordance with the Education Reform Act of 1993. The main purposes of the MCAS are to

- measure student, school, and district performance in meeting the state's learning standards as detailed in the Massachusetts curriculum frameworks;
- improve student achievement and classroom instruction by providing diagnostic feedback regarding the acquisition of skills and knowledge;
- help determine English language arts (ELA), mathematics, and science and technology/engineering (STE) competency for the awarding of high school diplomas;
- hold schools and districts accountable for the yearly progress they make toward meeting the goal, set by the federal No Child Left Behind (NCLB) Act, that all students will become proficient in reading and mathematics.

The purpose of this *2010 MCAS and MCAS-Alt Technical Report* is to document the technical quality and characteristics of the 2010 MCAS operational tests, and to present evidence of the validity and reliability of those tests' results. For all characteristics of the MCAS program that were modified in 2010, complete technical data and details are provided in this 2010 report. Technical reports for 1998 to 2009 are available on the Department of Elementary and Secondary Education (ESE) website at www.doe.mass.edu/mcas/tech/?section=techreports.

This 2010 report provides detailed information regarding test design and development, scoring, and analysis and reporting of 2010 MCAS results at the student, school, district, and state levels. This detailed information includes but is not limited to the following:

- Test administration
- 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 2010 MCAS test and its results.

1.1 Purpose of This Report

As mentioned previously, the *2010 MCAS and MCAS-Alt Technical Report* is designed to supplement the technical reports issued for previous MCAS administrations by providing information specific to the 2010 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.

1.2 Organization of This Report

Chapter 1 of this report provides a brief overview of what is documented within the report, including updates made to the MCAS program during 2010. Chapter 2 explains the guiding philosophy, purpose, uses, components, and validity of the state's assessment system. The report then includes two chapters that cover the test design and development, test administration, scoring, and analysis and reporting of results for the standard MCAS assessment (Chapter 3) and the alternate MCAS assessment (Chapter 4). Within these two chapters, there is much detail on 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.

1.3 Current Year Updates

In addition to changes detailed throughout this document, the following changes were made for the 2010 MCAS administration.

1.3.1 Reduction in Testing Time

In 2010, as part of an effort to reduce testing time, the MCAS ELA reading comprehension tests in grades 3–8 were shortened by eliminating one test session. See Section 3.2.2.3 for a description of the changes in the test design.

Similarly, the MCAS Mathematics tests in grades 3–8 were shortened by eliminating some of the matrix slots. (The tests are composed of common and matrix items. The matrix slots of each test form are used to field-test potential MCAS items or to equate this year's test to that of previous years by using previously administered items.) See Section 3.2.3.3 for a description of the changes in the test design.

See also Section 3.6.4.1 for a description of a special study that was conducted to evaluate the effects of the changes in test design noted above (and consequent reduction in the number of equating items) on the equating of the 2010 ELA reading comprehension and Mathematics tests in grades 3–8.

No test design changes were made to any other test or retest in 2010.

1.3.2 Competency Determination

Beginning with the class of 2010, to receive the Competency Determination (CD) required for high school graduation, students must

either

- earn a scaled score of at least 240 on both the grade 10 MCAS ELA and Mathematics tests or retests

or

- earn a scaled score between 220 and 238 on both tests or retests *and* fulfill the requirements of an Educational Proficiency Plan (EPP) (more information about EPP requirements can be found at www.doe.mass.edu/hsreform/epp)

AND

- earn a scaled score of at least 220 on one of the four high school MCAS STE end-of-course tests (Biology, Chemistry, Introductory Physics, Technology/Engineering).

Students must also meet all local graduation requirements.

1.3.3 Student Growth Percentile Scores

In 2010, for the first time, the ESE reported a student growth percentile (SGP) score in ELA and mathematics for students in grades 4–8 and 10 as part of the *Parent/Guardian Report*. This SGP score provided information about how much a student's MCAS achievement changed since the previous year, relative to other students with similar MCAS score histories. More detailed information about SGPs is provided in Section 3.8.

1.3.4 Reporting of Scaled Scores for Grade 3

In 2010, for the first time, scaled scores were reported for grade 3 ELA and mathematics results. The ESE's policy decision to use scaled scores makes score reporting at grade 3 consistent with the other grade levels.

1.3.5 Additional and Suspended Administrations

- **In April 2010**, a one-time, special testing opportunity was offered to students in the class of 2010 who had not yet taken or passed a high school MCAS STE test.
- **Since 2009**, the grades 5 and 7 History and Social Science pilot tests and the high school (grades 10–11) U.S. History pilot test have been suspended for budgetary reasons.

1.3.6 Updated Information About MCAS Test Participation Requirements

Updated, complete student participation requirements for all spring 2010 MCAS tests can be found in the *Spring 2010 Principal's Administration Manual*.

Student participation requirements for the November 2009 ELA and Mathematics retests, February 2010 Biology test, and March 2010 ELA and Mathematics retests can be found in the *Fall 2009/Winter 2010 Principal's Administration Manual*.

For a copy of either document, please call Student Assessment Services at 781-338-3625.

Chapter 2. The State Assessment System

2.1 Introduction

The 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 limited English proficient students;
- measure performance based on the Massachusetts curriculum framework learning standards;
- report on the performance of individual students, schools, and districts.

As required by the Education Reform Act, students must pass the grade 10 tests in English language arts, mathematics, and science and technology/engineering as one condition of eligibility for a high school diploma (in addition to fulfilling local requirements).

In addition, the MCAS program is used to hold schools and districts accountable, on a yearly basis, for the progress they have made toward the objective required by the No Child Left Behind (NCLB) Act that all students be proficient in reading and mathematics by 2014.

2.2 Guiding Philosophy

The MCAS and MCAS Alternate Assessment (MCAS-Alt) programs play a central role in helping all of the 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 lever 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, and for over a decade has served the Commonwealth well. The program continues to evolve, with recent and current improvements expected to

- respond to stakeholders' interests;
- reflect the vision and goals outlined by the governor's Readiness Project;
- respond to the Board of 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.

> Massachusetts is at a crossroads. Fifteen years after the passage of landmark education reform legislation, the Commonwealth is a national education leader. Standards-based reforms have yielded significant results…At the same time, this success masks persistent, complex problems that demand immediate attention. Despite the quantum leaps in academic rigor, our existing education system is not

> adequately preparing every student for success in life and work…Times have changed and so must the fundamental promise of public education. Today, our schools must ensure that high school graduates know and are capable of much more than ever before. Meeting this challenge requires the creation of a fully integrated, coherent and seamless education system.
>
> —*Ready for 21st Century Success: The New Promise of Public Education.* The Patrick Administration Education Action Agenda, June 2008.

In order for the vision of a fully integrated, coherent, and seamless education system to be realized, important enhancements and new components are necessary as the MCAS continues into its second decade.

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 conclusions about the level of students' achievement 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 Adequate Yearly Progress (AYP) toward meeting federal NCLB requirements
- 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
- making decisions about scholarships, including the John and Abigail Adams Scholarship
- providing diagnostic information to help all students reach higher levels of performance

---

[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 English language arts 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.

2.5 Components of the State Assessment System

The Massachusetts Education Reform Mandate

The Massachusetts Education Reform Act of 1993 specifies that the testing program must

- test *all* students who are educated with Massachusetts public funds, including students with disabilities and limited English proficient students;
- measure performance based on the Massachusetts curriculum framework learning standards (the current Massachusetts curriculum frameworks and the revision schedule are posted on the ESE's website at www.doe.mass.edu/frameworks);
- report on the performance of individual students, schools, districts, and the state.

As required by the Education Reform Act, students must earn a CD by passing grade 10 tests in ELA, mathematics, and STE as one condition of eligibility for a Massachusetts high school diploma.

The MCAS program is also used to comply with the standards and assessment requirements imposed by NCLB and is used as a core measure in the generation of AYP reports called for by NCLB.

2.6 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 [IRT] statistics, reliability, dimensionality, decision accuracy and consistency), and reporting. Information is described in detail in the 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).

Chapter 3. MCAS

3.1 Overview

MCAS tests have been administered to students in Massachusetts since 1998. In 1998, English language arts (ELA), mathematics, and science and technology/engineering (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 Competency Determination (CD) in ELA and mathematics as a condition for receiving a high school diploma. To fulfill the requirements of the No Child Left Behind (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 staff of the ESE with assistance and support from the assessment contractor (Measured Progress). 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 and 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 2010 MCAS 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 2010 MCAS administration also included retest opportunities in ELA and mathematics in November 2009 and March 2010 for students beyond grade 10 who had not yet passed the standard grade 10 test. A February Biology test was administered and, for 2010 only, an April retest in high school STE was offered, as the class of 2010 was the first class required to pass a high school STE test in order to earn a CD.

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 item type, may address part, all, or several of the indicators within a standard.

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 (MC)** 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 raw score point, and incorrect and blank responses are assigned a score of zero raw score points.
- **One-point short-answer (SA**) 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 merely selecting, an answer. One-point short-answer items are hand-scored as zero points (blank or incorrect) or one point (correct).
- **Two-point open-response (OR)** 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 are categorized as blanks and receive a zero raw score.
- **Two-point short-response (SR)** items are used in the grade 3 ELA 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 are categorized as blanks and receive a zero raw score.
- **Four-point open-response (OR)** 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.
- **Writing prompts (WP)** are administered to all students at grades 4, 7, and 10 as part of their ELA test. Students are required to write a draft composition. In a second session, students write a final composition based on that draft. Each composition is hand-scored by a professional scorer trained in the MCAS writing score point descriptions. Students receive two scores: one for topic development ranging from 0 to 6 points, the other for standard English conventions ranging from 0 to 4 points. Student reports include a score for each of these dimensions. Each student composition is scored by two different readers; the final score

is a combination of both sets of scores, so students may receive up to 20 points for their compositions.

3.2.1.3 Description of Test Design

The MCAS is structured using both *common* and *matrix* items. Common items are taken by 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 content area but ranges between 5 and 38 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 and have a small impact on testing time. 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 English Language Arts Test Specifications

3.2.2.1 Standards

The reading comprehension portion of the MCAS English Language Arts tests in grades 3–8 and 10 and the grade 10 retests measures the following learning standards from 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 tests at grades 4, 7, and 10 and the retests measures 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 are not assessable 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.*

3.2.2.2 Item Types

The MCAS reading comprehension portion of the ELA tests includes a mix of multiple-choice and open-response items. Two-point short-response questions 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. 2010 MCAS: English Language Arts Item Types and Score Points

Item Type	*Possible Raw Score Points*
Multiple-choice (MC)	0 or 1
Short-response (SR)	0, 1, or 2
Open-response (OR)	0, 1, 2, 3, or 4
Writing prompt (WP)	0 to 20

3.2.2.3 Test Design

In 2010, as part of an effort to reduce testing time, the MCAS ELA reading comprehension tests in grades 3–8 were shortened by eliminating one session. Table 3-2 describes the changes in the test design. See Section 3.6.4.1 for a description of a special study that was conducted to evaluate the effects of the change in test design (and consequent reduction in the number of equating items) on the equating of the 2010 tests.

Table 3-2. 2010 MCAS: Comparison of 2009 and 2010 English Language Arts Test Designs

	# of Sessions		*Minutes*		*Common Points*		*Matrix Points*	
Grade	*2009*	*2010*	*2009*	*2010*	*2009*	*2010*	*2009*	*2010*
3	3	2	150	120	48	48	12	14
4–8	3	2	135	120	52	52	20	14

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

Grade 3 Reading Comprehension Test

The common portion of the test includes two long passages and three short passages. Each long passage is 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, for a total of 16 multiple-choice and two 2-point short-response items. The grade 3 ELA test contains a total of 48 common points and 14 matrix points distributed across two testing sessions.

Grades 4–8 Reading Comprehension Tests

The common portion of each test includes two long passages and three short passages. Each long passage is 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. Grades 4–8 reading comprehension tests contain 52 common points and 14 matrix points per form distributed across two testing sessions.

Grade 10 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 sessions.

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 divided into 12 points for topic development and 8 points for standard English conventions.

English Language Arts Retest

A retest was offered to students 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 retest consists of common items only.

Table 3-3. 2010 MCAS: Distribution of English Language Arts Common and Matrix Items by Grade and Item Type

Grade and Test			*Items per Form*								*Total Matrix Items Across Forms*							
			Common				*Matrix*				*Equating Positions*				*Field-Test Positions*			
Grade	*Test*	*# of Forms*	*MC*	*SR*	*OR*	*WP*	*MC*	*SR*	*OR*	*WP*	*MC*	*SR*	*OR*	*WP*	*MC*	*SR*	*OR*	*WP*
3	Reading	15	36	4	1		10		1[a]		30		3[a]		120		12	
4	Reading	15	36		4		10		1		30		3		120		12	
4	Composition	2[b]				1												
5	Reading	15	36		4		10		1		30		3		120		12	
6	Reading	15	36		4		10		1		30		3		120		12	
7	Reading	15	36		4		10		1		30		3		120		12	
7	Composition	2[b]				1												
8	Reading	15	36		4		10		1		30		3		120		12	
10	Reading	38	36		4		12		2		96[c]		16[c]		360		60	
10	Composition	2[b]				1												
Retest[d]	Reading	1	36		4													
Retest[d]	Composition	1				1												
Retest[d]	Reading	1	36		4													
Retest[d]	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]While the grade 10 test is pre-equated, additional matrix items were included as potential backup equating items.
[d]ELA retests consist of common items only.

3.2.2.4 Blueprints

Table 3-4 shows the test specifications—the approximate distribution of common item points across the *Massachusetts English Language Arts Curriculum Framework* strands—for the MCAS 2010 ELA tests.

Table 3-4. 2010 MCAS: English Language Arts Common Point Distribution by Strand and 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 MCAS ELA 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 an 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 recognizes basic information presented in the text(s).
- Level II (Infer/Analyze) – Level II items require that the test taker understands a given text by making inferences and drawing conclusions related to the text(s).
- Level III (Evaluate/Apply) – Level III items require that the test taker understands multiple points of view and is able to project his/her own judgments or perspectives on the text(s).

Stated another way:

- Level I items require that students read the lines;
- Level II items require that students read between the lines; and
- Level III items require that students read beyond the lines.

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

3.2.2.6 Passage Types

Reading passages include both long and short texts—word counts vary between long passages 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, however, are treated as short passages regardless of length.

Passages were collected from published work; no passages were specifically written for the MCAS ELA tests. Passages can be broken down into the following passage types:

- Literary passages – These 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 – These 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 operational 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 operational form includes one long and three short literary passages and two long informational passages.

In grades 3–8, long passages are tested with 10 multiple-choice items and one open-response item. In grade 3, the 4-point open-response item for long passages may be replaced by two 2-point short-response items. Short passages are tested with five or six multiple-choices items and one or no open-response items. In grade 10, long passages are tested with eight multiple-choice items and one open-response item. Short passages are tested with four or five multiple-choice items and one or no open-response items.

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 percent of the passages used are written by authors found in Appendices A and B of the *Massachusetts 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. Measured Progress uses a variety of readability formulas to aid in the selection of passages appropriate for the intended audience. In addition, Massachusetts teachers use grade-level expertise to contribute to the selection of passages as members of the Assessment Development Committees.

Items related to these 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 and/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.

3.2.3 Mathematics Test Specifications

3.2.3.1 Standards

The MCAS Mathematics tests at grades 3–8 and grade 10 measure the learning standards of the five strands of the *Massachusetts Mathematics Curriculum Framework*:

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

3.2.3.2 Item Types

The MCAS 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. 2010 MCAS: Mathematics Item Types and Score Points

Item Type	*Possible Raw Score Points*
Multiple-choice (MC)	0 or 1
Short-answer (SA)	0 or 1
2-point open-response (OR)*	0, 1, or 2
Open-response (OR)	0, 1, 2, 3, or 4

*Only grade 3 Mathematics uses 2-point open-response items.

