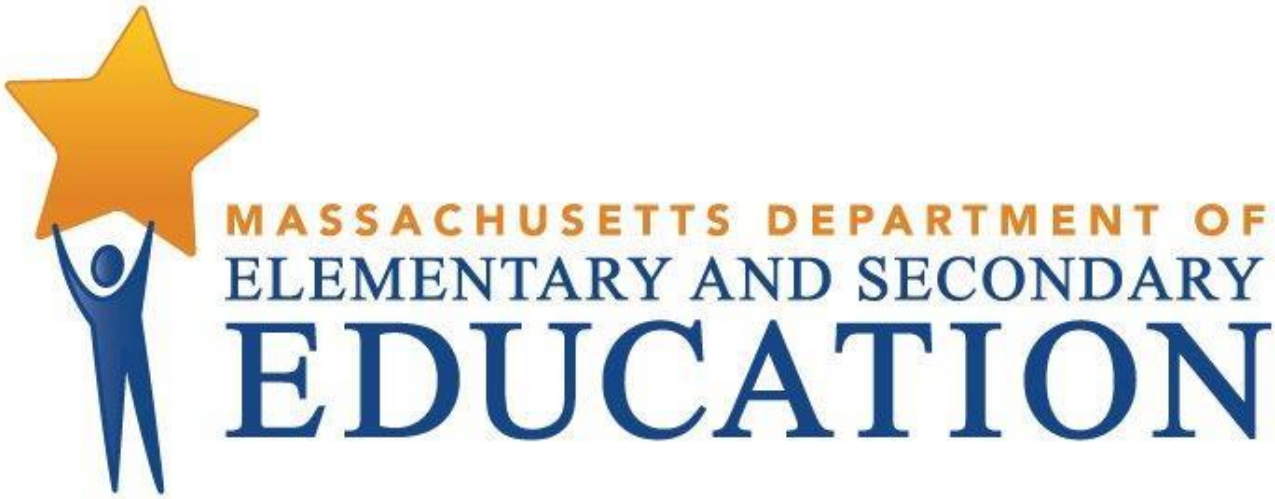


2016 MCAS and MCAS-Alt Technical Report

100 EDUCATION WAY, DOVER, NH 03820 (800) 431-8901
WWW.MEASUREDPROGRESS.ORG

This document was prepared by the
Massachusetts Department of Elementary and Secondary Education
Jeff Wulfson
Acting Commissioner

The Massachusetts Department of Elementary and Secondary Education, an affirmative action employer, is committed to ensuring that all of its programs and facilities are accessible to all members of the public. We do not discriminate on the basis of age, color, disability, national origin, race, religion, sex, sexual orientation, or gender identity.

Inquiries regarding the Department's compliance with Title IX and other civil rights laws may be directed to the Human Resources Director, 75 Pleasant St., Malden, MA 02148 781-338-6105.

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

TABLE OF CONTENTS

CHAPTER 1 OVERVIEW 1
1.1 Purposes of the MCAS 1
1.2 Purpose of This Report 1
1.3 Organization of This Report 2
1.4 Current Year Updates 2
1.4.1 Next-Generation MCAS Assessments 3
1.4.1.1 Background on the Transition to Next-Generation Assessments 3
1.4.2 Changes to the MCAS Assessments in 2016 4
1.4.3 Spring 2016 District Assessment Decision 4
1.4.4 Caution About Interpretation of 2016 MCAS Test Results for Grades 3–8 5
CHAPTER 2 THE STATE ASSESSMENT SYSTEM: MCAS 6
2.1 Introduction 6
2.2 Guiding Philosophy 6
2.3 Alignment to the Massachusetts Curriculum Frameworks 7
2.4 Uses of MCAS Results 7
2.5 Validity of MCAS and MCAS-Alt 7
CHAPTER 3 MCAS 8
3.1 Overview 8
3.2 Test Design and Development 8
3.2.1 Test Specifications 9
3.2.1.1 Criterion-Referenced Test 9
3.2.1.2 Item Types 9
3.2.1.3 Description of Test Design 11
3.2.2 ELA Test Specifications 11
3.2.2.1 Standards 11
3.2.2.2 Item Types 13
3.2.2.3 Test Design 13
3.2.2.4 Blueprints 17
3.2.2.5 Cognitive Levels 17
3.2.2.6 Reference Materials 17
3.2.2.7 Passage Types 17
3.2.3 Mathematics Test Specifications 18
3.2.3.1 Standards 18
3.2.3.2 Item Types 19
3.2.3.3 Test Design 20
3.2.3.4 Blueprints 22
3.2.3.5 Cognitive Levels 23
3.2.3.6 Use of Calculators, Reference Sheets, Tool Kits, and Rulers 24
3.2.4 STE Test Specifications 24
3.2.4.1 Standards 24
3.2.4.2 Item Types 25
3.2.4.3 Test Design 25
3.2.4.4 Blueprints 27
3.2.4.5 Cognitive and Quantitative Skills 28
3.2.4.6 Use of Calculators, Formula Sheets, and Rulers 29
3.2.5 Test Development Process 29
3.2.5.1 ELA Passage Selection and Item Development 31
3.2.5.2 Item Editing 33
3.2.5.3 Field-Testing Items 34
3.2.5.4 Scoring of Field-Tested Items 34
3.2.5.5 Data Review of Field-Tested Items 34
3.2.5.6 Item Selection and Operational Test Assembly 35
3.2.5.7 Operational Test Draft Review 36

3.2.5.8 Special Edition Test Forms 36
3.3 Test Administration 38
3.3.1 Test Administration Schedule 38
3.3.2 Security Requirements 39
3.3.3 Participation Requirements 40
3.3.3.1 Students Not Tested on Standard Tests 40
3.3.4 Administration Procedures 41
3.4 Scoring 41
3.4.1 Machine-Scored Items 42
3.4.2 Hand-Scored Items 42
3.4.2.1 Scoring Location and Staff 42
3.4.2.2 Benchmarking Meetings 43
3.4.2.3 Scorer Recruitment and Qualifications 43
3.4.2.4 Methodology for Scoring Polytomous Items 44
3.4.2.5 Scorer Training 47
3.4.2.6 Leadership Training 48
3.4.2.7 Monitoring of Scoring Quality Control 48
3.4.2.8 Interrater Consistency 49
3.5 Classical Item Analyses 51
3.5.1 Classical Difficulty and Discrimination Indices 51
3.5.2 DIF 54
3.5.3 Dimensionality Analysis 56
3.5.3.1 DIMTEST Analyses 57
3.5.3.2 DETECT Analyses 57
3.6 MCAS IRT Scaling and Equating 59
3.6.1 IRT 61
3.6.2 IRT Results 63
3.6.3 Equating 65
3.6.4 Achievement Standards 66
3.6.5 Reported Scaled Scores 67
3.7 MCAS Reliability 70
3.7.1 Reliability and Standard Errors of Measurement 70
3.7.2 Subgroup Reliability 71
3.7.3 Reporting Subcategory Reliability 72
3.7.4 Reliability of Achievement Level Categorization 72
3.7.5 Decision Accuracy and Consistency Results 73
3.8 Reporting of Results 77
3.8.1 *Parent/Guardian Report* 77
3.8.2 Decision Rules 79
3.8.3 Quality Assurance 79
3.9 MCAS Validity 80
3.9.1 Test Content Validity Evidence 80
3.9.2 Response Process Validity Evidence 80
3.9.3 Internal Structure Validity Evidence 80
3.9.4 Validity Evidence in Relationships to Other Variables 81
3.9.5 Efforts to Support the Valid Use of MCAS Data 81
CHAPTER 4 MCAS-ALT 86
4.1 Overview 86
4.1.1 Background 86
4.1.2 Purposes of the Assessment System 86
4.1.3 Format 87
4.2 Test Design and Development 87
4.2.1 Test Content 87
4.2.1.1 Access to the Grade-Level Curriculum 88
4.2.1.2 Assessment Design 89
4.2.1.3 Assessment Dimensions (Scoring Rubric Areas) 90
4.2.1.4 MCAS-Alt Grade-Level and Competency Portfolios 91
4.2.2 Test Development 91
4.2.2.1 Rationale 91

4.2.2.2 Role of the Advisory Committee 92
4.3 Test Administration 92
4.3.1 Evidence Collection 92
4.3.2 Construction of Portfolios 93
4.3.3 Participation Requirements 94
4.3.3.1 Identification of Students 94
4.3.3.2 Participation Guidelines 94
4.3.3.3 MCAS-Alt Participation Rates 96
4.3.4 Educator Training 96
4.3.5 Support for Educators 97
4.4 Scoring 97
4.4.1 Scoring Logistics 98
4.4.2 Selection, Training, and Qualification of Scorers 98
4.4.3 Scoring Methodology 100
4.4.3.1 ELA–Writing Scoring Methodology 104
4.4.4 Monitoring the Scoring Quality 105
4.4.5 Scoring of Grade-Level Portfolios in Grades 3 Through 8 and Competency Portfolios in High School 106
4.5 MCAS-Alt Classical Item Analyses 107
4.5.1 Difficulty 107
4.5.2 Discrimination 108
4.5.3 Structural Relationships Between Dimensions 110
4.5.4 Differential Item Functioning 111
4.6 Bias/Fairness 111
4.7 Characterizing Errors Associated With Test Scores 112
4.7.1 MCAS-Alt Reliability 112
4.7.2 Subgroup Reliability 113
4.7.3 Interrater Consistency 114
4.8 MCAS-Alt Comparability Across Years 115
4.9 Reporting of Results 118
4.9.1 Primary Reports 118
4.9.1.1 Portfolio Feedback Forms 118
4.9.1.2 Parent/Guardian Report 118
4.9.2 Decision Rules 118
4.9.3 Quality Assurance 118
4.10 MCAS-Alt Validity 119
4.10.1 Test Content Validity Evidence 119
4.10.2 Internal Structure Validity Evidence 119
4.10.3 Response Process Validity Evidence 119
4.10.4 Efforts to Support the Valid Reporting and Use of MCAS-Alt Data 120
4.10.5 Summary 121

REFERENCES 123

APPENDICES 126

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 α-Plots and b-Plots
Appendix L Achievement Level Score Distributions
Appendix M Raw to Scaled Score Look-Up Tables

Appendix N MCAS Scaled Score Distributions
Appendix O Interrater Consistency
Appendix P Classical Reliability
Appendix Q Sample Reports—MCAS
Appendix R Analysis and Reporting Decision Rules—MCAS
Appendix S Sample Reports—MCAS-Alt
Appendix T Analysis and Reporting Decision Rules—MCAS-Alt
Appendix U ELA–Writing Scoring Rubrics—MCAS-Alt
Appendix V PARCC Try-Out Item-Level Descriptive Statistics

Chapter 1 Overview

1.1 Purposes of the MCAS

The Massachusetts Education Reform Mandate

The Massachusetts Education Reform Act of 1993 requires the establishment of a statewide testing program. The Act specifies that the testing program must

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

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

Since 1998, the Massachusetts Comprehensive Assessment System (MCAS) has been the Commonwealth's program for student assessment, developed in accordance with the Massachusetts Education Reform Act of 1993. To fulfill the requirements of the Act, the MCAS is designed to

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

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

1.2 Purpose of This Report

The purpose of this *2016 MCAS and MCAS-Alt Technical Report* is to document the technical quality and characteristics of the 2016 MCAS operational tests, to present evidence of the validity and reliability of test score interpretations, and to describe modifications made to the program in 2016. Technical reports for 1998 to 2015 are available on the ESE website at www.doe.mass.edu/mcas/tech/?section=techreports. The *2016 MCAS and MCAS-Alt Technical*

Report is designed to supplement the technical reports issued for previous MCAS administrations by providing information specific to the 2016 MCAS test administrations. Previous technical reports, as well as other documents referenced in this report, provide additional background information about the MCAS program and its development and administration.

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

1.3 Organization of This Report

This report provides detailed information regarding test design and development, scoring, and analysis and reporting of 2016 MCAS results at the student, school, district, and state levels. This detailed information includes, but is not limited to, the following:

- an explanation of test administration
- an explanation of equating and scaling of tests
- statistical and psychometric summaries:
    - item analyses
    - reliability evidence
    - validity evidence

In addition, the technical appendices contain detailed item-level and summary statistics related to each 2016 MCAS test and its results.

Chapter 1 of this report provides a brief overview of what is documented within the report, including updates made to the MCAS program during 2016. Chapter 2 explains the guiding philosophy, purpose, uses, components, and validity of MCAS. The next two chapters cover the test design and development, test administration, scoring, and analysis and reporting of results for the standard MCAS assessment (Chapter 3) and the MCAS Alternate Assessment (Chapter 4). These two chapters include information about the characteristics of the test items, how scores were calculated, the reliability of the scores, how scores were reported, and the validity of the results. Numerous appendices, which appear after Chapter 4, are referenced throughout the report.