3.2.3.3 Test Design

In 2010, as part of an effort to reduce testing time, the MCAS Mathematics tests in grades 3–8 were shortened by eliminating some of the matrix slots. Table 3-6 describes the changes in the test design; the grade 10 Mathematics test was unchanged. See Section 3.6.4.1 for a description of a special study that was conducted to evaluate the effects of the change in test design (and consequent reduction in the number of equating items) on the equating of the 2010 test.

Table 3-6. 2010 MCAS: Comparison of 2009 and 2010 MCAS Mathematics Test Designs

Grade	*# of Sessions*		*Minutes*		*Common Points*		*Matrix Points*	
	2009	*2010*	*2009*	*2010*	*2009*	*2010*	*2009*	*2010*
3	2	2	60	45	40	40	10	7
4–6	2	2	60	45	54	54	12	7
7–8	2	2	60	45	54	54	16	12

The tests are composed of common and matrix items. The matrix slots of each test form are used to field-test potential MCAS 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. 2010 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*			*Equating Positions*			*Field-Test Positions*		
		MC	*SA*	*OR*	*MC*	*SA*	*OR*	*MC*	*SA*	*OR*	*MC*	*SA* [a]	*OR* [a]
3	18	26	6	4[b]	2	1	1	13	3	2	23	9	6
4	21	32	6	4	2	1	1	16	3	2	26	9	6
5	21	32	6	4	2	1	1	16	3	2	26	9	6
6	21	32	6	4	2	1	1	16	3	2	26	9	6
7	21	32	6	4	2	2	2	16	3	2	26	15	6
8	21	32	6	4	2	2	2	16	3	2	26	15	6
10	32	32	4	6	7	1	2	64[c]	8[c]	12[c]	160	24	30
Retest[d]	1	32	4	6									
Retest[d]	1	32	4	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. So at grade 4, there were actually 21 SA slots and 21 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]While the grade 10 test is pre-equated, additional matrix items were included as potential backup equating items.
[d]Mathematics retests consist of common items only.

3.2.3.4 Blueprints

Table 3-8 shows the test specifications—the approximate distribution of common item points across the *Massachusetts Mathematics Curriculum Frameworks* strands—for the 2010 MCAS Mathematics tests.

Table 3-8. 2010 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	35%	35%	33%	33%	26%	26%	20%
Patterns, Relations, and Algebra	20%	20%	26%	26%	28%	28%	30%
Geometry	12.5%	12.5%	13%	13%	13%	13%	15%
Measurement	12.5%	12.5%	13%	13%	13%	13%	15%
Data Analysis, Statistics, and Probability	20%	20%	15%	15%	20%	20%	20%
Total	100%	100%	100%	100%	100%	100%	100%

3.2.3.5 Cognitive Levels

Each item on the MCAS 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 an item correctly. Each of the three cognitive levels used in the MCAS Mathematics test is listed and described below.

- Level I – Recall and Recognition - Test items in this category 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 - Test items in this category require students to engage in mathematical reasoning beyond simple recall, a more flexible thought process, and enhanced organization of thinking skills. The items demand that students make a decision about the approach needed, represent or model a situation, and/or use one or more nonroutine procedures to solve a well-defined problem.
- Level III – Judgment and Synthesis - Test items in this category 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 in all grades. Level III is best represented by open-response items. An attempt is made to include cognitive level III items at each grade.

3.2.3.6 Use of Calculators and Reference Sheets

Beginning at grade 7, the second session of each Mathematics test is a calculator session. All items included in this session are calculator neutral (calculators are permitted but not required to answer the question) or calculator active (students should use calculators to answer the question). There are no specific limitations on the type of calculator students may use.

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 test 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. These tool kits contain manipulatives to answer specific questions. The tool kits are designed for specific items and therefore change annually. They are published with the released items. All students in grades 3–8 receive rulers for use on the Mathematics test. Students may keep the rulers after test administration.

3.2.4 Science and Technology/Engineering Test Specifications

3.2.4.1 Standards

Grades 5 and 8

The MCAS Science and Technology/Engineering tests at grades 5 and 8 measure the learning standards of the four strands of the *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 MCAS 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 high school MCAS STE tests are eligible but not required to take the MCAS test in the course they are studying. All students are required to take one of the four high school MCAS STE 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. Any grade 11 or 12 student who has not yet passed an STE test is eligible to take any of the four STE tests.

Testing opportunities for high school STE are given in February (Biology only) and June (Biology, Chemistry, Introductory Physics, and Technology/Engineering).

In April 2010, high school seniors who had not yet passed an MCAS STE test were given the opportunity to take an additional STE test in Biology, Chemistry, Introductory Physics, or Technology/Engineering. This opportunity will not be offered in subsequent MCAS administrations.

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 MCAS STE tests include multiple-choice and open-response items. Open-response items are more complex, requiring 8–10 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. 2010 MCAS: STE Item Types and Score Points

Item Type	*Possible Raw Score Points*
Multiple-choice (MC)	0 or 1
Open-response (OR)	0, 1, 2, 3, or 4

The high school Biology test includes modules. A module is composed of a single stimulus (a graphic or a written scenario) and a group of associated items. Each module consists of four multiple-choice items and one open-response item, and appears only in the high school Biology test.

3.2.4.3 Test Design

The MCAS 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. 2010 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[c]	15	40[a]	5[a]	12[b]	2[b]	NA	NA	180	28
	Chemistry[c]	5	40	5	20	2	NA	NA	100	10
	Introductory Physics[c]	10	40	5	12	2	NA	NA	120	15
	Technology/ Engineering[c]	5	40	5	20	2	NA	NA	100	10

[a]High school Biology common items may include a module consisting of 4 MC items and 1 OR item. These are included in the overall counts.
[b]High school Biology matrix items include one matrix module per form consisting of 4 MC items and 1 OR item. These are included in the overall matrix counts.
[c]High school STE tests are pre-equated, and there are no extra equating items.

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. 2010 MCAS: STE Common Point Distribution by Strand and Grade

Framework Strand	*Grade 5*	*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. 2010 MCAS: High School Biology Common Point Distribution by Strand

MCAS Reporting Category	*Percent of Raw Score Points*	*Related Frameworks 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. 2010 MCAS: High School Chemistry Common Point Distribution by Strand

MCAS Reporting Category	*Percent of Raw Score Points*	*Related Frameworks 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. 2010 MCAS: High School Introductory Physics Common Point Distribution by Strand

MCAS Reporting Category	*Percent of Raw Score Points*	*Related Frameworks 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. 2010 MCAS: High School Technology/Engineering Common Point Distribution by Strand

MCAS Reporting Category	*Percent of Raw Score Points*	*Related Frameworks 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 MCAS 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 rates each item based on the complexity of the mental processing a student must use to answer an 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. 2010 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 and Formula Sheets

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 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.

3.2.5 Test Development Process

Table 3-17 details the test development process.

Table 3-17. 2010 MCAS: Test Development Process Overview

Development Step	*Details of the Process*
Select reading passages	Test developers find potential passages and present them to the ESE, then to Assessment Development Committee (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 test.
Benchmark open-response items and compositions	ESE and MP staff determine appropriate benchmark papers for training of scorers of open-response 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 present to provide test characteristic curves and statistical information.
Operational test items	Items become part of the common item set and are used to provide 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/grade 10 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 3.2.2.6 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 MCAS ELA tests. Passages must be of interest to students; have a clear beginning, middle, and ending; support the development of unique assessment items; and be free of bias and sensitivity issues before they can be considered for the reading comprehension portion of the MCAS.

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 the Department of Elementary and Secondary Education

ESE Passage Review

The ESE content staff reviews potential passages before presenting the passages for Assessment Development Committee (ADC) review. Passages are reviewed for

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

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

ESE Item Review

All items and scoring guides are reviewed by the ESE content specialists 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 both the subject and pedagogy?
- 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 to 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?

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

Review by Assessment Development Committees

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 a specific 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

English language arts 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 treat passages in the same way as items in terms of their recommendations:

- accept
- accept with 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. 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.

Passage Review

All passages undergo a review by the Bias and Sensitivity Review Committee before they are approved for development. Committee members consider all passages in terms of gender, race, ethnicity, geography, religion, sexual orientation, culture, and social appropriateness and make recommendations as to whether to accept or reject those passages. They review the passages to ensure that students taking the test are not disadvantaged because of issues not related to the construct being tested. All recommendations to reject passages are accompanied by explanations as to the nature 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 for passages are not developed until the passages have been accepted by the Bias and Sensitivity Review Committee and approved by the ESE.

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.) Records of comments made by the Bias and Sensitivity Review Committee about specific items are kept with the items. The Bias and Sensitivity Review Committee makes the following recommendations regarding items:

- accept
- accept with edits (the committee identifies the nature of the issue causing 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 the outcome of the recommendations, the items can move to the next step in the development process. For a list of committee members, see Appendix A.

Review by External Content Expert Reviewers

When items are selected to be on the field-test portion of the MCAS, the items are submitted to expert reviewers for their feedback. The task of the expert reviewer is to consider the accuracy of the content of the item. 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 all affiliated with institutions of higher education either in 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 except for content.

3.2.5.3 Item Editing

After the ADC meetings to review new items, the items are edited in accordance with the ESE's decisions regarding the recommendations of the ADCs. Once the items have been developed and then reviewed by the ESE, the items are reviewed by Measured Progress editors. Measured Progress editors review and edit the items to ensure adherence to style guidelines in the *Chicago Manual of Style, 15th ed.*, 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 and ELA Passages

Items that have made it through the requisite reviews listed above are then 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, 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. Open-response items are 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 samples are scored.

3.2.5.6 Data Review of Field-Tested Items

Data Review by the Department of Elementary and Secondary Education

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

Data Review by Assessment Development Committees

The ADCs meet to review the items with their statistics. The ADCs consider the items and make one of the following recommendations regarding the field-tested items:

- 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 needs to go back into the field-test eligible pool to be re-field-tested. ADCs consider the following statistics 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 the differential item functioning by comparing the following subgroups of test takers:

- female/male

- black/white
- hispanic/white
- limited English proficient and formerly limited English proficient/native English speakers

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

3.2.5.7 Item and ELA Passage Selection and Operational Test Assembly

Measured Progress 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 the meeting, the test developers and psychometricians at Measured Progress considered 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 multiple-choice and constructed-response items 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 quality psychometric characteristics.
- **"Cueing" items**. Items are reviewed for any information that might "cue" or provide information that would help to answer another item.

The test developers then sort and lay out 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 single 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 Measured Progress test developers and psychometricians to construct the final versions of the tests.

3.2.5.9 Special Test Forms

All MCAS 2010 operational tests and retests were available in the following formats:

- Large-print – Form 1 of the operational test is used for translation into a large-print version. The large-print version 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.

The following special test formats were created only for the grade 10 MCAS Mathematics test and were made available to the students indicated:

- American Sign Language video – This video contains only the common items found in the operational test.
- Spanish/English version – This form of the 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.

Schools ordered special forms in advance of the testing. In order to be eligible to receive a special form, a student needs to have an IEP or a 504 plan, or have a 504 plan in development.

3.3 Test Administration

3.3.1 Test Administration Schedule

The standard MCAS tests were administered during three periods in the spring of 2010:

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

The 2010 MCAS administration also included retest opportunities in ELA and Mathematics for students in grades 11 and 12 who had not previously passed one or both grade 10 tests. Retests were offered in November 2009 and March 2010.

An additional high school (grades 9–12) end-of-course STE test in Biology was administered in February 2010, as both a standard test and a retest.

An additional one-time, special end-of-course testing opportunity was offered in April 2010 to students in the class of 2010 who had not yet taken or passed an MCAS high school STE test.

The grades 5 and 7 History and Social Science pilot tests and the high school (grades 10–11) U.S. History pilot test were suspended in 2009.

Table 3-18 shows the complete 2009–2010 MCAS test administration schedule.

Table 3-18. 2010 MCAS: Test Administration Schedule

Grade and Content Area	*Test Administration Date(s)*	*Deadline for Return of Materials to Contractor*
Retest Administration Windows		
November 4–10, 2009		
ELA Composition Retest	November 4	November 13
ELA Reading Comprehension Retest Sessions 1 and 2 Session 3	 November 5 November 6	
Mathematics Retest Session 1 Session 2	 November 9 November 10	
March 1–5, 2010		
ELA Composition Retest	March 1	March 9
ELA Reading Comprehension Retest Sessions 1 and 2 Session 3	 March 2 March 3	
Mathematics Retest Session 1 Session 2	 March 4 March 5	
March–April 2010 Test Administration Window		
Grades 3–8 ELA Reading Comprehension	March 22–April 12	April 14
Grades 4, 7, and 10 ELA Composition	March 23	
Grade 10 ELA Reading Comprehension Sessions 1 and 2 Session 3	 March 24 March 25	
Grades 4, 7, and 10 ELA Composition Make-Up	April 7	
May 2010 Test Administration Window		
Grades 3–8 Mathematics	May 10–27	May 28
Grades 5 and 8 Science and Technology/Engineering	May 11–27	
Grade 10 Mathematics Session 1 Session 2	 May 17 May 18	
High School (Grades 9–12) End-of-Course Science and Technology/Engineering Test Administration Windows		
February 1–2, 2010		
Biology	February 1–2	February 5
April 14, 2010		
Biology	April 14	April 15
Chemistry		
Introductory Physics		
Technology/Engineering		

Grade and Content Area	Test Administration Date(s)	Deadline for Return of Materials to Contractor
June 2–3, 2010		
Biology	June 2–3	June 8
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, establishing and following a document tracking system, and lists of approved and unapproved resource materials, can be found in the *Spring 2010 Principal's Administration Manual*, the *Fall 2009/Winter 2010 Principal's Administration Manual*, and all *Test Administrator's Manuals*.

3.3.3 Participation Requirements

In spring 2010, 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 broken arms
- students with limited English proficiency (LEP)

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 were 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:

- LEP 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 Alternate Assessment (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 with limited English proficiency, and/or students educated in alternate settings, can be found in the *Spring 2010 Principal's Administration Manual* and the *Fall 2009/Winter 2010 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

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 2010 Principal's Administration Manual* and the *Fall 2009/Winter 2010 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

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

Student identification information, demographic information, school contact information, and student answers to multiple-choice questions were converted to alphanumeric format. This information was not visible to scorers. Digitized student responses to short-answer, open-response, and writing-prompt test 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. 2010 MCAS: Summary of Scoring Locations and Scoring Shifts

Measured Progress Scoring Center, Content Area	*Grade(s)*	*Shift*	*Hours*
Menands, NY			
English Language Arts composition	7	Day	8:00 a.m.–4:00 p.m.
English Language Arts composition	10	Night	5:30 p.m.– 10:30 p.m.
Science and Technology/Engineering: Biology	HS (9–12)	Day	8:00 a.m.–4:00 p.m.
Science and Technology/Engineering: Biology	HS (9–12)	Night	5:30 p.m.–10:30 p.m.
Science and Technology/Engineering: Introductory Physics	HS (9–12)	Night	5:30 p.m.–10:30 p.m.
Longmont, CO			
English Language Arts reading comprehension	4, 7, 8, 10	Day	8:00 a.m.–4:00 p.m.
English Language Arts 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.

Measured Progress Scoring Center, Content Area	*Grade(s)*	*Shift*	*Hours*
Mathematics	4, 5, 6	Night	5:30 p.m.–10:30 p.m.
Dover, NH			
Science and Technology/Engineering: Chemistry	HS (9–12)	Day	8:00 a.m.–4:00 p.m.
Science and Technology/Engineering: Technology/Engineering	HS (9–12)	Day	8:00 a.m.–4:00 p.m.
Louisville, KY			
English Language Arts composition	4	Day	8:00 a.m.–4:00 p.m.
Science and Technology/Engineering	5, 8	Night	5:30 p.m.–10:30 p.m.

The following staff members were involved with scoring the 2010 MCAS responses:

- The **MCAS scoring project manager (SPM)** was located in Dover, New Hampshire, and oversaw communication and coordination of 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 scorers.

3.4.2.2 Benchmarking Meetings

Samples of student responses to field-test items were read, scored, and discussed by members of Scoring Services, Curriculum and Assessment, and the ESE 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 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, or anyone 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 the 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 scorer background across all scoring shifts at all scoring locations.

Table 3-20. 2010 MCAS: Summary of Scorer Background 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	218	10.4
Bachelor's degree	1241	59.2
Master's degree/doctorate	637	30.4
Teaching Experience		
No teaching certificate or experience	1083	51.7
Teaching certificate or experience	846	40.4
College instructor	167	8.0
Scoring Experience		
No previous experience as reader	1109	52.9
1–3 years experience	642	30.6
3+ years experience	345	16.5

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; and open-response items requiring a longer or more complex response, with assigned scores of 0–4, or, for ELA composition, 1–4 and 1–6.

The sample below of a 4-point mathematics OR scoring guide was one of the many different item-specific MCAS scoring guides used in 2010. The task associated with this scoring guide asked students to design four different gardens, each with a different shape.

Table 3-21. 2010 MCAS: 4-point Open-Response 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 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 readers and iScore staff by matching all responses with the correct item and/or pulling the actual test booklet to look at the student's original work.

Scorers may have also "flagged" a response as a "Crisis" response, which was 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, and/or sexual abuse;
- presence of dark thoughts or serious depression;
- sexual knowledge well outside of 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, in which each response was scored only once, or double-blind scored, in which each response was independently read and scored by two separate readers. In double-blind scoring, neither reader knew whether or not 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 readers (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 a senior reader (SR) or quality assurance coordinator (QAC).

Polytomous items on all high school tests (ELA, Mathematics, and STE), as well as the ELA composition at grades 4, 7, and 10, are 100 percent double-blind scored. Polytomous items on the ELA reading comprehension, Mathematics, and STE tests at grades 3–8 are 10 percent double-blind scored.

Above and beyond the 10 or 100 percent double-blind scoring, SRs, at random points throughout the scoring shift, engaged in read-behind scoring for each of the readers at his or her table. This process involved SRs viewing responses recently scored by a particular reader, and, without knowing the reader's score, assigning his or her own score to that same response. The SR would then compare scores and advise or counsel the reader 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. 2010 MCAS: Read-Behind and Double-Blind Resolution Charts

Read-Behind Scoring*			
Reader #1	*Reader #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.			

Double-Blind Scoring*			
Reader #1	*Reader #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 reader scores are identical or adjacent, the highest score is used as the final score. If reader scores are neither identical nor adjacent, the resolution score is used as the final score.			

Writing English Conventions Double-Blind Scoring*			
Reader #1	*Reader #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*				
Reader #1	*Reader #2*	*QAC/SR Resolution*	*Chief Reader*	*Final*
6	6	-	-	12
6	5	-	-	11
6	4	4	-	8
6	4	5	-	11
6	2	4	4	8
6	2	4	3	6
6	2	3	-	5

* 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. 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

Chief readers had overall responsibility for ensuring that readers scored responses consistently, fairly, and only 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 several possible modes of delivery. 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 more robust computer servers, scorers were trained on some items via a remote location; that is, the chief reader 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 would continue uninterrupted, either through the 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. Actual 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 matrix 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 more typical 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 percent exact and 90 percent exact plus adjacent agreement (at least 7 out of the 10 were exact score matches and either 0 or 1 discrepant) on either of two potential qualifying sets.

3.4.2.6 Leadership Training

Chief readers also had overall responsibility for ensuring that scoring leadership (QACs and SRs) scored consistently, fairly, and only according to the approved scoring guidelines. Scoring leadership must have met or surpassed the higher qualification standard of at least 80 percent exact and 90 percent exact plus adjacent, or for grade 10 leadership, at least 80 percent exact and 100 percent 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 percent exact and 90 percent 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 (CRs), an end-of-shift report combining RAs and RBs

Recalibration assessments 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 4 of the 5 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 2 or 3 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 0 or 1 out of the 5 exact were typically reassigned to another item or sent home for the day.

Embedded committee-reviewed responses (CRRs) were responses approved by the chief reader 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 percent exact and 90 percent 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.

Read-behinds involved responses that were first read and scored by a reader, then read and scored by an SR. Senior readers would, at various points during the scoring shift, command iScore to forward the next 1, 2, or 3 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. Read-behinds were performed at least 10 times for each full-time day shift reader and at least 5 times for each evening shift and partial-day shift reader. Readers who fell below the 70 percent exact and 90 percent exact plus adjacent score match standard were counseled, given extra monitoring assignments such as additional RBs, and allowed to resume scoring.

Double-blind readings 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 percent exact and 90 percent 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 a senior reader.

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/senior reader RBs. Once the SR completed the minimum number of required RBs for a reader, the reader's overall percentages on the CRs were automatically calculated. If the CR at the end of the scoring shift listed individuals who were still below the 70 percent exact/90 percent 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 CR, 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 2010. 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 constructed-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. Constructed-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 constructed-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 do. 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 constructed-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.

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 constructed-response). The mean difficulty 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. 2010 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*
English Language Arts	3	ALL	41	0.77	0.12	0.42	0.06
		MC	36	0.80	0.09	0.42	0.05
		OR	5	0.58	0.12	0.42	0.11
	4	ALL	42	0.75	0.12	0.42	0.10
		MC	36	0.78	0.09	0.39	0.07
		OR	6	0.57	0.15	0.60	0.05
	5	ALL	40	0.76	0.12	0.42	0.07
		MC	36	0.78	0.10	0.41	0.05
		OR	4	0.54	0.04	0.53	0.06
	6	ALL	40	0.75	0.12	0.40	0.08
		MC	36	0.77	0.11	0.38	0.06
		OR	4	0.55	0.04	0.57	0.04
	7	ALL	42	0.75	0.11	0.43	0.10
		MC	36	0.77	0.10	0.40	0.05
		OR	6	0.63	0.11	0.64	0.03
	8	ALL	40	0.77	0.09	0.43	0.08
		MC	36	0.78	0.08	0.41	0.06
		OR	4	0.63	0.06	0.62	0.02
	10	ALL	42	0.77	0.10	0.41	0.13
		MC	36	0.79	0.08	0.36	0.08
		OR	6	0.65	0.11	0.66	0.02
Mathematics	3	ALL	36	0.74	0.11	0.43	0.08
		MC	26	0.77	0.10	0.41	0.06
		OR	10	0.68	0.11	0.47	0.10
	4	ALL	42	0.71	0.11	0.41	0.09
		MC	32	0.74	0.10	0.39	0.06
		OR	10	0.62	0.11	0.50	0.10
	5	ALL	42	0.69	0.14	0.47	0.08
		MC	32	0.70	0.15	0.44	0.06
		OR	10	0.64	0.10	0.54	0.10
	6	ALL	42	0.71	0.12	0.47	0.09
		MC	32	0.73	0.12	0.45	0.07
		OR	10	0.65	0.10	0.55	0.12

Content area	Grade	Item type	Number of items	p-*Value*		*Discrimination*	
				Mean	*Standard deviation*	*Mean*	*Standard deviation*
Mathematics	7	ALL	42	0.68	0.14	0.47	0.11
		MC	32	0.68	0.15	0.44	0.07
		OR	10	0.69	0.11	0.57	0.15
	8	ALL	42	0.71	0.14	0.47	0.12
		MC	32	0.74	0.13	0.43	0.10
		OR	10	0.61	0.11	0.59	0.10
	10	ALL	42	0.66	0.11	0.47	0.13
		MC	32	0.65	0.12	0.41	0.07
		OR	10	0.69	0.06	0.64	0.14
STE	5	ALL	42	0.71	0.14	0.35	0.11
		MC	38	0.73	0.13	0.33	0.08
		OR	4	0.54	0.15	0.56	0.05
	8	ALL	42	0.66	0.14	0.38	0.10
		MC	38	0.68	0.13	0.36	0.08
		OR	4	0.50	0.11	0.54	0.06
Biology	HS	ALL	45	0.67	0.14	0.41	0.10
		MC	40	0.70	0.11	0.38	0.08
		OR	5	0.45	0.11	0.59	0.07
Chemistry	HS	ALL	45	0.62	0.13	0.43	0.11
		MC	40	0.63	0.13	0.41	0.08
		OR	5	0.49	0.05	0.67	0.06
Introductory Physics	HS	ALL	45	0.62	0.14	0.40	0.11
		MC	40	0.64	0.13	0.37	0.07
		OR	5	0.44	0.06	0.65	0.07
Technology/ Engineering	HS	ALL	45	0.61	0.17	0.33	0.11
		MC	40	0.64	0.15	0.31	0.10
		OR	5	0.38	0.06	0.51	0.08

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 because of 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 constructed-response items because multiple-choice items can be answered correctly by guessing. Similarly, discrimination indices for the four-point constructed-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 in 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 near-zero discrimination indices, but none was 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 constructed-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 because of 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, MCAS items were evaluated in terms of differential item functioning (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 could 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 constructed-response items. Dorans and Holland (1993) suggested that index values between -0.05 and 0.05 should be considered negligible. The preponderance 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.l0 (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]

---

[2] It should be pointed out here that 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.

For the 2010 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 versus female
- white versus black
- white versus Hispanic
- no disability versus disability
- not LEP/FLEP versus LEP/FLEP
- not low-income versus low-income

The tables in Appendix G present the number of items classified as either “low” or “high” DIF, overall and by group favored. Overall, a moderate number of items exhibited low DIF and very few exhibited high DIF; in addition, 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 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 3–8 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 2010 administration. A total of 20 tests was analyzed. The results for these analyses are reported below, including a comparison with the results from 2009.

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 overall possible conditioning scores. When a test is strictly unidimensional, all conditional covariances are expected to take on values near zero, indicating statistically independent item responses for examinees with equal expected scores. Nonzero 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 2010 (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 69,000 student examinees per test. For the high school tests, Mathematics and ELA each had over 69,000 student examinees, Biology had over 50,000, Introductory Physics had over 18,000, and Chemistry and Technology/Engineering had approximately 2,100 each. Because DIMTEST was limited to using 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 data set to see if the null hypothesis of unidimensionality would be rejected. DETECT was then applied to each data set 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 data set. Because strict unidimensionality is an idealization that almost never holds exactly for a given data set, the statistical rejections in the DIMTEST results were not surprising. Indeed, because of the very large sample sizes involved in most of the data sets (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 displays the multidimensionality effect-size estimates from DETECT.

Table 3-24. 2010 MCAS: Multidimensionality Effect Sizes by Grade and Content Area

Content Area	*Grade*	*Multidimensionality Effect Size*	
		2009	*2010*
ELA	3	0.11	0.12
	4	0.16	0.20
	5	0.12	0.14
	6	0.14	0.14
	7	0.16	0.13
	8	0.19	0.19
	10	0.18	0.12
	Average	**0.15**	**0.15**
Mathematics	3	0.14	0.12
	4	0.18	0.11
	5	0.19	0.19
	6	0.12	0.14
	7	0.17	0.12
	8	0.16	0.16
	10	0.17	0.16
	Average	**0.16**	**0.14**
STE	5	0.13	0.09
	8	0.14	0.13
	Biology (grades 9–12)	0.07	0.11
	Chemistry (grades 9–12)	0.10	0.10
	Introductory Physics (grades 9–12)	0.14	0.13
	Technology/Engineering (grades 9–12)	0.16	0.17
	Average	**0.12**	**0.12**

The DETECT values indicate very weak multidimensionality for all the tests for 2010. The ELA test forms (average effect size of about 0.15) and the Mathematics test forms (average of about 0.14) tended to show slightly greater multidimensionality than did the STE test forms (average of about 0.12). Also shown in Table 3-24 are the values reported in last year's dimensionality analyses. The averages last year for Mathematics and ELA were 0.16 and 0.15, respectively, and the average for the STE tests was 0.12. Thus, last year's results are seen to be 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 (MC) and constructed-response (CR) item types. Inspection of the DETECT clusters indicated that MC-CR separation generally occurred much more strongly with ELA than with Mathematics or STE, a pattern that has been consistent across all four years of dimensionality analyses for the MCAS tests. Specifically, for ELA, except for grade 3, every grade had one set of clusters dominated by MC items and another set of clusters dominated by CR items. This particular pattern within ELA has occurred in all four years of the MCAS dimensionality analyses. Of the seven Mathematics tests, only grades 5 and 7 showed evidence of consistent separation of MC and CR items (in grade 8, three out of four CR items were clustered separately from the MC items). Of the six STE tests, no test showed strong MC-CR separation, although the grade 8 test did have three out of four CR items clustered separately from the MC items. In comparison to past years, no single grade has had consistent MC-CR separation every year within the Mathematics or STE content areas.

Thus, a tendency is suggested for MC and CR to sometimes measure statistically separable dimensions, especially in regard to the ELA tests. This has been consistent across all four years of MCAS analyses. However, it is important to emphasize that the degree of violation of unidimensional local independence has been quite similar across the three content areas over the four years of analysis. Also, 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 STE. 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 2010 analyses the violations of local independence, as evidenced by the DETECT effect sizes, were 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 has been consistently low over the years, and the patterns with respect to the MC and CR 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. During the course of these psychometric analyses, a number of quality control procedures and checks on the processes were implemented. 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); and evaluation of the scaling results (e.g., parallel processing by the Psychometrics and Research and Data Analysis departments; 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 student reports (Measured Progress Department of Psychometrics and Research, *2009–2010 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 item was flagged as a result of the delta analyses) and what action was taken. The number of items identified for evaluation was generally fairly typical across the grades and content areas. Descriptions of the evaluations and results are included in the Item Response Theory Results and Equating Results sections below.

Table 3-25. 2010 MCAS: Items That Required Intervention During IRT Calibration and Equating