1.4 Current Year Updates

The 2016 MCAS assessments marked a transition from the legacy MCAS tests (administered from 1998 to 2016) to the next-generation MCAS tests, which will be introduced in 2017. Many of the changes reported in this section were made in response to this transition.

In addition, 2016 was a "choice" testing year where the majority of students in the state at grades 3–8 took assessments developed by the Partnership for Advancement of Readiness for College and Careers (PARCC), and a smaller percentage of students took the legacy MCAS assessments. Sections 1.4.3 and 1.4.4 provide information on the numbers of students taking each assessment and on the implications of the testing choice for the 2016 test results.

1.4.1 Next-Generation MCAS Assessments

On November 17, 2015, the Massachusetts Board of Elementary and Secondary Education voted to endorse the use of next-generation MCAS assessments starting in 2017. The next-generation MCAS assessments are designed to build upon the best aspects of the legacy MCAS assessments and will include innovative items developed by PARCC. The assessments will include the following elements:

- high-quality test items aligned to the Massachusetts learning standards
- new item types that more deeply assess both skills and knowledge; for example:
  - Writing to text in ELA
  - Solving complex problems in mathematics
- achievement levels that send clear signals to students, parents, and educators about readiness for work at the next level
- online and paper test administrations, with a goal of phasing in online testing so that computer-based tests are administered statewide in 2019
- online student accessibility features and accommodations

The next-generation MCAS assessments will be phased in. In 2017, all students in grades 3–8 will take the next-generation assessments in ELA and mathematics. Next-generation ELA and mathematics assessments will be administered at grade 10 for the first time in 2019. In STE, next-generation assessments will be administered to students in grades 5 and 8 in 2019, with the first administration for high school students still to be determined.

Additional information on the next-generation MCAS assessments is available at www.doe.mass.edu/mcas/nextgen/resources.html.

1.4.1.1 Background on the Transition to Next-Generation Assessments

The Board's vote of November 2015 was the culmination of a multi-year process to develop a plan for transitioning Massachusetts to next-generation assessments. Following are some key milestones from that process:

- **2011**: Massachusetts joins PARCC, a multi-state consortium formed to develop a new set of assessments for ELA and mathematics.
- **2013**: The Board votes to conduct a two-year "test drive" of the PARCC assessments in order to decide whether Massachusetts should adopt them in place of the existing MCAS assessments in ELA and mathematics.
- **2014**: The PARCC assessments are field-tested in a randomized sample of schools in Massachusetts and in the other consortium states.
- **Spring 2015**: Massachusetts districts (including charter schools and vocational-technical high schools) are given the choice of administering either PARCC or MCAS to their students in grades 3–8. Roughly half of the students at those grade levels take the MCAS assessments, and roughly half take PARCC.
- **November 2015**: Commissioner Mitchell Chester recommends to the Board that the state transition to a next-generation MCAS that would be administered for the first time in spring 2017 and that would utilize both MCAS and PARCC test items. The Board votes to endorse his recommendation.

1.4.2 Changes to the MCAS Assessments in 2016

The transition plan approved by the Board called for a small number of PARCC items to be included on the 2016 MCAS ELA and mathematics tests at grades 3–8. The PARCC items were used in order to provide MCAS test-takers the opportunity to experience PARCC items while the next-generation assessment was being developed.

- In mathematics, six PARCC items were included at each grade level, with three items placed at the end of each session. The item types included multiple-choice, multiple-select, constructed-response, and composite items (two-part items comprising two item types).
- In ELA, the PARCC narrative writing task (four evidence-based selected-response items and one prose-constructed-response item) was administered as a separate, timed portion at the end of session 2.

Of the PARCC item types included on the tests, several were new to the MCAS program, including multiple-select items, composite items, and evidence-based selected-response items. All PARCC items were paper-based items, and all were placed at the end of test sessions so that they would not influence student results on the MCAS items. Results on the PARCC items were reported but did not count toward students' MCAS scores. State results from the PARCC items appear in Appendix V.

Along with the inclusion of the PARCC items, two other changes were made to the 2016 ELA and mathematics tests at grades 3–8:

- Because the PARCC narrative writing task was administered at each grade level, the MCAS composition at grades 4 and 7 was eliminated.
- Field-test items were not included on the 2016 tests at grades 3–8 because the legacy phase of the MCAS testing program was ending.

Changes to the equating design for the 2016 ELA and mathematics tests at grades 3–8 are described in section 3.6.3.

1.4.3 Spring 2016 District Assessment Decision

For the spring 2016 test administrations at grades 3–8, the Department offered a revised "choice" process in which districts that administered PARCC in spring 2015 were required to do so again, while districts that administered MCAS in spring 2015 were allowed to choose MCAS or PARCC. Table 1-1 provides a summary of districts' assessment decisions for both the spring of 2016 and the spring of 2015.

Table 1-1. 2016 MCAS: Assessment Choices for Spring 2016 vs. Spring 2015

	Number of Public Districts	MCAS			PARCC		
		# of Districts	*% of Students*	*# of Students*	*# of Districts*	*% of Students*	*# of Students*
2016 Grades 3–8	360	118	**28%**	121,000	243	**72%**	306,000
2015 Grades 3–8	359	165	46%	207,500	194	54%	227,000

Of the districts that switched from MCAS to PARCC in 2016, many were larger urban districts, while the remaining districts taking MCAS tended to be smaller suburban districts. Consequently, as shown in Table 1-2, the demographic differences between the MCAS test-takers and PARCC test-

takers were more significant in 2016 than in 2015. The demographic characteristics of the students taking MCAS in 2016 also veered significantly from the state characteristics. For example, among MCAS test-takers in grades 3–8, the average proportion of students classified as economically disadvantaged fell from 27% in 2015 to 22% in 2016 (across the state in 2016, the percentage was 32%). At the same time, the proportion of white students among MCAS test-takers rose from 65% in 2015 to 72% in 2016 (across the state in 2016, the percentage was 60%).

Table 1-2. 2016 MCAS: Demographic Characteristics of MCAS and PARCC Students, 2016 vs. 2015

	Test	% Economically Disadvantaged	% English Learner	% White	% Black	% Asian	% Hispanic
Spring 2016	MCAS	**22%**	4%	**72%**	4%	7%	9%
Grades 3–8	PARCC	36%	11%	55%	9%	5%	20%
Spring 2015	MCAS	**27%**	8%	**65%**	5%	6%	16%
Grades 3–8	PARCC	30%	9%	60%	11%	5%	18%

1.4.4 Caution About Interpretation of 2016 MCAS Test Results for Grades 3–8

In 2015, the groups taking MCAS and PARCC differed only slightly in size and student demographics. Because of this, the Department was able to identify samples of 2015 MCAS and PARCC test-takers that were representative of all students in the state. These representative samples were then used in generating and reporting the 2015 test results. For example, the MCAS representative sample was used to equate the 2015 MCAS ELA and mathematics tests for grades 3–8 to the tests taken the previous year. The representative sample was also used to estimate "state" results for context on *Parent/Guardian* reports and public reports, as well as for summaries of item- and standard-level statistics.

In 2016, because so few urban students participated in MCAS, the Department was unable to draw a sample of MCAS test-takers that was representative of the state. And without a valid representative sample, the Department was unable to estimate statewide results. Consequently, the Department is not reporting aggregate statewide results for grades 3–8 in ELA and mathematics in 2016.

Readers of this report should be aware that the MCAS results in ELA and mathematics presented here for grades 3–8 are based on a non-representative sample and should not be compared with results in prior years. In addition, because a smaller number of students in grades 3–8 took MCAS in ELA and mathematics in 2016 than in prior years, reliability coefficients and other statistics may be impacted by the lower number (N) sizes.

All MCAS results reported for grade 10, and those reported for STE, were unaffected by these changes, and trends for these grades and subjects were maintained.

Although the 2016 MCAS test administration did not provide state-level student achievement results in grades 3–8 in ELA and mathematics, it did provide students, parents, and educators with achievement results aligned to the Massachusetts learning standards. Educators can continue to use this information to inform educational programming. Students and parents can use MCAS results to track achievement and plan interventions, if necessary (to obtain extra help or tutoring, for example).

Chapter 2 The State Assessment System: MCAS

2.1 Introduction

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

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

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

2.2 Guiding Philosophy

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

Since the first administration of the MCAS tests in 1998, the ESE has gathered evidence from many sources suggesting that the assessment reforms introduced in response to the Massachusetts Education Reform Act of 1993 have been an important factor in raising the academic expectations of all students in the Commonwealth and in making the educational system in Massachusetts one of the country's best.

The MCAS testing program has been an important component of education reform in Massachusetts for over 15 years. The program continues to evolve. As described in section 1.4, Massachusetts is transitioning in 2017 from the legacy MCAS to next-generation MCAS assessments that will

- align MCAS items with the current and revised Massachusetts Academic Learning Standards;
- incorporate innovations in assessment, such as online testing, technology-enhanced item types, and upgraded accessibility and accommodation features;
- provide achievement information that sends clear signals about readiness for work at the next level; and
- ensure the MCAS measures the knowledge and skills students need to meet the challenges of the 21st century.

2.3 Alignment to the Massachusetts Curriculum Frameworks

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. All MCAS tests are designed to measure MCAS performance levels based on performance level descriptors derived from the Massachusetts curriculum frameworks. Therefore, the primary inferences drawn from the MCAS test results are about the levels of students' mastery of the standards contained in the Massachusetts curriculum frameworks.

2.4 Uses of MCAS Results

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

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

2.5 Validity of MCAS and MCAS-Alt

Validity information for the MCAS and MCAS-Alt assessments 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 [DIF], item response theory [IRT] statistics, reliability, dimensionality, decision accuracy and consistency [DAC]); and reporting. Validity information is described in detail in sections of this report and is summarized for each of the assessment components in their respective Validity subsections (section 3.9 for MCAS and section 4.10 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 ESE staff with assistance and support from the assessment contractor, Measured Progress (MP). Massachusetts educators play a key role in the MCAS through service on a variety of committees related to the development of MCAS test items, the development of MCAS performance level descriptors, and the setting of performance standards. The program is supported by a five-member national Technical Advisory Committee (TAC) as well as measurement specialists from the University of Massachusetts–Amherst.

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

3.2 Test Design and Development

The 2016 MCAS test administration included operational tests in the following grades and content areas:

- grades 3–8 and grade 10 ELA, including a composition component at grade 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 2016 MCAS administration also included retest opportunities in ELA and mathematics in November 2015 and March 2016 for students beyond grade 10 who had not yet passed the standard grade 10 tests. A February biology test was also administered. This test could be taken as a retest or as a first experience of MCAS science for students in block-scheduled science classes who completed their biology class in January.

In 2016, the grades 3–8 ELA and mathematics tests included PARCC items. These items were scored for informational purposes only and did not count toward the MCAS results. The ESE included the PARCC sections to allow students to "try out" the different kinds of questions that will

appear on the state's next-generation MCAS tests, which will be administered beginning in spring 2017. These PARCC items were placed at the end of test sessions, as described in section 1.4.2. To accommodate the PARCC items, no field-test items or matrix equating items were included on the 2016 ELA and mathematics tests at grades 3–8.

In 2016, Massachusetts districts were given the choice of administering either PARCC or MCAS assessments in ELA and mathematics to their students in grades 3–8. As described in section 1.4.3, 28% of students at those grades took the MCAS tests.

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. The MCAS assesses only the content and processes described in the Massachusetts curriculum frameworks. In 2011, Massachusetts adopted new curriculum standards in mathematics and ELA. In 2012, all new items were double-coded to the older standards and the new standards with the older standard considered the primary standard. In 2013, items continued to be double-coded to both the new and the former standards, but in 2013 the new Massachusetts standards were considered the primary standards. Starting in 2014, all new test items for grades 3–8 ELA and mathematics were coded only to the 2011 standards. At grade 10, items continued to be double-coded with the first or primary code coming from the 2011 Massachusetts standards. At all grade levels, older ELA and mathematics items used in tests (those developed before the 2011 standards were adopted) were also coded to the new standards. All items on the STE tests were coded to the *2006 Massachusetts Science and Technology/Engineering Curriculum Framework*.

3.2.1.2 Item Types

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

- **Multiple-choice** items are used to provide breadth of coverage within a content area. Because they require no more than one 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 composition component of the ELA assessment. Each multiple-choice item requires that students select the single best answer from four response options. Multiple-choice items are aligned to one primary standard. They are machine-scored; correct responses are worth one score point, and incorrect and blank responses are assigned zero score points. Though considered as wrong responses, blanks are disaggregated from the incorrect responses.
- **One-point short-answer** mathematics items are used to assess students' skills and abilities to work with brief, well-structured problems that have one or a very limited number of solutions (e.g., mathematical computations). Short-answer items require approximately two minutes for most students to answer. The advantage of this type of item is that it requires students to demonstrate knowledge and skills by generating, rather than selecting, an answer. One-point short-answer items are hand-scored and assigned one point (correct) or zero points (blank or incorrect). The blanks are disaggregated from the incorrect responses.