Item Number	*Content Area*	*Grade*	*Reasons*	*Action*
270156	ELA	03	*c* parameter	*c* = 0
279400	ELA	03	*c* parameter	*c* = 0
220200	ELA	03	Delta Analysis; Position Change	Removed from equating
231830	ELA	03	IRT Plot Outlier; Position Change	Removed from equating
255585	ELA	04	*c* parameter	*c* = 0
255596	ELA	04	*c* parameter	*c* = 0
268949	ELA	04	*c* parameter	*c* = 0
268952	ELA	04	*c* parameter	*c* = 0
256056	ELA	04	*c* parameter	*c* = 0
260416	ELA	04	*c* parameter	*c* = 0
287624	ELA	05	*c* parameter	*c* = 0
287625	ELA	05	*c* parameter	*c* = 0
276125	ELA	05	*c* parameter	*c* = 0
270993	ELA	05	*c* parameter	*c* = 0
270193	ELA	05	*c* parameter	*c* = 0
280144	ELA	05	*c* parameter	*c* = 0
208724	ELA	05	Delta Analysis; IRT Plot Outlier	Removed from equating
257668	ELA	06	*c* parameter	*c* = 0
277124	ELA	06	*c* parameter	*c* = 0
277127	ELA	06	*c* parameter	*c* = 0
277142	ELA	06	*c* parameter	*c* = 0
277144	ELA	06	*c* parameter	*c* = 0
257301	ELA	06	*c* parameter	*c* = 0
203343	ELA	06	Delta Analysis; Position Change	Removed from equating
207499	ELA	06	Delta Analysis; Position Change	Removed from equating
203352	ELA	06	IRT Plot Outlier; Position Change	Removed from equating
279211	ELA	07	*c* parameter	*c* = 0
280270	ELA	07	*c* parameter	*c* = 0
280279	ELA	07	*c* parameter	*c* = 0
271298	ELA	07	*c* parameter	*c* = 0
271310	ELA	07	*c* parameter	*c* = 0
266568	ELA	07	*c* parameter	*c* = 0
266565	ELA	07	*c* parameter	*c* = 0
275961	ELA	08	*c* parameter	*c* = 0
275968	ELA	08	*c* parameter	*c* = 0
275981	ELA	08	*c* parameter	*c* = 0
271334	ELA	08	*c* parameter	*c* = 0
276060	ELA	08	*c* parameter	*c* = 0
276087	ELA	08	*c* parameter	*c* = 0
227774	ELA	08	*c* parameter	*c* = 0
207609	ELA	08	Delta Analysis	Removed from equating
280204	ELA	10	Delta Analysis	Retained for equating
280204	ELA	10	*c* parameter	*c* = 0
280216	ELA	10	*c* parameter	*c* = 0

continued

Item Number	Content Area	Grade	Reasons	Action
280221	ELA	10	*c* parameter	*c* = 0
280228	ELA	10	*c* parameter	*c* = 0
253867	ELA	10	*c* parameter	*c* = 0
279390	ELA	10	*c* parameter	*c* = 0
279553	ELA	10	*c* parameter	*c* = 0
279560	ELA	10	*c* parameter	*c* = 0
279561	ELA	10	*c* parameter	*c* = 0
268741	ELA	10	*c* parameter	*c* = 0
268740	ELA	10	*c* parameter	*c* = 0
268739	ELA	10	Delta Analysis	Retained for equating
276594	ELA	10	*c* parameter	*c* = 0
276600	ELA	10	*c* parameter	*c* = 0
254865	ELA	10	*c* parameter	*c* = 0
203481	MAT	03	*c* parameter	*c* = 0
260977	MAT	03	*c* parameter	*c* = 0
218517	MAT	03	Delta Analysis	Removed from equating
221120	MAT	05	Position Study	Removed from equating
260821	MAT	05	Position Study	Removed from equating
223157	MAT	06	*c* parameter	*c* = 0
265870	MAT	06	*c* parameter	*c* = 0
227450	MAT	06	IRT Plot Outlier	Removed from equating
273841	MAT	06	a parameter	a set to initial
219513	MAT	07	*c* parameter	*c* = 0
208618	MAT	07	IRT Plot Outlier	Removed from equating
276326	MAT	08	*c* parameter	*c* = 0
228118	MAT	08	*c* parameter	*c* = 0
288136	MAT	08	*c* parameter	*c* = 0
261484	MAT	10	*c* parameter	*c* = 0
254107	MAT	10	Delta Analysis	Removed from equating
229964	MAT	10	Delta Analysis	Removed from equating
229594	MAT	10	Delta Analysis	Removed from equating
273779	STE	05	*c* parameter	*c* = 0
281799	STE	05	*c* parameter	*c* = 0
281818	STE	05	*c* parameter	*c* = 0
264809	STE	05	*c* parameter	*c* = 0
281840	STE	05	*c* parameter	*c* = 0
273774	STE	05	*c* parameter	*c* = 0
264848	STE	05	*c* parameter	*c* = 0
281820	STE	05	*c* parameter	*c* = 0
282046	STE	08	*c* parameter	*c* = 0
229491	STE	08	*c* parameter	*c* = 0
208026	BIO	10	Delta Analysis	Retained—Pre-Equated
273830	BIO	10	*c* parameter	*c* = 0
243541	BIO	10	*c* parameter	*c* = 0

3.6.1 Item Response Theory

All MCAS items were calibrated using item response theory (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 specific 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 2010 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 test might have too few examinees to support the 3PL model in future administrations.

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

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

where
i indexes the items,
j indexes students,
a represents item discrimination,
b represents item difficulty,
c is the pseudo guessing parameter,
ξ_i 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:

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

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 (ICTC) 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,
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:

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

where
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
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.

$$P_i(1|\theta_j) = \sum_{k}^{m+1} w_{ik} P_{ik}(1|\theta_j)$$

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 Results

The tables in Appendix H give the IRT item parameters and associated standard errors of all common items on the 2010 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) was 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 shows graphs of the test characteristic curves (TCCs) and test information functions (TIFs), which are defined below. Note that, because of the use of the one-parameter 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

$$E(X \mid \theta_j) = \sum_{i=1}^{n} P_i\left(1 \middle| \theta_j\right),$$

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
$E(X \mid \theta_j)$ 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 do students of low ability. Most TCCs are "S-shaped": 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 the 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.

Table 3-26. 2010 MCAS: Number of Newton Cycles Required for Convergence

Content Area	*Grade*	*Cycles*
English Language Arts	3	69
	4	42
	5	41
	6	45
	7	60
	8	43
	10	46
Mathematics	3	31
	4	34
	5	46
	6	27
	7	34
	8	16
	10	50
STE	5	35
	8	34
Biology	HS	20
Chemistry	HS	20
Introductory Physics	HS	20
Technology/Engineering	HS	200

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 regarding how the test development supports successful equating annually.

The 2010 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). 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 2010 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 (that is, naturally occurring groups are assumed). Comparability is instead evaluated by utilizing a set of anchor items (also called equating items). However, the equating items are designed to mirror the common test in terms of item types and distribution of emphasis. Subsets of the equating items are matrixed across forms.

Item parameter estimates for 2010 were placed on the 2009 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 2009 and 2010 MCAS tests should have the same item parameters. After the item parameters for each 2010 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 non-equating 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.

3.6.4 Equating Results

Prior to equating the 2010 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. These items were scrutinized, and a decision was made as to whether to include the 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 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. With 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. All effect sizes were well below the criterion value for excluding an item as an equating item, 0.80.

Finally, *a*-plots and *b*-plots, which show IRT parameters for 2010 plotted against the values for 2009, 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 (2011) will be scaled to the 2010 administration using the same equating method described above.

3.6.4.1 MCAS Special Equating Item Evaluation Study

As described in Section 3.2 of this report, some changes were made to the test design for grades 3–8 in 2010. In order to ensure robust and accurate equating of the 2010 test to the operational scale, a special equating item analysis study was conducted to supplement the normal equating item evaluation analyses described above.

First, the delta analysis and *b*-plot analysis were repeated with additional items included, beyond the set of equating items. Specifically, all items that had been previously administered were included in the analyses. There were between 12 and 24 extra items available for each test. Using a larger pool of items to rerun the analyses provided greater power in detecting items that were potentially problematic for inclusion as equating items. As a result of expanding the pool of items used for the analyses, two additional items were flagged for removal from the equating set, both from the grade 5 Mathematics test (see Table 3-25).

In addition to the extended *b*-plot and delta analyses, a position effect study was conducted. The position effect study tracked the change in relative position in the test booklet for items that were administered in 2010 as well as in a previous year. Relative position was defined as position relative to the start of the session in which the item was administered. Therefore, an item that was administered in the third position of the second session was identified as having a relative position of 3. Any item that moved more than 5 relative positions between the previous administration and the current administration was flagged for careful scrutiny. Very few items from the ELA reading comprehension test, and no items from the STE test, moved more than 5 relative positions. On average, across grade levels, just over 10 items moved more than 5 relative positions in the Mathematics tests. Of the items that moved more than 5 relative positions, two were eliminated that had not already been flagged by the *b*-plot, delta, or rescore analyses, and the retained ones had statistical properties that were consistent with items that moved fewer than 5 relative positions.

Tables L-1 to L-14 in Appendix L show the statistics for all previously administered items, both those used for equating and those used for the special evaluation study. For each item, the following details are provided: item number, standardized distance from the regression line of the delta value and *b*-parameter value, the relative old and new position of the item, the change in relative position, and whether the item was used as an equating item.

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. Note that these operational θ-metric cut scores will remain fixed throughout the assessment program unless standards are reset for any reason. Also shown in the table are the cutpoints on the reporting score scale (2007 Standard Setting Report).

Table 3-27. 2010 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*
English Language Arts	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	HS	-0.962	-0.129	1.043	200	220	240	260	280
Chemistry	HS	-0.134	0.425	1.150	200	220	240	260	280
Introductory Physics	HS	-0.714	0.108	1.133	200	220	240	260	280

Content Area	Grade	Theta				Scaled score			
		Cut 1	Cut 2	Cut 3	Min	Cut 1	Cut 2	Cut 3	Max
Technology/ Engineering	HS	-0.366	0.201	1.300	200	220	240	260	280

Appendix M shows performance 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 readily 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 performance 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 performance 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 performance level scores. Students' raw scores (i.e., total number of points) on the 2010 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 Fahrenheit or Celsius scales, or the same distance can be expressed in either miles or kilometers, student scores on the 2010 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' performance 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. 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 score*s 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 performance level. Table 3-28 shows the slope and intercept terms used to calculate the scaled scores for each grade, content area, and performance level. Note that the values in Table 3-28 will not change unless the standards are reset.

Appendix N 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 O 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. 2010 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*
English Language Arts	3	5.3918	229.1229	13.7552	243.2737	14.6413	243.4846	10.6838	247.9487
	4	5.2725	225.9368	16.7645	238.8768	13.2890	239.1096	14.0056	237.9832
	5	5.7614	228.8437	15.5400	243.8539	14.2857	243.5429	10.8225	247.5325
	6	5.2883	227.2979	18.1653	245.0681	11.9689	243.3393	12.4378	242.6866
	7	5.7824	228.8413	17.5593	246.8481	10.8108	244.2162	12.9870	241.0390
	8	6.7385	231.2263	19.4363	252.3810	10.9529	246.9770	11.0436	246.8691
	10	3.8002	221.5733	25.0627	230.3759	19.1205	232.6577	12.7389	241.7834
Mathematics	3	6.0274	226.0937	21.6450	241.8831	17.8891	241.5564	10.1574	249.5277
	4	5.7566	224.9449	15.2905	233.1346	23.2829	229.5460	11.8203	244.5390
	5	5.8183	224.1542	22.6244	236.1538	22.7531	236.1320	10.2512	249.2465
	6	5.9539	223.0365	26.9542	233.7466	22.7273	234.7273	10.5932	248.2203
	7	5.5864	222.7094	26.7023	232.9506	21.5983	234.2981	11.0497	246.8508
	8	5.3017	221.6859	27.1739	228.6413	22.7273	230.5000	11.7509	244.7474
	10	4.1034	220.7755	32.8407	226.2069	32.3625	226.4078	10.1937	249.4190
STE	5	5.0250	225.6783	16.3934	238.5246	20.0000	238.2000	10.4712	248.5864
	8	4.4558	222.2279	19.2308	229.6154	14.9254	231.9403	17.8571	226.4286
Biology	HS	4.8493	224.6651	24.0096	243.0972	17.0648	242.2014	10.2197	249.3408
Chemistry	HS	4.3728	220.5860	35.7782	224.7943	27.5862	228.2759	10.8108	247.5676
Introductory Physics	HS	4.3347	223.0950	24.3309	237.3723	19.5122	237.8927	10.7124	247.8629
Technology/ Engineering	HS	7.3692	222.6971	35.2734	232.9101	18.1984	236.3421	11.7647	244.7059

3.6.6.1 Reported Scores for High School Biology

An issue arose with the scoring and reporting of the February 2010 high school Biology test. Specifically, two of the cut scores that were used for reporting differed slightly from the cuts identified at the time of form construction. Because the high school STE tests are pre-equated, the operational cuts should not have differed from the original cuts. The issue was discovered after official MCAS reports were delivered to schools and districts. After being alerted to the issue, Measured Progress psychometric staff investigated and found that, in the process of importing item parameters from the SQL database where they are loaded into the program used to run item calibrations, the difficulty value for one constructed-response item was incorrectly transferred.

As a result of the incorrect item parameter value, the scaled scores associated with 11 of the 61 raw score points were two points higher than they should have been (because MCAS scaled scores are reported as even integers). Consequently, students who scored at those raw score points received slightly higher reported scores than they should have. The overall impact was that 723 students (4.6 percent of the tested population) were incorrectly categorized into the *Needs Improvement* performance level (instead of *Failing*), and 109 students (0.7 percent) were incorrectly categorized as *Advanced* (instead of *Proficient*). For all students, the revised score was within one standard error of the original score. There was no impact on the cut between *Needs Improvement* and *Proficient*. Because no students were disadvantaged as a result of the slight elevations in scaled scores, and because there was no effect on the number of students categorized as *Proficient* or above, the ESE decided not to issue corrected score reports.

The incident occurred because the high school Biology test is pre-equated and was not processed within Measured Progress's standard psychometric systems. After being alerted to the issue, Psychometrics and Research staff thoroughly checked all of the pre-equated MCAS tests and verified that the problem was unique to Biology. In the future, to ensure that any data transfer problems will be found and corrected prior to reporting, psychometric staff will implement an additional quality control step comparing the cuts identified at the time of form construction with those obtained for the operational forms.

3.7 MCAS Reliability

Although an individual item's performance is an important focus for evaluation, a complete evaluation of an assessment must also address the way items function together and complement one another. Tests that function well provide a dependable assessment of the student's level of ability. Unfortunately, no test can do this perfectly. A variety of factors can contribute to a given student's score being either 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 knew the answer. Collectively, extraneous factors that affect a student's score are referred to as measurement error. 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 other students 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 or 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 made are small on average and students' scores on such a test will consistently represent their ability) are described as reliable.

There are a number of ways to estimate an assessment's reliability. One possible approach 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. (This is referred to as "test-retest reliability.") A potential 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. A solution to the "remembering items" problem 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 is known as "alternate forms reliability," because an alternate form of the test is used in each administration.) This approach, however, does not address the problem 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 generally preclude the use of parallel forms reliability indices. One way to address the latter two problems 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. By doing this, the problems associated with an intervening time interval and with creating and administering two parallel forms of the test are alleviated. 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, α (alpha), 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 2010 MCAS:

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

where
i indexes the item,
n is the total number of items,
$\sigma^2_{(Y_i)}$ 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 standard errors of measurement (SEMs) for each content area and grade. (Statistics are based on common items only.) Generally, reliability estimates are in acceptable ranges and consistent with results obtained for previous administrations of the test.

Table 3-29. 2010 MCAS: Raw Score Descriptive Statistics, Cronbach's Alpha, and Standard Errors of Measurement (SEM) by Content Area and Grade

Content Area	*Grade*	*Number of Students*	*Raw score*			*Alpha*	*SEM*
			Maximum	*Mean*	*Standard Deviation*		
English Language Arts	3	69,417	48	35.20	7.95	0.90	2.51
	4	69,561	72	49.81	10.62	0.90	3.39
	5	69,769	52	36.74	8.48	0.90	2.69
	6	71,008	52	36.68	8.31	0.89	2.74
	7	70,146	72	50.88	10.61	0.91	3.24
	8	71,039	52	38.13	9.19	0.90	2.84
	10	69,753	72	52.12	10.23	0.90	3.25
Mathematics	3	69,347	40	29.53	7.64	0.90	2.42
	4	69,591	54	37.26	10.26	0.90	3.26
	5	69,660	54	37.04	11.61	0.92	3.33
	6	70,950	54	37.59	11.59	0.92	3.28
	7	70,168	54	36.84	11.79	0.92	3.29
	8	70,897	54	36.45	12.10	0.92	3.52
	10	69,792	60	39.88	13.32	0.92	3.68
STE	5	69,828	54	36.24	8.94	0.87	3.28
	8	70,879	54	33.70	9.81	0.88	3.43
Biology	HS	55,581	60	35.48	11.92	0.91	3.55
Chemistry	HS	2,094	60	30.50	14.31	0.93	3.77
Introductory Physics	HS	18976	60	33.85	11.85	0.91	3.64
Technology/ Engineering	HS	2178	60	31.66	10.49	0.87	3.78

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 and/or content area.

3.7.2 Inter-Rater Consistency

Section 3.4.2 of this report describes in detail 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 percent (for compositions and all high school tests) or 10 percent (all other constructed-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 that required retraining or other intervention and are presented here as evidence of the reliability of the MCAS. A summary of the inter-rater consistency results are presented in Table 3-30 below. Results in the table are collapsed 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 P. These inter-rater consistency statistics are the result of the processes implemented to ensure valid and reliable hand-scoring of items as described in detail in Section 3.4.2.