- **Two-point open-response** items are used in the grade 3 mathematics test instead of the 4-point open-response items used in all other tests. 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 completely correct, one-point responses are partially correct, and responses with a score of zero are completely incorrect. Blank responses also receive a score of zero. The blanks are disaggregated from the incorrect responses.
- **Two-point short-response** items are used only 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 receive a score of zero. The blanks are disaggregated from the incorrect responses.
- **Four-point open-response** items typically require students to use higher-order thinking skills—such as evaluation, analysis, and summarization—to construct satisfactory responses. Four-point open-response items are administered in all content areas at all grades except grade 3 mathematics. (The grade 3 ELA assessment includes one 4-point open-response item as well as four 2-point short-response items.) Open-response items take most students approximately 5 to 10 minutes to complete. Open-response items are hand-scored by scorers trained in the specific requirements of each question scored. Students may receive up to four points per open-response item. Totally incorrect or blank responses receive a score of zero. The blanks are disaggregated from the incorrect responses.
- **Writing prompts** are administered to all students in grade 10 as part of the ELA test. The writing assessment consists of two sessions separated by a 10-minute break. During the first session, students write a draft composition. In the second session, students write a final composition based on that draft. Each composition is hand-scored by trained scorers. Students receive two scores: one for topic development (0 to 6 points) and the other for standard English conventions (0 to 4 points). Student reports include a score for each of these dimensions. Each student composition is scored by two different scorers; the final score is a combination of both sets of scores, so students may receive up to 20 points for their compositions. These 20 composition points amount to 28% of a student's overall ELA score in grade 10, the grade in which the writing prompts are administered.

In addition to the item types above, the following item types were used in the grades 3–8 ELA and mathematics tests in the PARCC item section.

- **Multiple-select** items were used in the grades 3–8 mathematics tests. Students select 1–3 correct answers from a set of answer options. The items are machine-scored; correct responses are worth one score point, and incorrect and blank responses are assigned zero score points.
- **Constructed-response** items were used in the grades 3–8 mathematics tests. Constructed-response items are hand-scored by scorers trained in the specific requirements of each question scored. Students may receive up to three or four points per constructed-response item. Totally incorrect or blank responses receive a score of zero.
- **Evidence-based selected-response (EBSR)** items were used in the grades 3–8 ELA tests. EBSR items are two-part multiple-choice or multiple-select items. Students respond by selecting the single best answer from four response options in the first part. Students then respond to the second part by selecting evidence from the stimulus that supports the answer from the first part. (In the grade 8 ELA test, one EBSR provides six response options in both

parts, with two correct responses in each part.) The items are machine-scored; correct responses are worth two score points, partially correct answers are worth one score point, and incorrect and blank responses are assigned zero score points. Students must answer the first part correctly in order to receive one or two score points.

- **Prose-constructed-response** items were used in the grades 3–8 ELA tests. Prose-constructed-response items are hand-scored by scorers trained in the specific requirements of each question scored.
- **Composite two-part** items were used in the grades 3–8 mathematics tests. Each part of these two-part items is scored separately and the sum of the scores is the score earned on the item.

3.2.1.3 Description of Test Design

The MCAS assessment instruments are structured using both common and matrix items. Identical common items are administered to all students in a given grade. Student scores are based on student performance on common items only. 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. Equating and field-test items are divided among the multiple forms of the test for each grade and content area. The number of test forms varies by grade and content area but ranges between 1 and 32 forms. Each student takes only one form of the test and therefore answers a subset of field-test items and/or equating items. Field-test and equating items are not distinguishable to test-takers. Because all students participate in the field test, an adequate sample size (approximately 1,800 students per item) is obtained to produce reliable data that can be used to inform item selection for future tests.

As noted above, the 2016 ELA and mathematics assessments at grades 3–8 omitted the field-test and equating sections in order to accommodate the inclusion of PARCC items.

3.2.2 ELA Test Specifications

3.2.2.1 Standards

The MCAS ELA tests measure learning standards from the *2011 Massachusetts Curriculum Framework for English Language Arts and Literacy*.

The following standards are assessed on the grades 3–8 ELA tests and on the reading comprehension portion of the grade 10 ELA test.

Anchor Standards for Reading

- Key Ideas and Details (Standards 1–3)
- Craft and Structure (Standards 4–6)
- Integration of Knowledge and Ideas (Standards 7–9)

Anchor Standards for Language

- Conventions of Standard English (Standards 1 and 2)
- Knowledge of Language (Standard 3)
- Vocabulary Acquisition and Use (Standards 4–6)

The composition portion of the grade 10 ELA test assesses the following standards. (As noted in section 1.4.2, the composition portion of the ELA tests at grades 4 and 7 was omitted in 2016 to accommodate the PARCC items.)

Anchor Standards for Writing

- Text Types and Purposes (Standard 1)
- Production and Distribution of Writing (Standards 4 and 5)

For grade-level articulation of these standards, please refer to the *2011 Massachusetts Curriculum Framework for English Language Arts and Literacy*. This assessment year, 2016, the MCAS ELA assessments were fully aligned to the 2011 standards. Items field-tested in the 2016 assessment for future use were aligned only to the 2011 Massachusetts ELA standards except for the grade 10 ELA items, which were aligned to both the 2001/2004 standards and the 2011 standards listed on the previous page. The 2001/2004 standards assessed on the grade 10 ELA assessment are listed below.

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

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 November 2015 and March 2016 ELA retests were aligned to both the 2001/2004 and 2011 Massachusetts ELA standards.

3.2.2.2 Item Types

The grades 3–8 ELA tests and the reading comprehension portion of the grade 10 ELA test use a mix of multiple-choice and open-response items. The grade 3 test also includes some 2-point short-response items. Additionally, grade 10 students take a composition test as part of their ELA test administration. PARCC item types were also included in grades 3–8.

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. 2016 MCAS: ELA Item Types and Score Points

Item Type	Possible Raw Score Points
Multiple-choice	0 or 1
Short-response[1]	0, 1, or 2
Open-response	0, 1, 2, 3, or 4
Writing prompt[2]	0 to 20

[1] Only grade 3 includes 2-point short-response items (along with one 4-point open-response item).
[2] Only administered at grade 10.

PARCC Item Types	Possible Raw Score Points
Evidence-based selected-response	0, 1, or 2
Prose-constructed-response	0 to 12[3] 0 to 15[4]

[3] grades 3–5
[4] grades 6–8

3.2.2.3 Test Design

The ELA tests at grades 3–8 have traditionally been made up of a reading comprehension portion (all grades) and a composition portion (grades 4 and 7 only). In 2010, as part of an effort to reduce testing time, the reading comprehension tests at grades 3–8 were shortened by eliminating one session, going from three sessions to two sessions. In 2016, the composition was discontinued at grades 4 and 7, for the reasons noted in section 1.4.2. With the discontinuation of the composition, the tests at grades 3–8 are no longer being called "ELA reading comprehension tests"; they are simply called "English Language Arts (ELA) tests."

Table 3-2 shows the current design of the ELA tests at grades 3–8. MCAS tests are untimed, and the times below are approximate.

Table 3-2. 2016 MCAS: ELA Test Designs, Grades 3–8

Grade	# of Sessions	Minutes per Session	Common Points	PARCC Item Points
3	2	60	48	20
4–5	2	60	52	20
6–8	2	60	52	23

The grade 10 ELA test continues to be made up of a reading comprehension portion (three sessions, each approximately 45 minutes in length) and a composition portion.

Grade 3 ELA Test

The common portion of this test includes two long passages and three short passages. Each long passage item set typically includes 10 multiple-choice items and either one 4-point open-response item or two 2-point short-response items. Each short passage item set generally includes five or six multiple-choice items and one short-response item (or no short-response item). No 4-point open-response items are associated with short passages on the 2016 grade 3 ELA assessment. The test contains a total of 48 common points distributed across two testing sessions.

Grades 4–8 ELA Tests

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

Grade 10 ELA Reading Comprehension Test

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

Grade 10 ELA Composition

Students in grade 10 must also complete the composition portion of the MCAS. The composition portion of the ELA test consists of one writing prompt with a total value of 20 points (12 points for topic development and 8 points for standard English conventions). The composition score accounts for 28% of a student's total raw score for ELA. As in previous years, the 2016 composition at grade 10 assessed literary analysis.

ELA Retests

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

Distribution of Common and Matrix Items

Table 3-3 lists the distribution of ELA common and PARCC items in each grade-level test and in the retests.

Table 3-3. 2016 MCAS: Distribution of ELA Common and PARCC Items by Grade and Item Type

Grade and Test			Items per Form					
			Common				PARCC[a]	
Grade	*Test*	*# of Forms*	*MC*	*SR*	*OR*	*WP*	*EBSR*	*PCR*
3	English Language Arts	1	36	4	1		4	1
4	English Language Arts	1	36		4		4	1
5	English Language Arts	1	36		4		4	1
6	English Language Arts	1	36		4		4	1
7	English Language Arts	1	36		4		4	1
8	English Language Arts	1	36		4		4	1

Grade and Test			Items per Form								Total Matrix Positions 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*
10	Reading Comprehension	25	36		4		12		2		36[c]		6[c]		264		44	
	Composition	2[b]				1												
Retest[d]	Reading Comprehension	1	36		4													
	Composition	1				1												
	Reading Comprehension	1	36		4													
	Composition	1				1												

[a] PARCC items did not count toward students' MCAS scores.
[b] The ELA composition is field-tested out of state.
[c] The grade 10 ELA test is pre-equated; however, in 2016, because of the 2015 rescaling of ELA items, equating items were added to the test for the third year in a row.
[d] ELA retests consist of common items only.

3.2.2.4 Blueprints

Table 3-4 shows the test specifications—the percentage of common item points aligned to the Massachusetts ELA curriculum framework strands—for the MCAS 2016 ELA tests.

Table 3-4. 2016 MCAS: Distribution of ELA Item Points by Percentage Across Strands by Grade

Framework Strand	Percent of Raw Score Points at Each Grade						
	3	*4*	*5*	*6*	*7*	*8*	*10*
Language	19	15	12	12	10	13	7
Reading	81	85	88	88	90	87	65
Writing							28
Total	100	100	100	100	100	100	100

3.2.2.5 Cognitive Levels

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

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

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

3.2.2.6 Reference Materials

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

3.2.2.7 Passage Types

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

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

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

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

The reading comprehension portion of the MCAS ELA test is designed to include a set of passages with a balanced representation of male and female characters; races and ethnicities; and urban, suburban, and rural settings. It is important that passages be of interest to the age group being tested.

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 their grade-level expertise when participating in passage selection as members of the Assessment Development Committees (ADCs).

Items based on ELA reading passages require students to demonstrate skills in both literal comprehension (cognitive level 1), in which the answer is stated explicitly in the text, and inferential comprehension (cognitive levels 2 and 3), in which the answer is implied by the text or the text must be connected to relevant prior knowledge to determine an answer. Items focus on the reading skills reflected in the content standards and require students to use reading skills and strategies to answer correctly.

Items coded to the language standards use the passage as a stimulus for the items. There are no standalone multiple-choice, short-response, or open-response items on the MCAS ELA assessments. All vocabulary, grammar, and mechanics questions on the MCAS ELA tests are derived from a passage. The 2016 ELA composition writing prompts are not associated with a specific reading passage.

3.2.3 Mathematics Test Specifications

3.2.3.1 Standards

For grades 3–8, all items on the 2016 mathematics assessments were aligned to the *2011 Massachusetts Curriculum Framework for Mathematics*. The items on the 2016 grade 10 mathematics assessment were aligned to 2011 standards that matched content in the 2000/2004 standards.

The 2011 standards are grouped by domains at grades 3–8 and by conceptual categories at the high school level.

Domains for Grades 3–5

- Operations and Algebraic Thinking
- Number and Operations in Base Ten
- Number and Operations—Fractions
- Measurement and Data
- Geometry

Domains for Grades 6 and 7

- Ratios and Proportional Relationships
- The Number System
- Expressions and Equations
- Geometry
- Statistics and Probability

Domains for Grade 8

- The Number System
- Expressions and Equations
- Functions
- Geometry
- Statistics and Probability

High School Conceptual Categories

- Number and Quantity
- Algebra
- Functions
- Geometry
- Statistics and Probability

3.2.3.2 Item Types

The grade 10 test and the MCAS portion of the mathematics tests in grades 3–8 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. PARCC item types were also included in grades 3–8. 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. 2016 MCAS: Mathematics Item Types and Score Points

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

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

PARCC Item Type	Possible Raw Score Points
Multiple-choice	0 or 1
Multiple-select	0 or 1
Constructed-response **	0, 1, 2, 3, or 4
Composite – 2 part	0, 1, or 2

** Constructed-response items are worth 3 or 4 points.

3.2.3.3 Test Design

In 2010, as part of an effort to reduce testing time, the mathematics tests in grades 3–8 were shortened by eliminating some of the matrix slots. Table 3-6 shows the current designs.

Table 3-6. 2016 MCAS: Mathematics Test Designs, Grades 3–8

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

The mathematics tests typically comprise common and matrix items. The matrix slots in each test form are used to field-test potential items or to equate the current year's test to that of previous years by using previously administered items. As noted in section 3.2, the 2016 mathematics assessments at grades 3–8 omitted the field-test and equating sections in order to accommodate the inclusion of PARCC items. Table 3-7 shows the distribution of common and PARCC items on the 2016 mathematics tests.

Table 3-7. 2016 MCAS: Distribution of Mathematics Common and PARCC Items by Grade and Item Type

Grade	# of Forms	Items per Form						
		Common			*PARCC*			
		MC	*SA*	*OR*	*MC*	*MS*	*CR*	*Comp*
3	1	26	6	4[a]	2	1	2	1
4	1	32	6	4	2	1	2	1
5	1	32	6	4	2	1	2	1
6	1	32	6	4	3	0	2	1
7	1	32	6	6	1	2	2	1
8	1	32	6	6	2	1	2	1

[a] Open-response items only at grade 3 are worth two points.

Grade	# of Forms	Items per Form						Total Matrix Items Across Forms								
		Common			*Matrix*			*Total Slots*			*Equating Slots*			*Field-Test Slots (available)*		
		MC	*SA*	*OR*	*MC*	*SA*	*OR*	*MC*	*SA*	*OR*	*MC*	*SA*	*OR*[b]	*MC*	*SA*[a]	*OR*[a]
10	32	32	4	6	7	1	2	224	32	64	21[c]	3[c]	4[c]	203	29	60
Retest[d]	1	32	4	6												
	1	32	4	6												

[a] The number of field-test slots available may not be the number of unique field-test items. In many instances, items are repeated in different forms of the test (especially in the case of short-answer and open-response items).
[b] Open-response items only at grade 3 are worth two points.
[c] The grade 10 mathematics test is pre-equated.
[d] Mathematics retests consist of common items only.

3.2.3.4 Blueprints

Tables 3-8 through 3-10 show the test specifications—the distribution of common item points across the Massachusetts mathematics curriculum framework domains—for the 2016 MCAS mathematics tests in grades 3–8.

Table 3-8. 2016 MCAS: Mathematics Common Point Distribution by Domain, Grades 3–5

Domain	Percent of Raw Score Points at Each Grade		
	3	*4*	*5*
Operations and Algebraic Thinking	33	25	20
Number and Operations in Base Ten	15	20	25
Number and Operations – Fractions	15	20	25
Geometry	12	15	10
Measurement and Data	25	20	20
Total	100	100	100

Table 3-9. 2016 MCAS: Mathematics Common Point Distribution by Domain, Grades 6 and 7

Domain	Percent of Raw Score Points at Each Grade	
	6	*7*
Ratios and Proportional Relationships	19	20
The Number System	18	22
Expressions and Equations	30	20
Geometry	15	20
Statistics and Probability	18	18
Total	100	100

Table 3-10. 2016 MCAS: Mathematics Common Point Distribution by Domain, Grade 8

Domain	Percent of Raw Score Points
The Number System	5
Expressions and Equations	30
Functions	25
Geometry	30
Statistics and Probability	10
Total	100

Table 3-11 shows the distribution of common item points in the grade 10 mathematics test across the strands of the *2000 Massachusetts Mathematics Framework*. Table 3-12 represents the distribution of common points for the same test using reporting categories that are based on the conceptual categories in the *2011 Massachusetts Mathematics Framework*. The difference between the two frameworks is that the 2000 category of Measurement is distributed among the 2011 Geometry and Statistics and Probability strands.

Table 3-11. 2016 MCAS: Mathematics Common Point Distribution by *2000 Mathematics Framework* Strand, Grade 10

Framework Strand	Percent of Raw Score Points
Number Sense and Operations	20
Patterns, Relations, and Algebra	30
Geometry	15
Measurement	15
Data Analysis, Statistics, and Probability	20
Total	100

Table 3-12. 2016 MCAS: Mathematics Common Point Distribution by Reporting Category, Grade 10*

Reporting Category	Percent of Raw Score Points
Number and Quantity	20
Algebra and Functions	30
Geometry	30
Statistics and Probability	20
Total	100

* Reporting categories are based on conceptual categories. Only content in the 2011 standards that matches content in the 2000 standards was assessed.

3.2.3.5 Cognitive Levels

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

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

Cognitive Levels I and II are represented by items in all grades. Level III is best represented by open-response items. An attempt is made to include cognitive Level III items at each grade.

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

The second session of the grades 7, 8, and 10 mathematics tests is a calculator session. All items included in this session are either calculator neutral (calculators are permitted but not required to answer the question) or calculator active (students are expected to use a calculator to answer the question). Each student taking the mathematics test at grade 7, 8, or 10 had access during Session 2 to a calculator with at least four functions and a square root key.

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

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

Rulers are provided to students in grades 3–8.

3.2.4 STE Test Specifications

3.2.4.1 Standards

Grades 5 and 8

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

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

High School

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

physics, and technology/engineering). Students who pass one MCAS STE assessment may not take other MCAS STE assessments.

The high school STE tests measure the learning standards of the strands listed in Tables 3-16 through 3-19.

3.2.4.2 Item Types

The STE tests include multiple-choice and open-response items. Open-response items are more complex, requiring 5–10 minutes of response time. Each type of item is worth a specific number of points in the student's total test score, as shown in Table 3-13.

Table 3-13. 2016 MCAS: STE Item Types and Score Points

Item Type	Possible Raw Score Points
Multiple-choice	0 or 1
Open-response	0, 1, 2, 3, or 4

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

3.2.4.3 Test Design

The STE tests comprise common and matrix items. Each form includes the full complement of common items, which is taken by all students, and a set of matrix items. Table 3-14 lists the distribution of common and matrix items in each STE test.

Table 3-14. 2016 MCAS: Distribution of STE Common and Matrix Items by Grade and Item Type

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

[a] Field-tested items are repeated in multiple forms so there are generally more field-test slots available than there are unique field-tested items.
[b] The common items on each high school biology form include a module consisting of four multiple-choice items and one open-response item that are included in the overall counts.
[c] High school STE tests are pre-equated; therefore, the entire set of matrix slots is available for field-testing.
[d] High school biology matrix items may include one matrix module per form consisting of four multiple-choice items and one open-response item. These are included in the overall matrix counts. If a module is not field-tested in a specific form, the spaces are used for standalone items.

3.2.4.4 Blueprints

Grades 5 and 8

Table 3-15 shows the distribution of common items across the four strands of the *2006 Massachusetts Science and Technology/Engineering Curriculum Framework*.

Table 3-15. 2016 MCAS: STE Common Point Distribution by Strand and Grade

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

High School

Tables 3-16 through 3-19 show the distribution of common items across the various content strands for the MCAS high school STE tests.

Table 3-16. 2016 MCAS: High School Biology Common Point Distribution by Strand

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

Table 3-17. 2016 MCAS: High School Chemistry Common Point Distribution by Strand

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

Table 3-18. 2016 MCAS: High School Introductory Physics Common Point Distribution by Strand

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

Table 3-19. 2016 MCAS: High School Technology/Engineering Common Point Distribution by Strand

MCAS Reporting Category	Percent of Raw Score Points	Related Framework Strand(s)
Engineering Design	20	▪ Engineering Design
Constructions 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 an STE test is assigned a cognitive skill according to the cognitive demand of the item. Cognitive skills are not synonymous with difficulty. The cognitive skill describes each item based on the complexity of the mental processing a student must use to answer the item correctly. Only one cognitive skill is designated for each 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 3-20 describes the cognitive skills used for the STE test items.

Table 3-20. 2016 MCAS: STE Cognitive Skill Descriptions

Cognitive Skill	Description
Remembering	▪ **Identify or define a basic concept or term with little or no context** ▪ Recall facts with little or no context *Does the item require recalling or remembering facts or definitions?*
Understanding	▪ **Describe, explain, or identify typical classroom examples for a science or technology/engineering concept** ▪ Recognize and differentiate representations and descriptions of familiar models *Does the item require the recognition or a description of a familiar concept?*
Applying	▪ **Describe, explain, or identify a science or technology/engineering concept presented in a novel situation** ▪ Draw conclusions by comparing and contrasting information in novel situations ▪ Draw conclusions by interpreting information/data (including simple graphs and tables) or make predictions based on data ▪ Solve quantitative problems where an equation must be rearranged to solve the problem ▪ Describe or explain multiple processes or system components in a novel situation *Does the item require drawing conclusions based on novel information or solving complex problems?*
Analyzing	▪ **Critically examine and interpret data** or maps to draw conclusions based on given information (Note: An item with a graph/diagram/table/map does not necessarily require the skill of analyzing—it depends on how the information needs to be interpreted.) *Does the item require critical examination of information to make conclusions?*
Creating	▪ **Generate** an explanation or conclusion by combining **two or more science or technology/engineering concepts** in a novel situation ▪ **Construct** models, graphs, charts, drawings, or diagrams **and generate explanations** or conclusions based on the information ▪ Propose solutions to scientific or engineering problems based on given criteria/constraints *Does the item require the synthesis of different concepts or skills to generate a solution?*

3.2.4.6 Use of Calculators, Formula Sheets, and Rulers

Formula sheets are provided to students taking the high school chemistry, introductory physics, and technology/engineering tests. These sheets contain reference information that students may need to answer certain test items. Students taking the chemistry test also receive a copy of the Periodic Table of the Elements to refer to 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, although the biology test was designed to be taken without the aid of a calculator.

3.2.5 Test Development Process

Table 3-21 details the test development process in chronological order.

Table 3-21. 2016 MCAS: Overview of Test Development Process

Development Step	Detail of the Process
Select reading passages (for ELA only)	Contractor's test developers find potential passages and present them to ESE test developers for initial approval; ESE-approved passages go to Assessment Development Committees (ADCs), comprised of experienced educators, and then to a Bias and Sensitivity Review Committee (Bias) for review and recommendations. ELA items are not developed until the passages have been reviewed by an ADC and Bias. With the ADC and Bias recommendations, the ESE makes the final determination as to which passages to use.
Develop items	Contractor's test developers develop draft items in ELA, mathematics, and STE aligned to specific Massachusetts standards.
ESE and educator review of items	1. Contractor sends draft items to ESE test developers for review. 2. ESE test developers review and edit items prior to presenting the items to ADCs. 3. ADCs review items and make recommendations. 4. Bias committee reviews items and makes recommendations. 5. ESE test developers make final decisions based on recommendations from ADCs and Bias.
Expert review of items	Experts from higher education and practitioners review all field-tested items for content accuracy. Each item is reviewed by at least two independent expert reviewers.
Benchmark open-response items and compositions	ESE and Contractor test developers meet to determine appropriate benchmark papers for training of scorers of field-tested open-response items and compositions. Scoring rubrics and notes are reviewed and edited during benchmarking meetings. During the scoring of field-tested items, Contractor will contact ESE test developers with any unforeseen issues.
Item statistics meeting	ADCs review field-test statistics and recommend items for the common-eligible status, for re-field-testing (with edits), or for rejection. Bias also reviews items with elevated differential item functioning (DIF) statistics and recommends to accept items to become common-eligible or to reject items.
Test construction	Before test construction, ESE provides target performance-level cut scores to the developers. Contractor proposes sets of common items (items that count toward student scores) and matrix items. Matrix items consist of field-test and equating items, which do not count toward student scores. Sets are sent by Contractor to ESE test developers. The common set of items is delivered with proposed cut scores, including TCCs and TIFs. ESE test developers and editorial staff review and edit proposed sets of items. Contractor and ESE test developers and editorial staff meet to review edits and changes to tests. Psychometricians are available to provide statistical information for changes to the common form.
Operational test items	Items become part of the common item set and are used to determine individual student scores.
Released items	Approximately 50% of the common items in grades 3–8 are released to the public, and the remaining items return to the common-eligible pools. One hundred percent of the common items are released from the spring Grade 10 ELA and mathematics tests and the high school biology and introductory physics tests. Common items from the high school chemistry and technology/engineering tests and the November and March high school mathematics and ELA retests are not released.