Table 3-30. 2010 MCAS: Summary of Inter-Rater Consistency Statistics Collapsed 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*
English Language Arts	3	4	3	27,462	84.24	15.49	0.78	0.27
		1	5	6,911	63.88	34.67	0.69	1.42
	4	1	4	68,040	73.21	26.36	0.69	1.30
		4	5	27,401	62.80	35.47	0.77	1.59
		1	6	68,040	66.87	32.17	0.76	1.30
	5	4	5	27,621	62.26	35.55	0.71	1.95
	6	4	5	27,965	63.70	34.81	0.74	1.34
	7	1	4	67,241	65.88	33.52	0.57	1.54
		4	5	27,157	64.03	34.61	0.77	1.20
		1	6	67,241	64.88	34.12	0.68	1.54
	8	4	5	27,500	58.96	38.65	0.76	2.04
	10	1	4	66,223	67.53	31.89	0.55	1.46
		4	5	269,895	62.18	35.67	0.77	2.30
		1	6	66,223	64.28	34.76	0.62	1.46
Mathematics	3	6	2	41,331	98.06	1.94	0.96	0.00
		4	3	27,665	93.71	6.21	0.92	0.07
	4	6	2	41,613	98.69	1.31	0.97	0.00
		4	5	27,679	74.72	22.84	0.89	2.28
	5	6	2	41,620	98.50	1.50	0.97	0.00
		4	5	27,788	76.90	20.69	0.92	2.35
	6	6	2	42,307	98.83	1.17	0.97	0.00
		4	5	28,234	82.88	15.32	0.93	1.75
	7	6	2	41,631	98.51	1.49	0.96	0.00
		4	5	27,766	87.15	12.10	0.95	0.75
	8	6	2	41,982	98.81	1.19	0.97	0.00
		4	5	27,840	85.95	12.80	0.96	1.23
	10	4	2	27,3209	98.72	1.28	0.97	0.00
		6	5	40,9955	85.14	14.03	0.94	0.78
STE	5	4	5	27,875	68.85	27.68	0.87	3.33
	8	4	5	28,179	71.27	24.62	0.87	3.99
Biology	HS	5	5	264,807	73.47	23.26	0.87	3.28
Chemistry	HS	5	5	9,753	82.67	15.72	0.94	1.54
Introductory Physics	HS	5	5	89,648	71.16	26.03	0.87	2.65
Technology/ Engineering	HS	5	5	9,835	67.98	26.61	0.84	4.70

3.7.3 Subgroup Reliability

The reliability coefficients discussed in the previous section were based on the overall population of students who took the 2010 MCAS. Appendix Q presents reliabilities for various subgroups of interest. Subgroup Cronbach’s α’s 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 of 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 Q 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.

3.7.4 Reporting Subcategory Reliability

Of even more interest are reliabilities for the reporting subcategories within MCAS content areas, 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 Q. Once again, as expected, because they are based on a subset of items rather than the full test, computed subcategory reliabilities were lower (sometimes substantially so) than were overall test reliabilities, 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 quality of the full test based on statistical comparisons among subtests.

3.7.5 Reliability of Performance Level Categorization

While related to reliability, the accuracy and consistency of classifying students into achievement categories are even more important statistics in a standards-based reporting framework (Livingston & Lewis, 1995). After the performance levels were specified and students were classified into those levels, decision accuracy and consistency (DAC) analyses were conducted to determine the statistical accuracy and consistency of the classifications. For the MCAS, students are classified into one of four performance levels: *Warning* (*Failing* at grade 10), *Needs Improvement*, *Proficient*, or *Advanced* (*Above Proficient* at grade 3). This section of the chapter explains the methodologies used to assess the reliability of classification decisions, and results are given. Section 3.2 describes the reporting categories in greater detail.

Accuracy refers to the extent to which decisions based on test scores match decisions that would have been made 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 classification decisions based on test scores match the decisions 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 given to the same group of students. In operational test programs, however, such a design is usually impractical. Instead, techniques have been developed to estimate both the accuracy and consistency of classification decisions based on a single administration of a test. The Livingston and Lewis (1995) technique was used for the 2010 MCAS because it is easily adaptable to all types of testing formats, including mixed-format tests.
The DAC estimates reported in Appendix R 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. Of course, 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 2010 MCAS, 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.

Another way to measure consistency is to use 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 performance 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 performance 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 performance 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 are other consistency estimates.

3.7.6 Decision Accuracy and Consistency Results

The decision accuracy and consistency 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 performance level are also given. For these calculations, the denominator is the proportion of students associated with a given performance level. For example, the conditional accuracy value is 0.74 for *Needs Improvement* for grade 3 Mathematics. This figure indicates that among the students whose true scores placed them in this classification, 74 percent would be expected to be in this classification when categorized according to their observed scores. Similarly, a consistency value of 0.66 indicates that 66 percent 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 used.

For some testing situations, the greatest concern may be decisions around level thresholds. For example, in testing done for NCLB accountability purposes, the primary concern is distinguishing between students who are proficient and those who are not yet proficient. In this case, the accuracy of the Approaches Standard/Meets Standard threshold is of greatest interest. For the 2010 MCAS, Table 3-32 provides accuracy and consistency estimates 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. 2010 MCAS: Summary of Decision Accuracy (and Consistency) Results by Content Area and Grade—Overall and Conditional on Performance Level

Content Area	*Grade*	*Overall*	*Kappa*	*Conditional on Performance Level*			
				*Warning**	*Needs Improvement*	*Proficient*	*Advanced*
English Language Arts	3	0.74 (0.67)	0.49	0.82 (0.73)	0.82 (0.75)	0.71 (0.68)	0.64 (0.47)
	4	0.80 (0.72)	0.59	0.80 (0.69)	0.81 (0.75)	0.77 (0.70)	0.85 (0.70)
	5	0.80 (0.72)	0.59	0.78 (0.66)	0.80 (0.74)	0.77 (0.70)	0.87 (0.74)
	6	0.80 (0.72)	0.57	0.77 (0.64)	0.75 (0.67)	0.80 (0.75)	0.86 (0.72)
	7	0.84 (0.78)	0.62	0.78 (0.65)	0.79 (0.72)	0.86 (0.83)	0.86 (0.71)
	8	0.80 (0.73)	0.55	0.79 (0.68)	0.75 (0.66)	0.82 (0.79)	0.79 (0.64)
	10	0.83 (0.76)	0.62	0.75 (0.57)	0.79 (0.72)	0.82 (0.78)	0.88 (0.78)
Mathematics	3	0.78 (0.70)	0.57	0.80 (0.71)	0.74 (0.66)	0.73 (0.66)	0.88 (0.77)
	4	0.77 (0.69)	0.55	0.80 (0.71)	0.82 (0.77)	0.68 (0.59)	0.82 (0.66)
	5	0.78 (0.70)	0.59	0.84 (0.78)	0.76 (0.69)	0.69 (0.60)	0.87 (0.77)
	6	0.79 (0.71)	0.6	0.83 (0.77)	0.74 (0.66)	0.71 (0.63)	0.89 (0.80)
	7	0.78 (0.70)	0.58	0.84 (0.79)	0.74 (0.66)	0.77 (0.71)	0.82 (0.66)
	8	0.78 (0.70)	0.6	0.84 (0.79)	0.74 (0.65)	0.71 (0.63)	0.89 (0.78)
	10	0.82 (0.76)	0.61	0.74 (0.62)	0.72 (0.63)	0.69 (0.60)	0.93 (0.89)
Science and Technology/Engineering	5	0.76 (0.66)	0.51	0.76 (0.64)	0.76 (0.70)	0.72 (0.63)	0.84 (0.67)
	8	0.79 (0.70)	0.55	0.81 (0.73)	0.78 (0.72)	0.79 (0.73)	0.59 (0.32)
Biology	HS	0.79 (0.71)	0.59	0.81 (0.73)	0.73 (0.64)	0.80 (0.74)	0.85 (0.73)
Chemistry	HS	0.81 (0.73)	0.64	0.86 (0.82)	0.74 (0.64)	0.73 (0.63)	0.90 (0.82)
Introductory Physics	HS	0.79 (0.71)	0.59	0.79 (0.70)	0.76 (0.68)	0.79 (0.72)	0.87 (0.75)
Technology/Engineering	HS	0.79 (0.71)	0.55	0.78 (0.69)	0.75 (0.68)	0.83 (0.76)	0.71 (0.41)

**Failing* on all high school tests

.

Table 3-32. 2010 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*
English Language Arts	3	0.98 (0.97)	0.01	0.01	0.92 (0.88)	0.04	0.04	0.84 (0.81)	0.13	0.03
	4	0.96 (0.95)	0.02	0.02	0.90 (0.86)	0.06	0.04	0.94 (0.91)	0.05	0.02
	5	0.97 (0.96)	0.01	0.02	0.90 (0.87)	0.05	0.04	0.92 (0.89)	0.06	0.02
	6	0.97 (0.96)	0.01	0.02	0.91 (0.87)	0.05	0.04	0.92 (0.89)	0.06	0.02
	7	0.98 (0.97)	0.01	0.01	0.92 (0.89)	0.04	0.04	0.94 (0.92)	0.04	0.02
	8	0.98 (0.97)	0.01	0.01	0.94 (0.92)	0.03	0.03	0.88 (0.84)	0.08	0.04
	10	0.99 (0.99)	0	0.01	0.93 (0.90)	0.04	0.04	0.91 (0.87)	0.06	0.03
Mathematics	3	0.96 (0.95)	0.02	0.02	0.92 (0.88)	0.05	0.04	0.90 (0.86)	0.07	0.03
	4	0.96 (0.95)	0.02	0.02	0.90 (0.86)	0.06	0.04	0.91 (0.87)	0.07	0.03
	5	0.95 (0.94)	0.02	0.02	0.92 (0.89)	0.05	0.03	0.91 (0.87)	0.06	0.03
	6	0.95 (0.94)	0.02	0.02	0.92 (0.89)	0.05	0.03	0.91 (0.88)	0.06	0.03
	7	0.95 (0.93)	0.03	0.03	0.92 (0.89)	0.05	0.03	0.91 (0.89)	0.06	0.02
	8	0.94 (0.92)	0.03	0.03	0.92 (0.89)	0.05	0.03	0.92 (0.89)	0.06	0.02
	10	0.97 (0.96)	0.01	0.02	0.93 (0.91)	0.04	0.03	0.92 (0.88)	0.05	0.03
Science and Technology/ Engineering	5	0.95 (0.93)	0.02	0.03	0.88 (0.84)	0.07	0.05	0.92 (0.89)	0.06	0.02
	8	0.94 (0.91)	0.03	0.03	0.89 (0.85)	0.06	0.04	0.96 (0.94)	0.04	0
Biology	HS	0.95 (0.93)	0.02	0.03	0.92 (0.89)	0.05	0.04	0.92 (0.89)	0.05	0.03
Chemistry	HS	0.92 (0.89)	0.05	0.03	0.94 (0.91)	0.04	0.03	0.94 (0.92)	0.04	0.02
Introductory Physics	HS	0.95 (0.93)	0.03	0.03	0.91 (0.88)	0.05	0.04	0.93 (0.91)	0.04	0.02
Technology/ Engineering	HS	0.92 (0.89)	0.04	0.04	0.89 (0.85)	0.07	0.04	0.97 (0.96)	0.02	0

* *Failing* on all high school tests.

The above indices are derived from Livingston and Lewis's (1995) method of estimating decision accuracy and consistency. It should be noted that 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; that is, it is more intuitive and interpretable for two parallel forms to have the same statistical distribution.

Note that, as with other methods of evaluating reliability, DAC statistics 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 3-32 and 3-33 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 performance against Massachusetts content standards. Consistent with this purpose, results on the MCAS were reported in terms of performance levels, which describe student performance in relation to these established state standards. There are four performance levels: *Warning* (at grades 3–8) or *Failing* (at high school), *Needs Improvement*, *Proficient*, and *Advanced* (at grades 4–10) or *Above Proficient* (at grade 3). Students receive a separate performance 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 the districts for distribution to the schools. The details of the reports are presented in the following sections. See Appendix R for a sample *Parent/Guardian Report* and sample Student Labels.

3.8.1 Unique Reporting Notes

New in 2010, growth percentiles were reported on the *Parent/Guardian Reports*. Growth percentiles were reported only for ELA and mathematics. The median growth percentile for the school and district were also calculated and reported on the *Parent/Guardian Report*. Students' previous two scaled scores for ELA and mathematics were also reported where available. The *Parent/Guardian Report* was redesigned in 2010 to accommodate reporting growth percentiles. Focus groups were held in different Massachusetts towns with participants from various backgrounds. Also new in 2010, scaled scores were reported in grade 3.

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-identifying information, a commissioner's letter to parents, general information about the test, and website information for parent resources. The inside portion contains the performance level and scaled score for each content area tested. If the student does not receive a scaled score, the reason is displayed under the heading "Performance Level." Historical scaled scores are reported where appropriate and available. A performance 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.

If the student is administered the ELA or Mathematics test with certain nonstandard accommodations, a note specific to each nonstandard accommodation informing the parent is printed on the report. The following nonstandard accommodations are reported:

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

At the high school level, there is an additional note stating whether or not a student has met the graduation requirement for each subject, 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 performance level are reported. There are two printed copies of the reports: one for the parent and one for the school. Sample reports are provided in Appendix R.

The front page of the report provides the following identifying information about the student:

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

Student results labels are produced for each student receiving a *Parent/Guardian Report*. The information reported on the label includes the following:

- student name
- grade
- birth date
- test date
- student ID (SASID)
- school code
- school name
- district name
- student's scaled score and performance 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 reported results for the MCAS are accurate relative to collected data and other pertinent information, 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 S.

3.8.4 Quality Assurance

Quality assurance measures are implemented throughout the entire process of analysis and reporting at Measured Progress. The data processors and data analysts working on the MCAS tests perform quality control checks of their respective computer programs and intermediate products. Moreover, when data are handed off to different units within the Data Services and Static Reporting department (DSSR), 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 data for accuracy. Once report designs have been approved by the ESE, reports are run using demonstration data generated to reflect all possible scenarios for decision rules. These reports are then approved by the ESE.

Another type of quality assurance measure is 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 on staff who uses the decision rules to independently program the reporting table. The production and quality assurance tables are compared, and only after there is 100 percent agreement are the tables released for report generation.

The 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 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 and 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 those 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 as indicated by decision rules. This set is necessary in order to check 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 psychometric checks and program management review. The appropriate sample reports are then sent to the ESE for review and signoff.

3.9 MCAS Validity

Because interpretations of test scores, and not tests themselves, are evaluated for validity, the purpose of the *2010 MCAS and MCAS-Alt Technical Report* is to describe several technical aspects of the MCAS tests in support of score interpretations (AERA et al., 1999). Each section contributes an important component in the investigation of score validation: test development and design; test administration; scoring, scaling, and equating; item analyses; reliability; and score reporting.

Standards for Educational and Psychological Testing (AERA et al., 1999) provides a framework for describing sources of evidence that should be considered when constructing a validity argument. The evidence around test content, response processes, internal structure, relationship to other variables, and consequences of testing speaks to different *aspects* of validity but are not distinct *types* of validity. Instead, each contributes to a body of evidence about the comprehensive validity of score interpretations.

Evidence on test content validity is meant to determine 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 this lens provided by the standards, evidence based on test content is extensively described in Sections 3.2 and 3.3. 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 are all components of validity evidence based on test content. 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. Items are presented to students in multiple formats (constructed-response and multiple-choice). 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 locally administered annual training sessions. To speak to student response processes, however, additional studies would be helpful and might include an investigation of students' cognitive methods using think-aloud protocols.

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), differential item functioning analyses, dimensionality analyses, reliability, standard errors of measurement, 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 in 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.

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*.

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. Performance 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.

The evidence presented in this chapter supports inferences made about student achievement on the content represented on the Massachusetts content standards for English language arts, mathematics, and science and technology/engineering. 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 the empirical and logical evidence supporting the technical quality of the MCAS Alternate Assessment (MCAS-Alt). The information presented documents the processes and procedures used to develop, 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 there is intentional flexibility built into the MCAS-Alt to maximize the instructional usefulness of the results, the procedures described in this report are designed to constrain unwanted variability where possible.

This chapter includes a separate section for each phase of the alternate assessment process. The sections provide a basis for the validity of the alternate assessment. That is, while each section taken individually is a key component of any technical report, together they document how the assessment appropriately evaluates the knowledge and skills of students with disabilities in the context of grade-level content standards.

This chapter is intended primarily for a technical audience, such as the ESE, its technical advisory committee, special education directors, and researchers. However, teachers and parents are also crucial parts of the alternate assessment system. Thus, this chapter is intended to be read more broadly than is the case for general education technical documents, although certain sections will require highly specialized knowledge and a solid understanding of measurement concepts. The chapter is organized using a construct validity framework; that is, all of the information presented is intended to support the inferences about students and/or schools from the assessment scores.

4.1.2 Purposes of the Assessment System

The MCAS is the state's testing program for students, implemented in response to the 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 learning standards in the Massachusetts curriculum frameworks. The curriculum for all students, including students with disabilities, must be aligned with these standards. The MCAS is designed to improve teaching and learning; 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 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;
- to determine how much of the academic curriculum a student has learned;
- to use the assessment results to provide challenging academic instruction;
- to include difficult-to-assess students in statewide assessment and accountability systems;
- to 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, with stakeholder involvement, between 1998 and 2000. It has been refined and enhanced each year since its 2001 implementation.

4.1.3 Format

The MCAS-Alt consists of a portfolio of "evidence" collected during the school year that documents the student's performance of the skills, knowledge, and concepts outlined in the state's curriculum frameworks. Alternate assessments allow the ESE to report results to parents, schools, and the public on the academic performance of *all* students with disabilities, and to assist schools in developing challenging programs of instruction for students with significant disabilities.

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

4.2 Test Design and Development

4.2.1 Test Content

MCAS-Alt assessments are required in all grades and content areas for which standard MCAS tests are administered. Specific MCAS-Alt requirements for students in each grade are listed below.

Grade	*English Language Arts Strands Required*	*Mathematics Strands Required*	*Science and Technology/Engineering 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 Science and Technology/ Engineering strands*
6	▪ Language (General Standard 4) ▪ Reading and Literature (General Standard 8)	▪ Number Sense and Operations ▪ Patterns, Relations, and Algebra	
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 Science and Technology/ Engineering 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

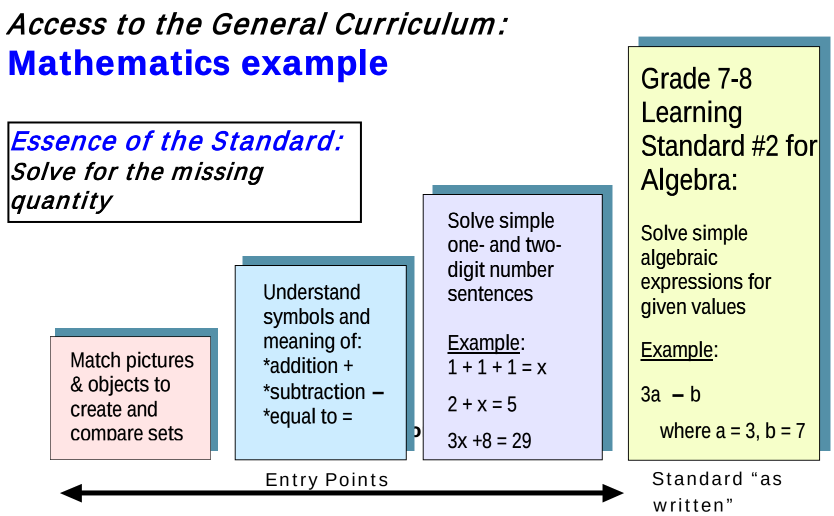
*Earth and Space Science, Life Science, Physical Science, Technology/Engineering

4.2.1.1 Access to the Grade-Level Curriculum

The ESE's *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities* can be 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. Targeted outcomes are based on *entry points* at the grade level in which the student is enrolled.

Most students with significant disabilities will be able to 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, even at the lowest levels of complexity, and may instead need to focus on goals that allow them to explore the tools, materials, and academic content by addressing targeted social, communication, and/or motor skills (*access skills*) practiced during academic activities. For example, a student may practice operating an electronic switch on cue to indicate whose turn is next during an STE activity; or reach, grasp, and release the materials used during a mathematics activity; or focus on a story read aloud for increasing periods of time.

Figure 4-1. 2010 MCAS-Alt: Access to the General Curriculum (Mathematics Example)



4.2.1.2 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 required pieces of primary evidence must be a data chart (e.g., field data chart, line graph, bar graph) that shows the following information, at minimum:

- the targeted skill based on the learning standard being assessed
- tasks performed by the student on eight distinct dates, with a description of each
- percentage of accuracy for each performance
- percentage of independence for each performance
- progress over time, beginning at a level that indicates 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. The data chart plus at least two additional pieces of primary evidence form the minimum "core set of evidence" required in each portfolio strand.

Each piece of primary evidence must be labeled with the

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

Supporting Documentation
In addition to the required primary evidence, supporting documentation (further elaborated in Section 4.2.1.4) may be included at the discretion of the teacher to show the context in which the activity was conducted. This documentation may include

- notes from teachers or peers describing the activity;
- photographs showing the context of the learning activity;
- self-evaluation or reflection sheets;
- work description labels.

4.2.1.3 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 strand of the portfolio and apply criteria in order to produce a score in each rubric area based on the evidence found in the portfolio. Scores are based on evidence of the following:

- **completeness** of all portfolio materials
- **level of complexity** at which the student addresses learning standards in the Massachusetts curriculum frameworks in the content area being assessed
- **accuracy** of the student's responses to questions, or of his or her 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-correcting, goal-setting)
- **generalized performance** of a skill in different instructional contexts, settings, or using different materials or methods of response

4.2.1.4 Supporting Documentation

Supporting documentation should be included, where needed, to adequately describe the student's performance, and must be clearly labeled with the student's name and date of completion. Supporting documentation provides additional descriptive information on (1) the setting and context in which the learning activity occurred; (2) the student's self-evaluation of his or her performance; or (3) other information describing the student's performance by the teacher, parent, other adult, or peer.

Supporting documentation may include any of the following submissions:

- **narrative descriptions** by the teacher or parent describing how the task or activity was conducted and/or what the student was asked to do
- **photographs** of the student engaged in specific tasks or relevant classroom or community activities that show how the student engaged in the instructional activity (i.e., the context of the activity)
- **tools, templates, or examples** made by the student
- **reflection sheet or 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

4.2.1.5 MCAS-Alt Grade-Level and Competency Portfolios

A small number of MCAS-Alt portfolios were submitted for students who address learning standards at or near grade-level expectations but were unable to participate in standard MCAS testing, even with accommodations. See the Participation Guidelines section of the *2010 Educator's Manual for MCAS-Alt* for a description and profile of the students who should be considered for the MCAS-Alt, and in particular for grade-level and competency portfolios.

MCAS-Alt English Language Arts (ELA), Mathematics, and Science and Technology/Engineering (STE) portfolios may be submitted to earn a Competency Determination (CD) for students in high school. Each competency portfolio is evaluated by a panel of content area experts to determine whether it provides evidence that the student's performance is equivalent to that of a student who received a performance level of *Needs Improvement* or higher on the standard MCAS test in that content area.

A small number of MCAS-Alt portfolios are submitted for students in grades 3 through 8 which address learning standards at or near grade-level expectations but are unable to participate in standard MCAS testing, even with accommodations. These portfolios are reviewed by a team of content area experts to determine if they meet the criteria to attain a performance level of *Needs Improvement* or higher.

For additional information on how these 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

Students with disabilities are required by federal and state laws to participate in the MCAS in order to assess their performance of the skills and knowledge of content described in the state's curriculum frameworks. Students with disabilities must either take MCAS tests, with or without accommodations, or participate in an alternate assessment if they cannot take the tests due to the severity of their disabilities.

Alternate assessments measure the academic performance of students with the most significant disabilities. Before 1998, learning was not assessed or reported for all students with disabilities. Since being required to participate in the assessment system, schools have been accountable for their performance, and students with disabilities have a greater chance of being considered when decisions are made to allocate school resources.

The alternate assessment results provide accurate and detailed feedback that can be used to identify challenging instructional goals for each student. The evidence submitted in a portfolio ensures that students with the most significant disabilities have an opportunity to "show what they know" and to receive instruction at a level that is challenging and attainable.

Using the curriculum resources provided by the ESE to improve and enhance instruction for students with disabilities, teachers have become adept at providing standards-based instruction at a level that challenges and engages each student, and they cite unanticipated gains in students' performance and understanding.

4.2.2.2 Role of Advisory Committee

The MCAS-Alt Advisory Committee meets twice annually (December and March) to discuss policy issues related to the alternate assessment and students with significant cognitive disabilities. The MCAS-Alt Advisory Committee has been a critical component to the development, implementation, and continued administration of the MCAS-Alt. This diverse group of stakeholders, including teachers, parents, student advocates, representatives of students who are blind and deaf, principals, private school directors, special education directors, IEP team chairpersons, and representatives from higher education has contributed to the overall vision and decision making regarding relevant issues that affect the assessment of students with disabilities. 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 a 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 whether the skill has been mastered. On each date, the data must indicate whether a correct response was given (percent of accuracy) and whether the student required a cue or prompt (percent of independence). Data can be 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, clearly illustrating how the task relates to the measurable outcome being assessed. Data charts may include performance data from any of the following:

A Collection of Work Samples
A percentage of accuracy and independence can be either charted for each work sample or summarized for a number of work samples on each date, provided all work is based on the same skill or measurable outcome.

Responses to Specific Tasks
The percentage of accuracy and independence must be charted each time an activity, task, or trial is conducted, provided these are related to the same skill or outcome. Multiple data recorded on a single date must be summarized for percent of accuracy and independence for each date.

4.3.2 Construction of Portfolios

The student's MCAS-Alt portfolio must include the elements listed below. All forms may be photocopied from originals found in the Required Forms section and/or the Product Description Labels and Blank Data Charts section of the *2010 Educator's Manual for MCAS-Alt*, or may be completed electronically by using the online MCAS-Alt Forms and Graphs program available at www.doe.mass.edu/mcas/alt.

- **artistic cover** (recommended but not required) designed and produced by the student and inserted in the front window of the three-ring portfolio binder
- **portfolio cover sheet** containing important information about the student
- **portfolio contents checklist** (optional) verifying the materials included in the portfolio
- **student's introduction to the portfolio** produced as independently as possible by the student using his or her primary mode of communication. The introduction may be written, dictated, or recorded on video or audio and should describe "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 child's portfolio or, at minimum, was invited to do so. In the event no signature was obtained, the school must include a record of its attempts to invite a parent, guardian, or primary care provider to view the portfolio.
- **consent form to photograph and audio/videotape a student** (required if images or recordings of the student are included in the portfolio). This form, provided in English and Spanish, must be signed by the parent or guardian before images or recordings of the student can be made. A signed copy of this form must be kept on file at the school.
- **weekly schedule** documenting the student's enrollment in a program of instruction, including participation in the general academic curriculum
- **strand cover sheet** related to the set of evidence that addresses a particular outcome in the required standard/strand
- **product description** (optional) attached to each piece of primary evidence that provides required information. If labels are not used, required information must be provided on teacher-designed labels or written directly on each piece.

The above required elements, along with all evidence and other documentation, are placed inside a white three-ring portfolio binder provided by the ESE for each student.

4.3.3 Participation Requirements

4.3.3.1 Identification of Students

All students whose education is publicly funded, including students with disabilities, must participate in MCAS. Students with significant disabilities who are unable to take the standard MCAS tests, even with accommodations, must take the MCAS-Alt.

A student with a disability has either an Individualized Education Program (IEP) provided under the Individuals with Disabilities Education Act or a plan provided under Section 504 of the Rehabilitation Act of 1973. All students with disabilities 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 to the ESE's Student Information Management System (SIMS).

4.3.3.2 Participation Guidelines

A student's IEP 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.

The student's team considers 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 in order 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 on the following page.

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 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 test 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.)	***Then*** 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, and even when 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 that content area.

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).

Students who address knowledge and skills at grade-level expectations, but who require an alternate assessment, can satisfy the Competency Determination requirement if they can demonstrate in their portfolio a level of achievement comparable to that of a student who has met the CD requirements on the standard grade 10 test or retest in that subject. Students who meet these requirements on the high school MCAS-Alt will be eligible to earn a CD. For further information on meeting the CD through the MCAS-Alt, see pages 25–32 of the *2010 Educator's Manual for MCAS-Alt*.

4.3.3.3 MCAS-Alt Participation Rates

Across all content areas, a total of 9,429 students, or 1.7 percent of the assessed population, participated in the 2010 MCAS-Alt in grades 3–10. A slightly higher relative proportion of students in grades 3–8 took MCAS-Alt compared with students in grade 10, and slightly more students were 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 Administrator Training

During the month of October a total of 882 administrators received training on the 2010 MCAS-Alt and other topics related to MCAS for students with disabilities, including assessment results, reporting, and accountability.

Topics were as follows:

- 2009 MCAS results
- graduation requirements, MCAS retests, Performance Appeals, and Education Proficiency Plans
- state accountability and AYP
- MCAS accommodations
- MCAS-Alt
- resources and changes for 2010

4.3.5 Educator Training

During the month of October 2009, a total of 2,755 educators received training on the 2010 MCAS-Alt. Educators attending the training had the option of attending one of two sessions. Session one was for educators who were new to the process. Topics of discussion for session one were as follows:

- Which students should take the MCAS-Alt?
- portfolio requirements for each grade and content area
- collecting data on measurable outcomes to assess a student's performance and progress
- using the *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities* (fall 2006)

Session two was for educators who had compiled and submitted alternate assessments previously. Topics of discussion for session two were as follows:

- Which students should take the MCAS-Alt?

- How did students perform on the 2009 MCAS-Alt?
- 2010 MCAS-Alt requirements/what’s new
- using the *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities* (fall 2006)
- frequent mistakes in compiling MCAS-Alt portfolios
- data collection requirements for MCAS-Alt

Additionally, during January 2010, a total of 1,096 educators received training, either for educators who were new to the process (and did not attend the overview training in the fall) or at roundtable discussions where teachers were able to review and discuss their students’ portfolios in an informal setting with their questions answered by expert teachers.

During March 2010, an additional 786 educators received training at roundtable discussions, where they were able to review and discuss their students’ portfolios and have their questions answered by expert teachers and attend an afternoon session on the scoring of portfolios.

4.3.6 Support for Teachers: Service Center

The MCAS Service Center provides accurate toll-free telephone support to district and school staff who have questions related to test administration, reporting, training, materials, and other relevant operations and logistics.

The Measured Progress project management team provided extensive training to the 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 2010. The ESE and Measured Progress closely monitored scorers to ensure that the score of each portfolio was accurate.

Through application of a universal scoring rubric and verification of the standards being assessed as found in the *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities* (fall 2006), evidence of the student’s performance was evaluated and scored against research-based criteria on how students with significant disabilities learn and demonstrate knowledge and skills. The MCAS-Alt Rubric for Scoring Portfolio Strands was developed with assistance from teachers and a statewide advisory committee. The criteria for scoring portfolios are listed and described in detail on the following pages.

The scoring of MCAS-Alt portfolios reflects the degree to which a student has learned, understood, and applied the knowledge and skills outlined in the Massachusetts curriculum frameworks. The MCAS-Alt portfolio measures progress over time, as well as the highest achievement attained by the student on the assessed skills, and reflects the degree to which cues, prompts, and other assistance was provided to the student.

4.4.1 Scoring Logistics

MCAS-Alt portfolios were reviewed and hand-scored by trained scorers according to the procedures described in this section. Scores were entered onto score forms designed by Measured Progress and the ESE; score forms were monitored for accuracy and completeness.