3.2.5.1 ELA Passage Selection and Item Development

All items used on the MCAS tests are developed specifically for Massachusetts and are directly linked to the Massachusetts 2011 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 3.2.2 for specific content standard alignment. Content not found in the curriculum frameworks is not subject to the statewide assessment.

ELA Reading Passages

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

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

ESE Passage Review

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

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

Passages that are approved by the ESE are presented to the ADCs as well as the Bias and Sensitivity Committee for review and approval. The ESE reviews all committee comments and recommendations and gives final approval to passages. Development of items with corresponding passages does not begin until the ESE has approved the passages.

ADC Passage Review

Each grade and content area has its own ADC that comprises between 10 and 12 educators from across the state who teach that content or that grade, in the case of elementary grades. ELA ADCs review ELA 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, that passage is likely to provide an unfair advantage to those students who are familiar with the work. Committee members choose one of the following recommendations for each new passage:

- accept

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

For passages recommended for acceptance, committee members provide suggestions for items that could be written. 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.

Bias and Sensitivity Committee Passage Review

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

Item Development and Review

ESE Item Review

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

- Alignment: Are the items aligned to the standards? Is there a better standard to which to align the item?
- Content: Does the item show a depth of understanding of the subject?
- Contexts: Are contexts used when appropriate? Are they realistic?
- Grade-level appropriateness: Are the content, language, and contexts appropriate for the grade level?
- Creativity: Does the item demonstrate creativity with regard to approaches to items and contexts?
- Distractors: Have the distractors for multiple-choice items been chosen based on common sources of error? Are they plausible?
- Mechanics: How well are the items written? Do they follow the conventions of item writing?
- Missed opportunities (for reading comprehension only): Were there items that should have been written based on the passage but were not written?

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

ADC Item Review

Once the ESE has reviewed new items and scoring guides and any requested changes have been made, the materials are submitted to ADCs for further review. Committees review new items for the characteristics listed on the previous page and provide insight into how standards are interpreted

across the state. Committees choose one of the following recommendations regarding each new item:

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

All ADC committee recommendations remain with the item in the comment field of the item card.

Bias and Sensitivity Committee Item Review

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

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

All Bias and Sensitivity Committee review comments are kept in the comment field of the item card. The comments stay with the item. The ESE disposition of its recommendation is also kept in the comment field and stays with the item.

Once the Bias and Sensitivity Review Committee has made its recommendations and the ESE has determined whether to act on the recommendations, ESE-approved items become "field-test eligible" and move to the next step in the development process.

External Content Expert Item Review

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

3.2.5.2 Item Editing

ESE content specialists review the recommendations of the expert reviewers and item committees and determine whether or not to accept the suggested edits. The items are also reviewed and edited by Measured Progress editors to ensure adherence to style guidelines in *The Chicago Manual of*

Style, to MCAS-specific style guidelines, and to sound testing principles. According to these principles, all 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; and
- exhibit high technical quality regarding psychometric characteristics.

3.2.5.3 Field-Testing Items

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

3.2.5.4 Scoring of Field-Tested Items

Each field-tested multiple-choice item is machine-scored. Each constructed-response item (short-answer, short-response, or open-response) is hand-scored. In order to train scorers, the ESE works closely with the scoring staff to refine the rubrics and scoring notes and to select benchmark papers that exemplify the score points and the variations within each score point. Approximately 1,800 student responses are scored per constructed-response field-tested item. As with the multiple-choice items, 1,800 student responses are sufficient to provide reliable results. See section 3.4 for additional information on scorers and scoring.

3.2.5.5 Data Review of Field-Tested Items

Data Review by the ESE

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

Data Review by ADCs

The ADCs meet to review the items with their field-test statistics. ADCs consider the following when reviewing field-test item statistics:

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

The ADCs make one of the following recommendations regarding each field-tested item:

- accept

- edit and field-test again (This is for mathematics and STE items only. Because ELA items are passage-based, items cannot be field-tested again individually. To address this matter, more than twice the number of items needed for the test are field-tested in ELA.)
- 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, the newly edited item returns to the field-test-eligible pool to be field-tested again.

Data Review by the Bias and Sensitivity Review Committee

The Bias and Sensitivity Review Committee also reviews the statistics for the field-tested items. The committee reviews only the items that the ADCs have accepted. The Bias and Sensitivity Review Committee pays special attention to items that show DIF when comparing the following subgroups of test-takers:

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

The Bias and Sensitivity Review Committee considers whether DIF seen in items is a result of item bias or is the result of uneven access to curriculum and makes recommendations to the ESE regarding the disposition of items based on the committee's item statistics. The ESE makes the final decision regarding the Bias and Sensitivity Committee recommendations.

3.2.5.6 Item Selection and Operational Test Assembly

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

- **Content coverage/match to test design and blueprints.** The test designs and blueprints stipulate a specific number of items per item type for each content area. Item selection for the embedded field test is based on the depth of items in the existing pool of items that are eligible for the common portion of the test. Should a certain standard have few items aligned to it, then more items aligned to that standard will be field-tested to ensure a range of items aligned to that standard are available for use.
- **Item difficulty and complexity.** Item statistics drawn from the data analysis of previously field-tested items are used to ensure similar levels of difficulty and complexity from year to year as well as high-quality psychometric characteristics. Since 2011, items can be reused if they have not been released. When an item is reused in the common portion of the test, the latest usage statistics accompany that item.
- **"Cueing" items.** Items are reviewed for any information that might "cue" or help the students answer another item.

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

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

3.2.5.7 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.

3.2.5.8 Special Edition Test Forms

Students With Disabilities

MCAS is accessible to students with disabilities through the provision of special edition test forms and a range of accommodations for students taking the standard tests. To be eligible to receive a special edition test form, a student needs to have a disability that is documented in an individualized education plan (IEP) or a 504 plan, or needs to have a 504 plan in development. All MCAS 2016 operational tests and retests were available in the following special editions for students with disabilities:

- Large-print – Form 1 of the operational test is translated into a large-print edition. The large-print edition contains all common and matrix items found in Form 1.

- Braille – This form includes only the common items found in the operational test. If an item shows bias against blind students (e.g., if it includes a complex graphic that a student taking the Braille test could not reasonably be expected to comprehend as rendered), then simplification of the graphic is considered, with appropriate rewording of the item text, as necessary. If a graphic such as a photograph cannot be rendered in Braille, or if the graphic is not needed for the student to respond to the item, the graphic is replaced with descriptive text or a caption, or eliminated altogether. Three-dimensional shapes that are needed by the

student to respond, and that are rendered in two dimensions in print, are rendered on the Braille test as "front view," "top view," and/or "side view," and are accompanied where necessary by a three-dimensional wooden or plastic manipulative wrapped in a Braille-labeled plastic bag.

Modifications to original test items for the Braille version of the test are made only when necessary, as determined by the Braille test subcontractor, blind consumers, and Department staff, and only when they do not provide clues or assistance to the student or change what the item is measuring. When successful modification of an item or graphic is not possible, all or part of the item is omitted, and may be replaced with a similar item.

- Electronic text reader CD – This CD, in Kurzweil format, contains only the common items found in the operational test. For items or passages that have graphics, captions and words in the graphics are read aloud to the student as they are seen in text. The items are not modified and are read aloud to the student as they are seen in the standard test booklet.

In addition, the grade 10 MCAS mathematics test is available to students who are deaf or hard-of-hearing in an American Sign Language DVD edition, which contains only the common items found in the operational test.

Appendix D details student accommodations that do not require a special test form. Students who have an IEP or are on a 504 plan are eligible to take the MCAS standard operational tests with these accommodations. After testing is completed, the ESE receives a list that includes the number of students who participated in MCAS with each accommodation. No identifying information is provided.

Spanish-Speaking Students

Spanish/English editions of the spring grade 10 mathematics test and the March and November mathematics retests were available for Spanish-speaking ELL students who had been enrolled in school in the continental United States for fewer than three years and could read and write in Spanish at or near grade level. The Spanish/English edition of the spring grade 10 mathematics test contains all common and matrix items found in Form 1 of the operational test.

Measured Progress employs two independent translators to complete the translation of the grade 10 mathematics test and the mathematics retests to Spanish. The translation process is as follows:

- A set of translation rules or parameters is generated taking the following into consideration: vocabulary, usage, and consistency over the years. These rules are provided to both translators.
- The first translator translates from English to Spanish. The second translator proofs the work of the first translator.
- Discrepancies between the two translations are resolved by a third party.
- The Publishing Department reviews the graphics in English and Spanish to ensure that they are consistent.
- The Spanish version is always on the left-hand page with the English version always on the right-hand page. Students taking the Spanish version of a mathematics test always have the English translation as part of their test.

- The script that the teacher reads when administering the test is also translated into Spanish while the *Test Administrator's Manual* is in English and Spanish.
- The translated test undergoes a publication and linguistics review at the ESE.

The Spanish/English editions of the grade 10 mathematics test and the mathematics retests are not available in any other special format.

3.3 Test Administration

3.3.1 Test Administration Schedule

The standard MCAS tests were administered during three periods in spring 2016:

- March–April
  - Grades 3–8 and 10 ELA
- May
  - Grades 3–8 and 10 mathematics
  - Grades 5 and 8 STE
- June
  - High school (grades 9–12) end-of-course STE
    - biology
    - chemistry
    - introductory physics
    - technology/engineering

The 2016 MCAS administration also included retest opportunities in ELA and mathematics for students in grades 11 and 12, and students who had exited high school and who had not previously passed one or both grade 10 tests. Retests were offered in November 2015 and March 2016.

An additional high school (grades 9–12) end-of-course STE biology test was administered in February 2016.

Table 3-22 shows the complete 2015–2016 MCAS test administration schedule.

Table 3-22. 2016 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, 2015		
ELA Composition Retest	November 4	November 17
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 2–8, 2016		
ELA Composition Retest	March 2	March 11

Grade and Content Area	Test Administration Date(s)	Deadline for Return of Materials to Contractor
ELA Reading Comprehension Retest Sessions 1 and 2 Session 3	 March 3 March 4	March 11
Mathematics Retest Session 1 Session 2	 March 7 March 8	
March–April 2016 Test Administration Window		
Grades 3–8 ELA	March 28–April 12	Grades 3–8: April 14 Grade 10: April 7
Grade 10 ELA Composition	March 22	
Grade 10 ELA Reading Comprehension Sessions 1 and 2 Session 3	 March 23 March 24	
Grade 10 ELA Composition Make-Up	March 31	
May 2016 Test Administration Window		
Grades 3–8 Mathematics	May 9–May 24	May 26
Grades 5 and 8 STE	May 10–May 24	
Grade 10 Mathematics Session 1 Session 2	 May 17 May 18	
High School (Grades 9–12) End-of-Course STE Test Administration Windows		
February 1–2, 2016		
Biology	February 1–February 2	February 8
June 1–2, 2016		
Biology	June 1–June 2	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, guidance for establishing and following a document tracking system, and lists of approved and unapproved resource materials, can be found in

the *Spring 2016 Principal's Administration Manual*, the *Fall 2015/Winter 2016 Principal's Administration Manual*, and all *Test Administrator's Manuals*.

3.3.3 Participation Requirements

In spring 2016, students educated with Massachusetts public funds were required by state and federal laws to participate in MCAS testing (districts were given the choice of administering either MCAS or PARCC assessments in ELA and mathematics in grades 3–8). 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 innovation schools
- students enrolled in a Commonwealth of Massachusetts Virtual School
- students enrolled in educational collaboratives
- students enrolled in private schools receiving special education that is publicly funded by the Commonwealth, including approved and unapproved private special education schools within and outside Massachusetts
- students enrolled in institutional settings receiving educational services
- students in mobile military families
- students in the custody of either the Department of Children and Families (DCF) or the Department of Youth Services (DYS)
- students with disabilities, including students with temporary disabilities such as a broken arm
- ELL students
- students who have been expelled but receive educational services from a district
- foreign exchange students who are coded as #11 under "Reason for Enrollment" in the Student Information Management System (SIMS)

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

3.3.3.1 Students Not Tested on Standard Tests

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

- ELL students in their first year of enrollment in U.S. schools, who are not required to participate in ELA testing
- students with significant disabilities who must instead participate in the MCAS-Alt (see Chapter 4 for more information)
- students with a medically documented absence who are unable to participate in make-up testing, including students participating in post-concussion "graduated reentry" plans who were determined to be not well enough for standard MCAS testing
- students in military families who enrolled in a Massachusetts school in grade 11 or later (the district may, in lieu of having the student participate in MCAS retests, submit to the

Department alternative evidence or information that demonstrates that the student has met the CD graduation standard in each required content area)

More details about test administration policies and student participation requirements at all grade levels, including requirements for earning a CD, requirements for students with disabilities or students who are ELLs, and students educated in alternate settings, can be found in the *Spring 2016 Principal's Administration Manual* and the *Fall 2015/Winter 2016 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 and test administration protocols
- reviewing plans for maintaining test security with the superintendent
- 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 during 2016 testing can be found in Appendix C. For a list of test accommodations, see Appendix D.)
- completing and ensuring the accuracy of information provided on the PCPA
- monitoring the ESE's website (www.doe.mass.edu/mcas) throughout the school year for important updates
- providing the ESE with correct contact information to receive important notices via fax and e-mail during test administration

More details about test administration procedures, including ordering test materials, scheduling test administration, designating and training qualified test administrators, identifying testing spaces, meeting with students, providing accurate student information, and accounting for and returning test materials, can be found in the *Spring 2016 Principal's Administration Manual* and the *Fall 2015/Winter 2016 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 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 secure server-to-server interface designed by Measured Progress.