Security was maintained at the scoring site, with access to unscored portfolios and completed score forms restricted to ESE and Measured Progress staff. MCAS-Alt scoring leadership staff at each site included a floor manager (FM) and table leaders (TLs). Each table leader managed a table with four to five scorers. The FM managed a group of tables within a range of grade 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 Measured Progress staff, plus four teacher consultants. The PLT met for two days in early April to accomplish the following:

- select sample portfolio strands to use for training, calibration, and qualification of scorers
- field-test the *2010 Guidelines for Scoring Student Portfolios*

On the first day, the group reviewed and scored approximately 200 portfolios using the draft of the 2010 guidelines, noting any scoring problems that arose during the review. All concerns were resolved either using the educator's manual or through additional scoring rules agreed upon by the PLT and subsequently addressed in the final 2010 guidelines.

Of the 200 portfolios reviewed, 60 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 60 double-scores, 48 were in exact agreement in all five scoring dimensions: Level of Complexity, Demonstration of Skills and Concepts, Independence, Self-Evaluation, and Generalized Performance.

These 48 samples were set aside and rescored. Of these 48 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, Measured Progress recruited 120 people to serve as scorers or TLs at the MCAS-Alt Scoring Center. All TLs and many scorers had worked on scoring projects for other test administrations and had a four-year college degree.

Training

Scorers were thoroughly trained in all rubric areas and score points through review and "mock scoring" of a sample set of student portfolios selected to illustrate clear examples of each rubric score point. Scorers were given detailed instructions regarding how to review each piece of evidence and tally the data using a strand organizer. Scorers were then taught to apply the resulting data to the rubrics (see Section 4.4.3) for Level of Complexity, Demonstration of Skills and Concepts, Independence, Self-Evaluation, and Generalized Performance. After some basic instructions regarding the assignment of rubric scores, trainers reviewed actual 2010 portfolio samples with scorers, discussing each piece of evidence and the score it should receive in each dimension. Trainers facilitated discussion and review among scorers to clarify the characteristics of each score point.

Scorer Qualification

Before scoring actual student portfolios, each scorer was required to demonstrate the ability to score by taking a qualifying assessment of 24 questions and scoring a sample portfolio of four strands (20 dimensions). The qualifying threshold score required on the assessment was 85 (21 correct of 24 total questions). The qualifying rate of accuracy on the sample portfolio was 85 percent exact agreement overall for the five scoring dimensions (Level of Complexity, Demonstration of Skills and Concepts, Independence, Self-Evaluation, and Generalized Performance); that is, exact agreement on 17 of 20 total scorable dimensions for the four strands.

Scorers who did not achieve the required accuracy rate on the qualifying assessment were retrained on another qualifying assessment. If they achieved an accuracy rate of at least 85 percent exact agreement, they 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 Table Leaders and Floor Managers

Table leaders were recruited, trained, and qualified prior to scoring by the ESE using the same methods and criteria used for scorers, except that they were required to score 90 percent or better on the qualifying test. TLs and FMs also received training in logistical, management, and security procedures.

Eight licensed Massachusetts educators who had led a table during the previous year's scoring institute were designated as M-resolvers. M-resolvers were there to assist in the training of the new TLs and to perform resolution scores on portfolios with scores of M.

The room was monitored by two floor managers, 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

A scorer was randomly assigned a portfolio by his or her TL. The first step in scoring is to ensure that the portfolio is complete. The scorer ensured that the required strands for each grade were submitted. Then each submitted strand was scored individually. A strand was considered complete if it included a data chart with at least eight different dates and two additional pieces of evidence related to the same measurable outcome.

To assist in the scoring, the scorer used a worksheet called the strand organizer. By filling in the worksheet, the strand organizer allowed a scorer to keep track of the evidence submitted for each strand, determine its completeness, and make the necessary calculations to come up with a final score for each rubric dimension.

Scorers sat at a table with four or five other scorers, all scoring the same grade-level portfolios and guided by a TL. TLs were experienced scorers who qualified at a higher threshold and who had received advanced training on the logistics of the scoring center. Scorers could ask TLs questions as they reviewed their portfolios. In the event that the TL could not answer a question, the FM assisted with the question. In the event that the FM was unable to answer a question, ESE staff was available to provide clarification.

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

MCAS-Alt 2010 score distributions for all scoring dimensions are provided in Appendix F.

***A.* Level of Complexity**

The score for Level of Complexity reflects how the student addressed curriculum framework learning standards. Using the *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities* (fall 2006) the scorer ensured that the student's measurable outcome was aligned with the intended learning standard and, if so, whether the work was at grade level, an entry point, or an the access skill.

Level of Complexity Score for Each Strand

Each strand was given a Level of Complexity score based on the scoring rubric for Level of Complexity (Table 4-1). Scorers assigned a Level of Complexity score based on the following criteria:

- whether the evidence was aligned with a learning standard in the required strand
- whether the evidence met grade-level performance expectations, was modified below grade-level expectations ("entry points"), or addressed skills of daily living in the context of academic instruction ("access skills")

Table 4-1. 2010 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 all portfolio evidence was correctly labeled with the following information:

- student’s name
- date of performance
- percentage of accuracy
- percentage of independence

If any piece of evidence was not labeled correctly, that piece was not scorable. If at least two pieces of correctly labeled primary evidence and a complete data chart, all documenting the student’s performance of the same skill, were not submitted, the strand received scores of M in <u>both</u> Demonstration of Skills and Concepts <u>and</u> Independence (see section C).

Each strand was scored for Demonstration of Skills and Concepts by first identifying the “final 1/3 time frame” on the data chart (or the final three points, if fewer than 12 points are listed on the chart).

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

- all data points in the final 1/3 time frame of the data chart;
- 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 4-2 below.

A score of M was also given if the data chart listed the percentages of accuracy and independence as 80–100 percent for the duration of the data collection period.

Table 4-2. 2010 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 performed the skill without cues or prompts during tasks or activities based on the learning standards being assessed.

Each strand was scored for Independence by first identifying the final 1/3 time frame on the data chart (or the final three points, if fewer than 12 points are listed on the chart).

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

- all data points during the final 1/3 time frame of the data chart;
- all other primary evidence in the portfolio 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 4-3 below.

A score of M was given in both Demonstration of Skills and Concepts and in Independence when the following primary evidence was *not* included in the strand:

- one data chart measuring a single skill or outcome based on the required learning standard or strand on at least eight different dates that shows the student's accuracy and independence on each task or trial; and
- two pieces of primary evidence, such as work samples, videos, or photographs, that measure the same skill (or address the same outcome) as the data chart, labeled with all required information.

A score of M was also given if the primary evidence listed above lists the percentages of accuracy and independence as 80–100 percent for the duration of the data collection period.

Table 4-3. 2010 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 shows the frequency of self-correction, self-monitoring, goal-setting, reflection, and overall awareness by the student of his or her own learning. The 2010 MCAS-Alt score distributions for Self-Evaluation are provided in Appendix F.

Self-Evaluation Score in Each Strand

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

Table 4-4. 2010 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 self-corrects, monitors, sets goals, and reflects in this content area—**multiple examples** of self-evaluation were found in this strand.

Combined Self-Evaluation Score

A final score for Self-Evaluation was given in the content area by combining the three individual strand scores according to Table 4-5 or, in the case of a two-strand portfolio, by combining the two individual strand scores according to Table 4-6. Descriptors of the overall content area scores are shown in Table 4-7.

Table 4-5. 2010 MCAS-Alt: Determination of Combined Self-Evaluation Score for Each Content Area: 3-Strand Portfolio

Strand Score 1	*Strand Score 2*	*Strand Score 3*	*Combined Content Area Score*
M	M	M	**M**
M	M	1	**1**
M	M	2+	**1**
M	1	1	**2**
M	1	2+	**2**
M	2+	2+	**2**
1	1	1	**3**
1	1	2+	**3**
1	2+	2+	**3**
2+	2+	2+	**4**

Table 4-6. 2010 MCAS-Alt: Determination of Combined Self-Evaluation Score for Each Content Area: 2-Strand Portfolio

Strand Score 1	*Strand Score 2*	*Combined Content Area Score*
M	M	**M**
M	1	**1**
M	2+	**1**
1	1	**2**
1	2+	**3**
2+	2+	**4**

Table 4-7. 2010 MCAS-Alt: Rubric for Combined Self-Evaluation Score in Each Content Area

Score Point				
M	*1*	*2*	*3*	*4*
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—evidence of self-evaluation was found in only one strand.	Student occasionally self-corrects, monitors, sets goals, and reflects in this content area—evidence of self-evaluation was found in two strands.	Student frequently self-corrects, monitors, sets goals, and reflects in this content area—for a three-strand portfolio, at least one example of self-evaluation was found in each strand; for a two-strand portfolio, two or more examples were found in only one strand.	Student self-corrects, monitors, sets goals, and reflects all or most of the time in this content area—two or more examples of self-evaluation were found in 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.

Generalized Performance Score in Each Strand

Scorers totaled the number of contexts and approaches in each strand to determine the score of either 1 or 2+, based on the rubric shown in Table 4-8.

Table 4-8. 2010 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**.

Combined Generalized Performance Score

A final Generalized Performance score was determined in the content area by combining the three scores for each strand, as shown in Table 4-9, or in the case of a two-strand portfolio, by combining the two individual strand scores, as shown in Table 4-10. Descriptors for the combined Generalized Performance scores are shown in Table 4-11.

Table 4-9. 2010 MCAS-Alt: Determination of Combined Generalized Performance Score for Each Content Area: 3-Strand Portfolio

Strand Score 1	*Strand Score 2*	*Strand Score 3*	*Resulting Overall Score*
1	1	1	**1**
1	1	2+	**2**
2+	2+	1	**3**
2+	2+	2+	**3**

Table 4-10. 2010 MCAS-Alt: Determination of Combined Generalized Performance Score for Each Content Area: 2-Strand Portfolio

Strand Score 1	*Strand Score 2*	*Resulting Overall Score*
1	1	**1**
1	2	**2**
2	2	**3**

Table 4-11. 2010 MCAS-Alt: Rubric for Combined Generalized Performance Score in Each Content Area

Score Point		
1	*2*	*3*
Student demonstrates knowledge and skills in one context; or uses one approach and/or method of response and participation in each strand in the content area.	Student demonstrates knowledge and skills in multiple contexts; or uses multiple approaches and/or methods of response and participation in only one strand in the content area.	Student demonstrates knowledge and skills in multiple contexts; or uses multiple approaches and/or methods of response and participation in two or more strands in the content area.

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

Grade-Level Portfolios in Grades 3 through 8

Each 2010 grade-level portfolio was evaluated by a panel of content area experts to determine whether it met *Needs Improvement* (or higher) performance level requirements. To receive a score of *Needs Improvement* or higher, the portfolio must have demonstrated the following:

- that the student showed knowledge and skills at the level comparable with a student who received scores of *Needs Improvement* or higher on the standard MCAS test
- that the student had independently and accurately addressed all required learning standards and strands described in the portfolio requirements

High School Competency Determination Portfolios

A student may earn a score of *Needs Improvement* or higher (in order to fulfill the state's Competency Determination requirement) by submitting an MCAS-Alt competency portfolio in ELA, Mathematics, and/or STE. Specific requirements for submission of competency portfolios are described in the *2010 Educator's Manual for MCAS-Alt*.

Each 2010 competency portfolio was evaluated by a panel of content area experts to determine whether it met *Needs Improvement* (or higher) performance level requirements. To receive a score of *Needs Improvement* or higher, the portfolio must have demonstrated the following:

- that the student showed knowledge and skills at the level comparable with a student who received scores of *Needs Improvement* or higher on the standard MCAS test
- that the student had independently and accurately addressed all required learning standards and strands described in the portfolio requirements

If the student's portfolio demonstrated a level of performance comparable to or higher than that of students who passed the standard grade 10 MCAS tests in ELA and Mathematics, the student was awarded a CD in that content area.

4.4.5 Monitoring the Scoring Quality

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

Scoring consistency and accuracy were maintained using the following methods, described below:

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

Double-Scoring

All portfolios for students in grades 9–12 were double-scored. Scorers at grades 3–8 had at least one of their portfolios double-scored each morning and afternoon, and every fifth portfolio thereafter. At least 20 percent of portfolios for students in grades 3–8 were double-scored.

Double-scoring refers to a portfolio being scored by two scorers at different tables, without knowledge by either scorer of the score assigned to the portfolio by the other.

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

Table 4-15 in Section 4.6.3 shows the percentages of inter-rater agreement for the 2010 MCAS-Alt.

Read-Behind Scoring

Read-behind scoring refers to a table leader's 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 table leader's score became the score of record.

Read-behinds (or double-scores) were performed on every scorer's first three portfolios. If those scores were consistent with the table leader's resolution scores, a read-behind (or a double-score) 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 table leader's resolution scores, the scorer was retrained. The table leader determined when a retrained scorer could resume scoring. Additionally, a read-behind (or a double-score) was performed on each subsequent portfolio for any scorer permitted to resume scoring, until consistency with the table leader's scores was established.

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

Scorer Tracking Forms

The table leader 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 providing the Project Leadership Team with a record of scorers' accuracy and consistency, scoring 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 cumulative accuracy and productivity for each scorer.

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 include an evaluation of each item. Both *Standards for Educational and Psychological Testing* (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. While the specific statistical criteria identified in these publications were developed primarily for general—not alternate—assessment, the principles and some of the techniques apply within the alternate assessment framework as well.

Both qualitative and quantitative analyses were conducted to ensure that MCAS-Alt items met these standards. Qualitative analyses are described in earlier sections of this report; this section focuses on the quantitative evaluations. The statistical evaluations discussed are 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 2010 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 performance level categorization. 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 were evaluated in terms of item difficulty according to standard classical test theory practices. "Difficulty" was defined as the average proportion of points achieved on an item and was measured by obtaining the average score on an item and dividing by the maximum score for the item. MCAS-Alt tasks are scored polytomously, such that a student can achieve a score of M, 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 on student progress relative to a standard rather than to differentiate among 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 are only cautiously applicable to the MCAS-Alt.

A desirable feature of an item is that the higher-ability students perform better on the item than do 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/content area combination is presented in Table 4-14. 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 4-14 should not be interpreted to mean that the students performed better on the MCAS-Alt than the students who took general assessments did on those tests.

Also shown in Table 4-14 are the mean discrimination values. As with the item difficulty values, because the nature and use of the MCAS-Alt are different from those 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 4-14 should be interpreted with caution.

Table 4-14. 2010 MCAS-Alt: 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*
English Language Arts	3	ALL	9	0.85	0.20	0.62	0.06
		OR	9	0.85	0.20	0.62	0.06
	4	ALL	9	0.85	0.19	0.53	0.07
		OR	9	0.85	0.19	0.53	0.07
	5	ALL	9	0.85	0.19	0.65	0.10
		OR	9	0.85	0.19	0.65	0.10
	6	ALL	9	0.85	0.19	0.63	0.06
		OR	9	0.85	0.19	0.63	0.06
	7	ALL	9	0.85	0.19	0.53	0.07
		OR	9	0.85	0.19	0.53	0.07
	8	ALL	6	0.85	0.20	0.67	0.13
		OR	6	0.85	0.20	0.67	0.13
	HS	ALL	9	0.83	0.18	0.47	0.16
		OR	9	0.83	0.18	0.47	0.16
Mathematics	3	ALL	12	0.85	0.20	0.63	0.04
		OR	12	0.85	0.20	0.63	0.04
	4	ALL	12	0.85	0.20	0.67	0.06
		OR	12	0.85	0.20	0.67	0.06
	5	ALL	9	0.85	0.19	0.65	0.07
		OR	9	0.85	0.19	0.65	0.07
	6	ALL	6	0.85	0.19	0.65	0.06
		OR	6	0.85	0.19	0.65	0.06
	7	ALL	12	0.85	0.19	0.66	0.09
		OR	12	0.85	0.19	0.66	0.09
	8	ALL	12	0.85	0.19	0.67	0.09
		OR	12	0.85	0.19	0.67	0.09
	HS	ALL	15	0.84	0.17	0.35	0.14
		OR	15	0.84	0.17	0.35	0.14
STE	5	ALL	12	0.85	0.19	0.46	0.08
		OR	12	0.85	0.19	0.46	0.08
	8	ALL	12	0.85	0.19	0.56	0.10
		OR	12	0.85	0.19	0.56	0.10
Biology	HS	ALL	12	0.83	0.18	0.43	0.22
		OR	12	0.83	0.18	0.43	0.22
Chemistry	HS	ALL	12	0.85	0.18	0.42	0.11
		OR	12	0.85	0.18	0.42	0.11
Introductory Physics	HS	ALL	12	0.82	0.17	0.53	0.31
		OR	12	0.82	0.17	0.53	0.31
Technology/ Engineering	HS	ALL	12	0.82	0.18	0.56	0.09
		OR	12	0.82	0.18	0.56	0.09

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 are provided in Appendix E; item difficulty and discrimination values are presented for each item. 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 performance 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) 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) 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 are shown in Table 4-15.