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

3.4.1 Machine-Scored Items

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

3.4.2 Hand-Scored Items

Once responses to constructed-response items were sorted into item-specific groups, they were scored one item at a time. Scorers within each group scored one response at a time. However, if there was a need to see a student's responses across all of the constructed-response items, scoring leadership had access to the student's entire answer booklet. Details on the procedures used to hand-score student responses are provided below.

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 3-23.

Table 3-23. 2016 MCAS: Summary of Scoring Locations and Scoring Shifts

Measured Progress Scoring Center, Content Area	Grade(s)	Shift	Hours
Longmont, CO			
ELA Reading Comprehension	3, 4, 5, 6, 7	Night	5:30 p.m.–10:30 p.m.
	8, 10	Day	8:00 a.m.–4:30 p.m.
Mathematics	3, 4, 5, 6, 7	Night	5:30 p.m.–10:30 p.m.
	8, 10	Day	8:00 a.m.–4:30 p.m.
ELA Writing/Composition	3, 4, 5, 6, 7,8	Day	8:00 a.m.–4:30 p.m.
Dover, NH			
STE	5	Night	5:30 p.m.–10:00 p.m.
Menands, NY			
ELA Composition		Day	8:00 a.m.–4:30 p.m.
	10	Night	5:30 p.m.–10:30 p.m.
STE	5	Night	5:30 p.m.–10:30 p.m.
	8	Day	8:00 a.m.–4:30 p.m.
STE: Biology	HS	Day	8:00 a.m.–4:30 p.m.
STE: Introductory Physics	HS	Night	5:30 p.m.–10:30 p.m.
STE: Chemistry	HS	Night	5:30 p.m.–10:30 p.m.
STE: Technology/Engineering	HS	Night	5:30 p.m.–10:30 p.m.

The following staff members were involved with scoring the 2016 MCAS responses:

- The **MCAS Scoring Project Manager (SPM)** was located in Dover, New Hampshire, and oversaw communication and coordination of MCAS scoring across all scoring sites.
- The **iScore Operations Manager** was located in Dover, New Hampshire, and coordinated technical communication across all scoring sites.

- A **Scoring Center Manager (SCM)** was located at each satellite scoring location and provided logistical coordination for his or her scoring site.
- A **Scoring Content Specialist** 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. Scoring Content Specialists monitored and read behind on-site and off-site Scoring Supervisors.
- Several **Scoring Supervisors,** selected from a pool of experienced **Scoring Team Leaders (STLs)**, participated in benchmarking, training, scoring, and cleanup activities for specified content areas and grade levels. Scoring Supervisors monitored and read behind STLs.
- **STLs,** selected from a pool of skilled and experienced scorers, monitored and read behind **scorers** at their scoring tables. STLs generally monitored 5 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 Measured Progress's Scoring Services Department and Content, Design & Development (CDD) Department as well as ESE staff members 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; and
- approve various individual responses and sets of responses (e.g., anchor, training) to be used to train field-test scorers.

3.4.2.3 Scorer Recruitment and Qualifications

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

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

Table 3-24 is a summary of scorers' backgrounds across all scoring shifts at all scoring locations.

Table 3-24. 2016 MCAS: Summary of Scorers' Backgrounds Across Scoring Shifts and Scoring Locations

Education	Scorers		Leadership	
	Number	*Percent*	*Number*	*Percent*
Less than 48 college credits	0	0	0	0
Associate's degree/more than 48 college credits	87	12.63	8	6.40
Bachelor's degree	370	53.70	57	45.60
Master's degree/doctorate	232	33.67	60	48.00
Teaching Experience				
No teaching certificate or experience	353	51.23	56	44.80
Teaching certificate or experience	273	39.62	54	43.20
College instructor	63	9.14	15	12.00
Scoring Experience				
No previous experience as scorer	192	27.87	3	2.40
1–3 years of experience	242	35.12	29	23.20
3+ years of experience	255	37.01	93	74.40

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 (mathematics only), with assigned scores of 0–1; short-response items (grade 3 ELA only), with assigned scores of 0–2; open-response items requiring a longer or more complex response, with assigned scores of 0–4 or 0–2 (grade 3 mathematics only); and the writing prompt for the ELA composition, with assigned scores of 1–4 and 1–6.

The sample below (Table 3-25) of a 4-point mathematics open-response scoring guide was one of the many different item-specific MCAS scoring guides used in 2016. The task associated with this scoring guide required students to design four different gardens, each with a different shape.

Table 3-25. 2016 MCAS: Four-Point Open-Response Item Scoring Guide – Grade 10 Mathematics

Score	Description
4	The student response demonstrates an exemplary understanding of the Statistics and Probability concepts involved in representing data on two quantitative variables on a scatter plot, and describing how the variables are related. The student interprets a scatter plot, finds and compares measures of center, and identifies a relationship between the variables.
3	The student response demonstrates a good understanding of the Statistics and Probability concepts involved in representing data on two quantitative variables on a scatter plot, and describing how the variables are related. 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 a fair understanding of the Statistics and Probability concepts involved in representing data on two quantitative variables on a scatter plot, and describing how the variables are related. 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 a minimal understanding of the Statistics and Probability concepts involved in representing data on two quantitative variables on a scatter plot, and describing how the variables are related.
0	The student response contains insufficient evidence of an understanding of the Statistics and Probability concepts involved in representing data on two quantitative variables on a scatter plot, and describing how the variables are related to merit any points.

Scorers 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.
- **Unreadable:** The text on the scorer's computer screen is too faint to see accurately.
- **Wrong Location:** The response seems to be a legitimate answer to a different question.

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

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

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

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

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

All polytomous items on all high school tests (ELA, mathematics, and STE) were 100% double-blind scored. Ten percent of polytomous items on the ELA reading comprehension and mathematics tests at grades 3–8, and on the grades 5 and 8 STE tests, were double-blind scored. In addition, grades 3–8 had PARCC-based writing items added to the test this year. These items were scored at a 10% double-blind rate.

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

Table 3-26 outlines the rules for instances when the two read-behind or two double-blind scores were not identical (i.e., adjacent or discrepant).

Table 3-26. 2016 MCAS: Read-Behind and Double-Blind Resolution Charts

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

Double-Blind Scoring*, 4-Point Item			
Scorer #1	*Scorer #2*	*Scoring Leadership Resolution*	*Final*
4	4	-	4
4	3	-	4
3	4	-	4
4	2	3	3
4	1	2	2
3	1	1	1
* If double-blind scores are adjacent, the higher score is used as the final score. If scorer scores are neither identical nor adjacent, the resolution score is used as the final score.			

Writing Standard English Conventions Double-Blind Scoring*			
Scorer #1	*Scorer #2*	*Scoring Leadership 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 scorer scores are summed to obtain the final score. The resolution score, if needed, is summed with an identical scorer score; or, if the resolution score is adjacent to scorer #1 and/or #2 but not identical with either, then the two highest adjacent scores are summed for the final score.			

Writing Topic Development Double-Blind Scoring*				
Scorer #1	*Scorer #2*	*Scoring Leadership Resolution*	*Scoring Content Specialist*	*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 scorer scores are summed to obtain the final score. The resolution score, if needed, is summed with an identical scorer score; or, if the resolution score is adjacent to scorer #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 Scoring Content Specialist assigns a fourth score, which is doubled to obtain the final score.				

3.4.2.5 Scorer Training

Scoring Content Specialists had overall responsibility for ensuring that scorers scored responses consistently, fairly, and according to the approved scoring guidelines. Scoring materials were carefully compiled and checked for consistency and accuracy. The timing, order, and manner in which the materials were presented to scorers 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.

Measured Progress uses a range of methods to train scorers to score MCAS constructed-response items. The five training methods are as follows:

- live face-to-face training in small groups
- live face-to-face training of multiple subgroups in one large area
- audio/video conferencing
- live large-group training via headsets (WebEx)
- recorded modules (used for individuals, small groups, or large groups)

Some training was conducted remotely. Scorers were trained on some items via computers connected to a remote location; that is, the trainer was sitting at a computer in one scoring center, and the scorers were sitting at their computers at a different scoring center. Interaction between scorers and trainers remained uninterrupted through instant messaging or two-way audio communication devices, or through the on-site scoring supervisors.

Scorers started the training process by receiving 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. Scorer 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 scorers carefully reviewed three different sets of actual student responses, some of which had been used to train scorers when the item was a field-test item:

- **Anchor sets** are ESE-approved sets consisting of two to three sample responses at each score point. Each response is typical, rather than unusual or uncommon; solid, rather than controversial; and true, meaning that these responses have scores that cannot be changed.
- **Practice sets** include unusual, discussion-provoking responses, illustrating the range of responses encountered in operational scoring (e.g., exceptionally creative approaches; extremely short or disorganized responses; responses that demonstrate attributes of both higher-score anchor papers and lower-score anchor papers or that show traits of multiple score points).
- **Qualifying sets** consist of 10 responses that are clear, typical examples of each of the score points. Qualifying sets are used to determine if scorers are able to score consistently according to the ESE-approved scoring rubric.
- Some items on the spring 2016 test were PARCC-developed items that were included on the MCAS test. Training materials for these items were provided by PARCC. While the anchor and practice sets were ESE reviewed, they were not ESE approved. For comparability, the training materials were used based on the scoring and rationale provided by PARCC, and were not altered by Measured Progress or the ESE.

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

3.4.2.6 Leadership Training

Scoring Content Specialists also had overall responsibility for ensuring that scoring leadership (Scoring Supervisors and STLs) continued their history of scoring consistently, fairly, and only according to the approved scoring guidelines. Once they have completed their item-specific leadership training, scoring leadership must have met or surpassed a qualification standard of at least 80% exact and 90% exact plus adjacent, or, for grade 10 leadership, at least 80% exact and 100% adjacent.

3.4.2.7 Monitoring of Scoring Quality Control

Once MCAS scorers 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 ensure they scored student responses as accurately and consistently as possible. If a scorer fell below the minimum standard on any of the quality-control tools, there was some form of scorer intervention, ranging from counseling to retraining to dismissal. Scorers were required to meet or exceed the minimum standard of 70% exact and 90% exact plus adjacent agreement on the following:

- recalibration assessments (Recals)
- embedded responses
- read-behind scoring (RBs)
- double-blind scoring (DBs)
- compilation reports, an end-of-shift report combining recalibration sets and RBs

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

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

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

DBs involved responses scored independently by two different scorers. Scorers 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. Scorers who fell below the 70% exact and 90% exact plus adjacent score match standard during the scoring shift were counseled, given extra monitoring assignments such as additional RBs, and likely allowed to resume scoring. Responses given discrepant scores by two independent scorers were read and scored by an STL.

Compilation reports combined a reader's percentage of exact, adjacent, and discrepant scores on the Recals with that scorer's percentage of exact, adjacent, and discrepant scores on the RBs. As the STL conducted RBs, the scorers' overall percentages on the compilation reports were automatically calculated and updated. If the compilation report at the end of the scoring shift listed individuals who were still below the 70% exact and 90% exact plus adjacent level, their scores for that day were voided. Responses with scores voided were returned to the scoring queue for other scorers to score.

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

3.4.2.8 Interrater Consistency

As described above, double-blind scoring was one of the processes used to monitor the quality of the hand-scoring of student responses for constructed-response items. All of the open-response and

composition items were double-scored on the high school test; for all other open-response items, 10% of student responses were randomly selected and scored independently by two different scorers. Results of the double-blind scoring were used during the scoring process to identify scorers who required retraining or other intervention, and they are presented here as evidence of the reliability of the MCAS tests. A summary of the interrater consistency results is presented in Table 3-27. Results in the table are organized across the hand-scored items by content area and grade. The table shows the number of score categories, the number of included scores, the percent exact agreement, percent adjacent agreement, correlation between the first two sets of scores, and the percent of responses that required a third score. This same information is provided at the item level in Appendix O. These interrater consistency statistics are the result of the processes implemented to ensure valid and reliable hand-scoring of items.

Table 3-27. 2016 MCAS: Summary of Interrater Consistency Statistics Organized Across Items by Content Area and Grade

Content Area	Grade	Number of		Percent*		Correlation	Percent of Third Scores
		Score Categories	*Included Scores*	*Exact*	*Adjacent*		
ELA	3	3	7,897	75.29	24.05	0.68	0.62
		5	1,966	67.85	31.08	0.73	0.97
	4	5	7,790	62.37	35.55	0.72	2.04
	5	5	7,820	61.89	35.91	0.75	2.14
	6	5	8,231	60.92	36.98	0.79	1.98
	7	5	7,876	61.10	36.96	0.73	1.82
	8	5	7,882	62.93	34.59	0.75	2.50
	10	4	64,803	76.89	22.72	0.65	0.83
		5	267,197	63.70	35.11	0.74	1.19
		6	64,803	69.93	29.47	0.67	0.83
Mathematics	3	2	11,787	98.46	1.54	0.96	0.00
		3	7,863	94.29	5.46	0.94	0.25
	4	2	11,669	98.33	1.67	0.96	0.00
		5	7,842	81.32	17.51	0.93	1.17
	5	2	11,851	97.79	2.21	0.94	0.00
		5	7,905	84.06	14.12	0.94	1.82
	6	2	12,352	98.38	1.62	0.97	0.00
		5	8,213	80.82	16.94	0.93	2.24
	7	2	11,984	98.78	1.22	0.96	0.00
		5	7,971	81.08	17.50	0.94	1.42
	8	2	11,928	98.95	1.05	0.97	0.00
		5	7,914	84.38	14.64	0.94	0.97
	10	2	272,627	98.50	1.50	0.97	0.00
		5	406,910	84.68	14.46	0.94	0.86
STE	5	5	27,371	70.17	26.44	0.89	3.40
	8	5	27,889	67.54	28.80	0.87	3.68
Biology	9–12	5	245,061	73.49	24.43	0.90	2.09
Chemistry	9–12	5	4,012	71.98	24.58	0.88	3.44
Introductory Physics	9–12	5	73,858	84.17	14.85	0.92	0.97
Technology/ Engineering	9–12	5	12,963	69.98	27.21	0.84	2.81

*Values may not equal 100% due to rounding.

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* (AERA et al., 2014) 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) DIF statistics, and (4) dimensionality analyses. The item analyses presented here are based on the statewide administration of the MCAS in spring 2016.[1] Note that the information presented in this section is based on the items common to all forms, since those are the items on which student scores are calculated. (Item analyses are also performed for field-test items, and the statistics are then used during the item review process and form assembly for future administrations.)

3.5.1 Classical Difficulty and Discrimination Indices

All multiple-choice and open-response items are evaluated in terms of item difficulty according to standard classical test theory practices. Difficulty is defined as the average proportion of points achieved on an item and is measured by obtaining the average score on an item and dividing it by the maximum possible score for the item. Multiple-choice items are scored dichotomously (correct vs. incorrect), so, for these items, the difficulty index is simply the proportion of students who correctly answered the item. Open-response items are scored polytomously, meaning that a student can achieve scores other than just 0 or 1 (e.g., 0, 1, 2, 3, or 4 for a 4-point open-response item). 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

[1] As described in section 1.4, the majority of students in grades 3–8 took the PARCC tests in ELA and mathematics in 2016, rather than the MCAS tests in those subjects. Statewide, 28% of students in grades 3–8 took the MCAS ELA and mathematics assessments. The statistics and analyses in this report are based on the results of the MCAS test-takers.

essentially zero for open-response items) to 0.90, with the majority of items generally falling between 0.4 and 0.7. However, on a standards-referenced assessment such as the MCAS, it may be appropriate to include some items with very low or very high item difficulty values to ensure sufficient content coverage.

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

Discrimination indices can be thought of as measures of how closely an item assesses the same knowledge and skills assessed by the other items contributing to the criterion total score on the assessment. When an item has a high discrimination index, it means that students selecting the correct response are students with higher total scores, and students selecting incorrect responses are associated with lower total scores. Given this, the item can discriminate between low-performing examinees and high-performing examinees. Very low or negative point-biserial coefficients computed after field-testing new items can help identify items that are flawed.

A summary of the item difficulty and item discrimination statistics for each grade and content area combination is presented in Table 3-28. (For grades 3–8 in ELA and mathematics, item statistics in Table 3-28 are for the MCAS portion of the tests.) Note that the statistics are presented for all items as well as by item type (multiple-choice and open-response). The mean difficulty (*p*-value) and discrimination values shown in the table are within generally acceptable and expected ranges and are consistent with results obtained in previous administrations. Note that the number of students included in the item statistics calculations for the mathematics and ELA tests at grades 3–8 was less than one-third the total number of state examinees because many school districts chose to administer the PARCC mathematics and ELA tests instead of the MCAS tests. Also, as explained in sections 1.4.3 and 1.4.4, the demographic characteristics of the students taking MCAS in 2016 were significantly different from those of previous years. Thus, comparisons to previous years should be made with caution at those grade levels.

Table 3-28. 2016 MCAS: Summary of Item Difficulty and Discrimination Statistics by Content Area and Grade

Content Area	Grade	Item Type	Number of Items	*p*-Value		Discrimination	
				Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
ELA	3	ALL	41	0.81	0.12	0.43	0.07
		MC	36	0.83	0.10	0.43	0.07
		OR	5	0.63	0.13	0.43	0.06
	4	ALL	40	0.77	0.15	0.40	0.06
		MC	36	0.81	0.12	0.39	0.05
		OR	4	0.48	0.04	0.53	0.04

continued

Content Area	Grade	Item Type	Number of Items	*p*-Value		Discrimination	
				Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
ELA	5	ALL	40	0.75	0.13	0.40	0.08
		MC	36	0.77	0.11	0.38	0.06
		OR	4	0.51	0.07	0.58	0.04
	6	ALL	40	0.78	0.12	0.41	0.09
		MC	36	0.81	0.10	0.39	0.07
		OR	4	0.56	0.09	0.62	0.04
	7	ALL	40	0.80	0.11	0.39	0.09
		MC	36	0.82	0.09	0.37	0.06
		OR	4	0.62	0.03	0.62	0.03
	8	ALL	40	0.77	0.11	0.42	0.10
		MC	36	0.79	0.09	0.39	0.07
		OR	4	0.61	0.04	0.64	0.02
	10	ALL	42	0.77	0.11	0.40	0.11
		MC	36	0.78	0.11	0.36	0.05
		OR	6	0.68	0.10	0.64	0.04
Mathematics	3	ALL	36	0.75	0.14	0.41	0.11
		MC	26	0.76	0.15	0.40	0.11
		OR	10	0.74	0.10	0.46	0.09
	4	ALL	42	0.74	0.15	0.43	0.11
		MC	32	0.77	0.16	0.39	0.08
		OR	10	0.66	0.09	0.55	0.09
	5	ALL	42	0.73	0.13	0.45	0.11
		MC	32	0.73	0.13	0.42	0.08
		OR	10	0.72	0.10	0.53	0.15
	6	ALL	42	0.72	0.17	0.45	0.11
		MC	32	0.75	0.15	0.42	0.09
		OR	10	0.63	0.18	0.55	0.13
	7	ALL	42	0.69	0.11	0.49	0.10
		MC	32	0.69	0.11	0.47	0.06
		OR	10	0.71	0.13	0.57	0.15
	8	ALL	42	0.70	0.13	0.49	0.11
		MC	32	0.71	0.12	0.45	0.08
		OR	10	0.67	0.14	0.59	0.14
	10	ALL	42	0.70	0.11	0.47	0.13
		MC	32	0.73	0.10	0.41	0.08
		OR	10	0.61	0.08	0.64	0.13
STE	5	ALL	42	0.69	0.14	0.39	0.10
		MC	38	0.71	0.12	0.37	0.08
		OR	4	0.52	0.17	0.59	0.05
	8	ALL	42	0.66	0.14	0.39	0.11
		MC	38	0.67	0.14	0.37	0.08
		OR	4	0.52	0.09	0.61	0.05
Biology	HS	ALL	45	0.73	0.13	0.42	0.11
		MC	40	0.77	0.10	0.39	0.06
		OR	5	0.46	0.06	0.69	0.02

continued

Content Area	Grade	Item Type	Number of Items	*p*-Value		Discrimination	
				Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
Chemistry	HS	ALL	45	0.67	0.14	0.45	0.11
		MC	40	0.69	0.13	0.42	0.08
		OR	5	0.48	0.12	0.69	0.06
Introductory Physics	HS	ALL	45	0.69	0.13	0.44	0.10
		MC	40	0.71	0.11	0.42	0.07
		OR	5	0.47	0.08	0.64	0.05
Technology/ Engineering	HS	ALL	45	0.65	0.14	0.37	0.11
		MC	40	0.67	0.13	0.35	0.09
		OR	5	0.45	0.09	0.54	0.11

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

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

In addition to the item difficulty and discrimination summaries presented above, these same statistics were also calculated at the item level along with item-level score point distributions. These classical statistics, item difficulty and discrimination, are provided in Appendix E for each item. On MCAS items, 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 are a small number of items with discrimination indices below 0.20, but none was negative. While it is acceptable to include items with low discrimination values or with very high or very low item difficulty values when their content is needed to ensure that the content specifications are appropriately covered, there were very few such cases on the MCAS. Item-level score point distributions are provided for open-response items in Appendix F; for each item, the percentage of students who received each score point is presented.

3.5.2 DIF

The *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004) explicitly states that subgroup differences in performance should be examined when sample sizes permit and that actions should be taken to ensure that differences in performance are attributable to construct-relevant, rather than irrelevant, factors. *Standards for Educational and Psychological Testing* (AERA et al., 2014) includes similar guidelines. As part of the effort to identify such problems, psychometricians evaluated MCAS items in terms of DIF statistics.

For the MCAS, the standardization DIF procedure (Dorans & Kulick, 1986) was employed to evaluate subgroup differences. (Subgroup differences denote significant group-level differences in performance for examinees with equivalent achievement levels on the test.) 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. However, if subgroup differences in performance can be traced to differential experience (such as geographical living conditions or access to technology), the inclusion of such items is reconsidered during the item review process.

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

For the 2016 MCAS administration, DIF analyses were conducted for all subgroups (as defined in the No Child Left Behind Act) for which the sample size was adequate. In all, six subgroup comparisons were evaluated for DIF:

- male/female
- white/black
- white/Hispanic
- no disability/disability
- not LEP-FLEP/LEP-FLEP
- not economically disadvantaged/economically disadvantaged

The tables in Appendix G present the number of items classified as either "low" or "high" DIF, in total and by group favored. Overall, a moderate number of items exhibited low DIF and several exhibited high DIF; the numbers were fairly consistent with results obtained for previous administrations of the test. Note that the number of students included in the DIF calculations for the mathematics and ELA tests at grades 3–8 was less than one-third the total number of state examinees

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

because many school districts chose to administer the PARCC mathematics and ELA tests instead of the MCAS tests. Thus, comparisons to previous years should be made with caution at these grade levels.

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 2015–16 administrations. A total of 20 tests were analyzed, and the results for these analyses are reported below, including a comparison with the results from 2014–15.

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 over all possible conditioning scores. When a test is strictly unidimensional, all conditional covariances are expected to take on values within random noise of zero, indicating statistically independent item responses for examinees with equal expected scores. 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 for pairs composed of items 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 (Roussos & Ozbek, 2006).

DIMTEST and DETECT were applied to the common items of the MCAS tests administered during spring 2016 (a total of 20 tests). The data for each grade were split into a training sample and a cross-validation sample. For mathematics and ELA, there were over 19,500 student examinees per test for all the elementary and middle school test administrations and over 67,500 students per test for the high school administration. For the science assessments, all the elementary and middle school administrations had over 68,500 students per test, while the high school administrations had over 50,500 for biology, over 15,000 for physics, over 2,500 for technology/engineering, and over 800 for chemistry. Note that the number of students who took the mathematics and ELA tests at grades 3–8 was less than one-third the total number of state examinees because many school districts chose to administer the PARCC mathematics and ELA tests instead of the MCAS tests. Because DIMTEST had an upper limit of 24,000 students, the training and cross-validation samples for the tests that had over 24,000 students were limited to 12,000 each, randomly sampled from the total sample. DETECT, on the other hand, had an upper limit of 500,000 students, so every training sample and cross-validation sample used all the available data. After randomly splitting the data into training and cross-validation samples, DIMTEST was applied to each 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 19,500 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 3-29 below displays the multidimensionality effect-size estimates from DETECT.