Table 4-15. 2010 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*
English Language Arts	3	2	0.21	0.11	0.30	0.03	0.05	0.03
	4	3	0.25	0.19	0.36	0.00	0.03	0.02
	5	2	0.18	0.19	0.33	0.03	0.04	0.01
	6	2	0.20	0.15	0.24	0.03	0.03	0.01
	7	3	0.23	0.24	0.29	0.02	0.01	0.03
	8	2	0.20	0.28	0.38	0.03	0.03	0.03
	HS	3	0.14	0.06	0.38	0.06	0.07	0.01
Mathematics	3	2	0.19	0.17	0.27	0.06	0.03	0.04
	4	2	0.21	0.26	0.36	0.04	0.06	0.05
	5	2	0.17	0.23	0.35	0.02	0.03	0.06
	6	2	0.21	0.17	0.29	0.03	0.05	0.01
	7	2	0.26	0.21	0.29	0.07	0.04	0.07
	8	2	0.15	0.23	0.39	0.01	0.04	0.00
	HS	5	0.06	-0.30	0.32	0.05	0.17	0.07
STE	5	4	0.23	0.18	0.29	0.08	0.05	0.04
	8	4	0.20	0.30	0.43	0.08	0.04	0.09
Biology	HS	4	-0.20	-0.35	0.51	0.53	0.39	0.29
Chemistry	HS	4	0.05	-0.11	0.08	0.06	0.05	0.18
Introductory Physics	HS	4	-0.20	-0.42	0.67	0.07	0.12	0.11
Technology/ Engineering	HS	4	0.20	0.20	0.45	0.10	0.06	0.13

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

The average correlations among the dimensions range from moderately strong and negative to moderately strong and positive. Note that a negative relationship in some cases may be expected. For example, a low or negative correlation between Level of Complexity and Demonstration of Skills and Concepts may not be surprising, whereas a 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 4-15 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 actions should be taken to make certain that differences in performance are due to 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) is used 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 to plan 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 ADC 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 teachers 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 are taught in order to produce the required evidence. It is required, and therefore virtually guaranteed, that 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 recalibration to verify and validate portfolio scores. These procedures, along with ESE review of each year’s MCAS-Alt results, confirm 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 performance level rather than a total raw score—it was not possible to run decision accuracy and consistency (DAC) analyses.

4.6.1 Reliability

In the previous section, individual item characteristics of the 2010 MCAS-Alt were presented. Although individual item performance can be 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. All measurements include some degree of measurement error; no academic assessment can measure student performance with perfect accuracy. Some students will receive scores that underestimate their true ability, while others receive scores that overestimate their true ability. Items that function well together produce assessments that have less measurement error (i.e., the error is small on average). Such assessments are described as "reliable."

There are a number of ways to estimate an assessment's reliability. One approach is to split all test items into two groups and then correlate students' scores on the two half-tests. This is known as a split-half estimate of reliability. If the two half-test scores correlate highly, the items on them are likely measuring very similar knowledge or skills. It 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 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), that avoids the shortcomings 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 2010 MCAS-Alt tests. The formula is as follows:

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

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

Table 4-16 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 4-16. 2010 MCAS-Alt: Raw Score Descriptive Statistics, Cronbach's Alpha, and Standard Errors of Measurement (SEM) by Content Area and Grade

Content Area	*Grade*	*Number of students*	*Raw score*			*Alpha*	*SEM*
			Maximum	*Mean*	*Standard deviation*		
English Language Arts	3	1,197	26	21.56	1.11	0.68	0.63
	4	1,362	39	32.08	2.20	0.80	0.97
	5	1,214	26	21.59	1.16	0.72	0.61
	6	1,190	26	21.61	1.14	0.70	0.63
	7	1,155	39	31.99	2.35	0.81	1.02
	8	1,120	26	21.48	1.34	0.75	0.66
	10	827	39	31.22	3.10	0.79	1.43
Mathematics	3	1,232	26	21.56	1.20	0.70	0.66
	4	1,342	26	21.54	1.25	0.74	0.63
	5	1,276	26	21.60	1.22	0.72	0.64
	6	1,260	26	21.62	1.20	0.72	0.64
	7	1,208	26	21.53	1.30	0.73	0.67
	8	1,220	26	21.56	1.38	0.75	0.69
	10	850	39	30.93	3.41	0.83	1.39
Science and Technology/Engineering	5	1,090	39	32.04	2.33	0.84	0.92
	8	1,053	39	31.88	2.62	0.90	0.83
Biology	HS	719	39	31.05	3.33	0.92	0.95
Chemistry	HS	68	39	31.76	2.84	0.65	1.67
Introductory Physics	HS	54	39	30.69	3.68	0.75	1.84
Technology/ Engineering	HS	73	39	30.81	3.31		

Note: no reliability or SEM values are reported for HS Technology/Engineering because there were some items for which only one student received a score.

An alpha coefficient toward the high end is taken to mean that the items are likely measuring very similar knowledge or skills; that is, that 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 took the 2010 MCAS-Alt. Subgroup Cronbach's α's were calculated using the formula defined above including only the members of the subgroup in question in the computations. These statistics are reported in Appendix Q. Note that statistics are only reported for subgroups with at least 10 students.

For several reasons, the results of 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 on the statistical distribution of the studied subgroup. For example, it can be readily seen in Appendix Q 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 Inter-Rater Consistency

Section 4.4 of this chapter describes in detail 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 percent of student responses in grades 3–8 and 100 percent in high school. Results of the double-blind scoring were used during scoring to identify scorers who required retraining or other intervention and are presented here as evidence of the reliability of the MCAS-Alt. A summary of the inter-rater consistency results is presented in Table 4-17. Results in the table are collapsed across the tasks by content area, grade, and number of score categories (5 for Level of Complexity and 4 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 percent of responses that required a third score. This same information is provided at the item level in Appendix P.

Table 4-17. 2010 MCAS-Alt: Summary of Inter-Rater Consistency Statistics Collapsed 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*
English Language Arts	3	4	4	962	97.71%	2.18%	0.89	2.29%
		2	5	507	96.45%	2.96%	0.77	5.13%
	4	6	4	984	97.36%	2.54%	0.88	2.85%
		3	5	534	97.00%	2.06%	0.77	5.24%
	5	4	4	728	99.18%	0.82%	0.96	0.96%
		2	5	390	95.64%	2.56%	0.42	5.38%
	6	4	4	780	98.85%	1.15%	0.91	1.92%
		2	5	406	97.54%	1.48%	0.68	4.19%
	7	6	4	1,244	96.78%	3.05%	0.85	3.46%
		3	5	668	97.31%	1.50%	0.54	4.34%
	8	4	4	860	98.26%	1.74%	0.95	3.14%
		2	5	462	94.16%	3.03%	0.43	8.87%
	10	6	4	4,888	97.28%	2.58%	0.91	3.89%
		3	5	2,785	94.72%	3.16%	0.58	7.11%
Mathematics	3	4	4	948	97.78%	2.00%	0.88	2.22%
		2	5	491	98.17%	1.63%	0.81	2.65%
	4	4	4	640	97.50%	2.34%	0.87	2.66%
		2	5	345	97.97%	2.03%	0.87	3.48%
	5	4	4	768	97.01%	2.73%	0.85	3.52%
		2	5	390	96.67%	2.82%	0.61	4.62%
	6	4	4	808	98.51%	1.36%	0.92	3.59%
		2	5	413	98.06%	1.69%	0.79	4.60%
	7	4	4	870	97.59%	2.41%	0.90	2.87%
		2	5	472	98.31%	1.69%	0.76	2.12%
	8	4	4	942	98.41%	1.59%	0.94	2.76%
		2	5	500	96.40%	3.60%	0.74	7.00%
	10	10	4	4,816	97.36%	2.45%	0.91	4.49%
		5	5	2,804	95.19%	3.14%	0.49	6.06%
Science and Technology/ Engineering	5	8	4	992	98.69%	1.31%	0.95	1.71%
		4	5	534	97.75%	1.69%	0.59	3.18%
	8	8	4	1,204	97.18%	2.66%	0.87	4.24%
		4	5	654	96.02%	3.21%	0.67	7.19%

Content Area	*Grade*	*Number of Items*	*Number of Score Categories*	*Number of Included Scores*	*Percent Exact*	*Percent Adjacent*	*Correlation*	*Percent of Third Scores*
Biology	HS	6	4	4,156	97.28%	2.53%	0.90	4.02%
		3	5	2,394	93.65%	3.30%	0.42	7.89%
Chemistry	HS	6	4	378	97.62%	1.85%	0.89	7.14%
		3	5	235	88.51%	4.26%	0.41	16.17%
Physics	HS	6	4	300	96.00%	3.33%	0.81	6.00%
		3	5	168	94.64%	4.76%	0.90	12.50%
Technology/ Engineering	HS	6	4	444	93.24%	6.76%	0.88	6.98%
		3	5	242	90.91%	4.96%	0.50	11.16%

4.7 MCAS-Alt Comparability Across Years

Issues of comparability across years are 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 as 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 ensured through the classification of students into performance level categories (Table 4-18), using a lookup table that remains consistent each year. The description of each performance level remains consistent, which ensures that the meaning of students' scores is comparable from one year to the next. Table 4-19 shows the performance level lookup table (i.e., the performance level corresponding to each possible combination of dimension scores), which is used each year to combine and tally the overall performance level from individual strand scores. In addition, performance level distributions are provided in Appendix M. The distributions include results for each of the last three years.

Table 4-18. 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.
Emerging (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 4-19. MCAS-Alt Strand Performance Level Lookup

Level of complexity	*Demonstration of skills*	*Independence*	*Performance Level*
1			1
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

Level of complexity	*Demonstration of skills*	*Independence*	*Performance Level*
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
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

Measured Progress created the following primary reports for the MCAS-Alt:

- portfolio feedback form
- *Parent/Guardian Report*

4.8.1.1 Portfolio Feedback Forms

One report is produced for each student who submitted an MCAS-Alt portfolio. Student's content area performance level(s), strand dimension scores, and comments relating to those scores are printed on the report. The portfolio feedback form is a preliminary score report intended for the teacher who submitted the portfolio. General portfolio comments are also included.

4.8.1.2 Parent/Guardian Report

The *Parent/Guardian Report* is generated for all students who submitted an MCAS-Alt portfolio. It provides background information about the MCAS-Alt assessment, participation requirements, the purpose of the assessment, an explanation of scores, and contact information for further questions. Performance levels are displayed for each subject relative to all possible performance 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 school records. Sample reports are provided in Appendix T.

4.8.2 Interpretive Materials and Workshops

The 2010 report was redesigned to incorporate the information found previously in a separate interpretive guide, which was not produced. Two parent focus groups provided feedback on the report design revisions before it was finalized.

4.8.3 Decision Rules

To ensure that reported results for the MCAS-Alt are accurate relative to the collected portfolio evidence, a document outlining all decision rules is prepared prior to each reporting cycle and is reviewed and approved by the ESE. The decision rules are observed in the analyses of the MCAS-Alt test data and in reporting content area results. The decision rules are included in Appendix U.

4.8.4 Quality Assurance

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

Another quality assurance measure is 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 on staff who uses the decision rules to independently program the reporting table. The production and quality assurance tables are compared, and only after there is 100 percent agreement are the tables released for report generation.

The 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 sampled students for this purpose is very specific and can 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 program management review. The appropriate sample reports are then sent to the ESE for review and signoff.

4.9 Validity

The purpose of this report is to describe several technical aspects of the MCAS-Alt in an effort to contribute to the accumulation of validity evidence to support MCAS-Alt score interpretations. Because the combination of a test and its scores is evaluated for validity, this report presents documentation to substantiate the intended interpretations (AERA, 1999). Each of the sections in this chapter contributes important information to the validity argument by addressing one or more of the following aspects of the MCAS-Alt: development, administration, scoring, item analyses, reliability, performance levels, and reporting.

The MCAS-Alt assessments are based on, and aligned with, the Massachusetts curriculum framework content standards in English language arts, mathematics, and science and technology/engineering. The MCAS-Alt results are intended to assist educators and parents to make inferences about student achievement on the ELA, mathematics, and STE content standards; to use results to make program and instructional improvement; and as a component of school accountability.

Standards for Educational and Psychological Testing (AERA et al., 1999) provides a framework for describing sources of evidence that should be considered when constructing a validity argument. These sources include evidence based on the following five general areas: test content, response processes, internal structure, relationship to other variables, and consequences of testing. Although each of these sources may address a different *aspect* of validity, together they contribute to a body of evidence regarding the comprehensive validity of score interpretations.

4.9.1 Evidence Based on Test Development and Structure

Evidence based on internal structure is presented in the discussions of item analyses and reliability in Sections 4.5 and 4.6. Analyses of the internal structure of the assessments include classical item statistics (item difficulty, item-test correlation), correlations among the dimensions (level of complexity, demonstration of skills and concepts, and independence), fairness/bias, and reliability, including alpha coefficients, inter-rater consistency, and decision accuracy and consistency.

4.9.2 Other 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 *2010 Educator's Manual for MCAS-Alt*. Efforts by the ESE to provide training, materials, and ongoing support maximize consistency across the state, which enhances the quality and reliability of the inferences made based on results; this, in turn, contributes to the validity of the assessment.

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

Information provided in Section 4.7 shows how reported scores, including performance levels, ensure comparability of students' scores across years, which, in turn, contributes to validity.

Efforts undertaken to provide the public with accurate and clear test score information (described in Section 4.8) include reporting of performance levels that give reference points for mastery at each grade level. Performance levels and their descriptors provide a useful and consistent way to interpret scores, thereby contributing to the validity of the assessment.

4.9.3 Summary

The evidence presented in this report supports inferences related to student achievement of the skills and content represented by the content standards of ELA, mathematics, and STE on the MCAS-Alt, for 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.

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.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan Publishing Company.

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., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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.

Kingston, N. K., Kahl, S. K., Sweeney, K. P., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting performance standards: concepts, methods, and perspectives* (pp. 219–248). Mahwah, NJ: Lawrence Erlbaum.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: methods and practices*. New York: Springer-Verlag.

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. (2010). *2009–2010 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

Appendix A	Committee Membership
Appendix B	Participation Rates
Appendix C	Accommodation Frequencies
Appendix D	Standard and Nonstandard Test Accommodations
Appendix E	Item-Level Classical Statistics
Appendix F	Item-Level Score Distributions
Appendix G	Differential Item Functioning Results
Appendix H	Item Response Theory Parameters
Appendix I	Test Characteristic Curves and Test Information Functions
Appendix J	Analysis of Equating Items
Appendix K	*a*-Plots and *b*-Plots
Appendix L	Relative Position Study (Grades 3–8)
Appendix M	Performance Level Score Distributions
Appendix N	Raw to Scaled Score Lookup Tables
Appendix O	MCAS Scaled Score Distributions
Appendix P	Inter-Rater Agreement
Appendix Q	Classical Reliability
Appendix R	Sample Reports—MCAS
Appendix S	Analysis and Reporting Decision Rules—MCAS
Appendix T	Sample Reports—MCAS-Alt
Appendix U	Analysis and Reporting Decision Rules—MCAS-Alt