Table 3-29. 2016 MCAS: Multidimensionality Effect Sizes by Grade and Content Area

Content Area	Grade	Multidimensionality Effect Size	
		2015	*2016*
ELA	3	0.10	0.10
	4	0.24	0.15
	5	0.14	0.16
	6	0.13	0.13
	7	0.20	0.14
	8	0.13	0.14
	10	0.19	0.21
	Average	**0.16**	**0.15**
Mathematics	3	0.13	0.17
	4	0.15	0.10
	5	0.15	0.17
	6	0.14	0.11
	7	0.15	0.10
	8	0.18	0.10
	10	0.16	0.08
	Average	**0.15**	**0.12**
STE	5	0.09	0.13
	8	0.07	0.13
Biology	9–12	0.08	0.09
Chemistry	9–12		0.09
Introductory Physics	9–12	0.12	0.07
Technology/Engineering	9–12	0.14	0.09
	Average	**0.10**	**0.10**

The DETECT values indicate very weak to weak multidimensionality for all the tests for 2015–16. The ELA test forms (average effect size of about 0.15) and the mathematics test forms (average of about 0.12) tended to show slightly greater multidimensionality than did the science test forms (average of about 0.10). Also shown in Table 3-29 are the values reported in last year's dimensionality analyses. (The value for last year's chemistry test is omitted because the results of the 2014–15 DIMTEST analyses indicated that the null hypothesis was retained at a significance level of 0.01.) The average DETECT values in 2014–15 for ELA and mathematics were 0.16 and 0.15, respectively, and the average for the science tests was 0.10. Thus, last year's results are very similar to those from this year.

The way in which DETECT divided the tests into clusters was also investigated to determine whether there were any discernable patterns with respect to the multiple-choice and constructed-response item types. Inspection of the DETECT clusters indicated that multiple-choice/constructed-response separation generally occurred much more strongly with ELA than with mathematics or science, a pattern that has been consistent across all previous years of dimensionality analyses for the MCAS tests. Specifically, for ELA every grade had one set of clusters dominated by multiple-choice items and another set of clusters dominated by constructed-response items. This particular pattern within ELA has occurred in all previous years of the MCAS dimensionality analyses. Of the seven mathematics tests, none showed evidence of consistent separation of multiple-choice and constructed-response. Of the six science tests, only grade 5 showed strong multiple-

choice/constructed-response separation. In comparison to past years, no single grade has had consistent multiple-choice/constructed-response separation every year within the mathematics or science content areas.

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

In summary, for the 2015–16 analyses the violations of local independence, as evidenced by the DETECT effect sizes, were either weak or very weak in all cases. Thus, these effects do not seem to warrant any changes in test design or scoring. In addition, the magnitude of the violations of local independence have been consistently low over the years, and the patterns with respect to the multiple-choice and constructed-response items have also been consistent, with ELA tending to display more separation than the other two content areas.

3.6 MCAS IRT Scaling and Equating

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

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

An equating report, which provided complete documentation of the quality-control procedures and results, was reviewed by the ESE and approved prior to production of the *Spring 2016 MCAS Tests Parent/Guardian Reports* (Measured Progress Psychometrics and Research Department, *2015–2016 MCAS Equating Report*, unpublished manuscript).

Table 3-30 lists items that required intervention either during item calibration or as a result of the evaluations of the equating items. For each flagged item, the table shows the reason it was flagged (e.g., the *c*-parameter could not be estimated; the delta analysis indicated that the item's *p*-value change was much greater than that for other equating items) and what action was taken. The number of items identified for evaluation was similar to the number identified in previous years and in other state tests, across the grades and content areas. Descriptions of the evaluations and results are included in the Item Response Theory Results and Equating sections of this document. Note that the

high school science tests are included in the table below, even though those tests are pre-equated and no changes to the equating items were implemented during the operational administration. The alerts and interventions listed for the high school science tests were implemented after the operational administration as part of the quality-control process for future administrations.

Table 3-30. 2016 MCAS: Items That Required Intervention During IRT Calibration and Equating

Content Area	Grade	IREF	Reason	Action
ELA	3	308814	*a*-parameter	skipped in initial calibration
	4	307616	*c*-parameter	set *c* = 0
		307616	*b*/*b* analysis	removed from equating
		309013	*c*-parameter	set *c* = 0
		312779	*c*-parameter	set *c* = 0
	5	283363	*c*-parameter	set *c* = 0
		307289	*c*-parameter	set *c* = 0
	6	303485	*c*-parameter	set *c* = 0
	7	297219	*c*-parameter	set *c* = 0
		297233	*c*-parameter	set *c* = 0
		309112	*c*-parameter	set *c* = 0
		314021	*c*-parameter	set *c* = 0
		314029	*c*-parameter	set *c* = 0
	8	293267	*c*-parameter	set *c* = 0
		293278	*c*-parameter	set *c* = 0
		302300	*c*-parameter	set *c* = 0
		302305	*c*-parameter	set *c* = 0
		303849	*c*-parameter	set *c* = 0
	10	299076	*c*-parameter	set *c* = 0
		309158	*c*-parameter	set *c* = 0
		309180	*c*-parameter	set *c* = 0
		309551	*c*-parameter	set *c* = 0
		314406	*c*-parameter	set *c* = 0
		314408	*c*-parameter	set *c* = 0
		314699	*c*-parameter	set *c* = 0
		314827	*c*-parameter	set *c* = 0
Mathematics	3	300728	*c*-parameter	set *c* = 0
		306309	*c*-parameter	set *c* = 0
	6	311709	*c*-parameter	set *c* = 0
	8	264788	*c*-parameter	set *c* = 0
		273670	*b*/*b* analysis	removed from equating
		307500	*c*-parameter	set *c* = 0
	10	287699	delta analysis	removed from equating
		294500	*c*-parameter	set *c* = 0
		303376	*c*-parameter	set *c* = 0
		303389	*c*-parameter	set *c* = 0
		308754	*c*-parameter	set *c* = 0
		311205	*c*-parameter	set *c* = 0
		312325	*c*-parameter	set *c* = 0
		313889	*c*-parameter	set *c* = 0
STE	5	281840	*c*-parameter	set *c* = 0
		291042	*c*-parameter	set *c* = 0

continued

Content Area	Grade	IREF	Reason	Action
STE	5	291419	*c*-parameter	set *c* = 0
		316897	*c*-parameter	set *c* = 0
	8	265218	*c*-parameter	set *c* = 0
		265300	*c*-parameter	set *c* = 0
		291711	*c*-parameter	set *c* = 0
		299480	*c*-parameter	set *c* = 0
		299482	*c*-parameter	set *c* = 0
		310260	*c*-parameter	set *c* = 0
Biology	10	290951	*c*-parameter	set *c* = 0
		299860	*c*-parameter	set *c* = 0
		301396	*c*-parameter	set *c* = 0
		301396	*b*/*b* analysis	retained for equating
		304882	*c*-parameter	set *c* = 0
		310079	*c*-parameter	set *c* = 0.22
Chemistry	10	246788	delta analysis	retained for equating
		294698	*b*/*b* analysis	retained for equating
Introductory Physics	10	261148	delta analysis	retained for equating
		295865	*c*-parameter	set *c* = 0
		295865	*b*/*b* analysis	retained for equating
		301717	*c*-parameter	set *c* = 0
Technology/Engineering	10	206772	delta analysis	retained for equating
		287802	*b*/*b* analysis	retained for equating

3.6.1 IRT

All MCAS items were calibrated using IRT. IRT uses mathematical models to define a relationship between an unobserved measure of student performance, usually referred to as theta (θ), and the probability ($P(\theta)$) 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 $P(\theta)$ (Hambleton & van der Linden, 1997; Hambleton & Swaminathan, 1985). The process of determining the mathematical relationship between θ and $P(\theta)$ 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 $P(\theta)$. 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. IRT has characteristics that may be preferable to those of raw scores for equating purposes because it specifically models examinee responses at the item level, and also facilitates equating to an IRT-based item pool (Kolen & Brennan, 2014).

For the 2016 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 technology/engineering, 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 technology/engineering because there was concern that the tests might have too few examinees to support the 3PL model in future administrations.

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

$$P_i(\theta_j) = P(U_i = 1|\theta_j) = c_i + (1 - c_i)\frac{\exp[Da_i(\theta_j - b_i)]}{1+\exp[Da_i(\theta_j - b_i)]},$$

where
U indexes the scored response on an item,
i indexes the items,
j indexes students,
α represents item discrimination,
b represents item difficulty,
c is the pseudo guessing parameter,
θ is the student proficiency, and
D is a normalizing constant equal to 1.701.

For high school technology/engineering, this reduces to the following:

$$P_i(\theta_j) = P(U_i = 1|\theta_j) = \frac{\exp[D(\theta_j - b_i)]}{1+\exp[D(\theta_j - b_i)]}.$$

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 to model the probability that a student's response falls at or above a particular ordered category, given θ. This implies that a polytomous item with $k + 1$ categories can be characterized by k item category threshold curves (ICTCs) of the two-parameter logistic form:

$$P^*_{ik}(\theta_j) = P(U_i \geq k|\theta_j) = \frac{\exp[Da_i(\theta_j - b_i + d_{ik})]}{1+\exp[Da_i(\theta_j - b_i + d_{ik})]},$$

where
U indexes the scored response on an item,
i indexes the items,
j indexes students,
k indexes threshold,
θ is the student ability,
α 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), which indicate the probability of responding to a particular category given θ, are derived by subtracting adjacent ICTCs:

$$P_{ik}(\theta_j) = P(U_i = \mathrm{k}|\theta_j) = P^*_{ik}(\theta_j) - P^*_{i(k+1)}(\theta_j),$$

where
i indexes the items,
j indexes students,
k indexes threshold,
θ is the student ability,
P_{ik} represents the probability that the score on item *i* falls in category *k*, and
P^*_{ik} represents the probability that the score on item *i* falls at or above the threshold *k*

($P_{i0}^{*} = 1$ and $P_{i(m+1)}^{*} = 0$).

The GRM is also commonly expressed as:

$$P_{ik}(\theta_j) = \frac{\exp[Da_i(\theta_j - b_i + d_k)]}{1+\exp[Da_i(\theta_j - b_i + d_k)]} - \frac{\exp[Da_i(\theta_j - b_i + d_{k+1})]}{1+\exp[Da_i(\theta_j - b_i + d_{k+1})]}.$$

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

$$E(U_i|\theta_j) = \sum_k^{m+1} w_{ik} P_{ik}(\theta_j),$$

where w_{ik} is the weighting constant and is equal to the number of score points for score category *k* on item *i*.

Note that for a dichotomously scored item, $E(U_i|\theta_j) = P_i(\theta_j)$. 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 IRT Results

The tables in Appendix H give the IRT item parameters and associated standard errors of all operational scoring items on the 2016 MCAS tests by grade and content area. Note that the standard errors for some parameters are equal to zero. In these cases, the parameter or parameters were not estimated because the parameter's value was fixed (see explanation below). In addition, Appendix I contains graphs of the TCCs and TIFs, which are defined below. Note that, because the PARCC narrative writing task was administered in grades 3–8, the MCAS composition was eliminated from the ELA tests at grades 4 and 7. Therefore the TCCs and TIFs of these two ELA tests in 2016 are not comparable to those of 2015. Because of the use of the 1PL model, a TIF is not provided for high school technology/engineering.

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

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

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

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

The TIF displays the amount of statistical information that the test provides at each value of θ_j. Information functions depict test precision across the entire latent trait continuum. There is an inverse relationship between the information of a test and its standard error of measurement (SEM).

For long tests, the SEM at a given θ_j is approximately equal to the inverse of the square root of the statistical information at θ_j (Hambleton, Swaminathan, & Rogers, 1991), as follows:

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

Compared to the tails, TIFs are often higher near the middle of the θ distribution where most students are located. This is by design. Test items are often selected with middle difficulty levels and high discriminating powers so that test information is maximized for the majority of candidates who are expected to take a test.

Table 3-30 lists items that were flagged based on the quality-control checks implemented during the calibration process. (Note that some items were flagged as a result of the evaluations of the equating items; those results are described below.) In all cases, items flagged during this step were identified because of the guessing parameter (*c*-parameter) being poorly estimated. Difficulty in estimating the *c*-parameter is not at all unusual and is well documented in psychometric literature (see, e.g., 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 3-31. The number of cycles required fell within acceptable ranges (less than 150) for all tests.

Table 3-31. 2016 MCAS: Number of Newton Cycles Required for Convergence

Content Area	Grade	Cycles	
		Initial	*Equating*
ELA	3	33	31
	4	18	26
	5	29	25
	6	30	22
	7	34	21
	8	31	24
	10	80	16
Mathematics	3	33	17
	4	25	17
	5	38	18
	6	51	23
	7	27	11
	8	27	44
	10	31	9
STE	5	43	103
	8	30	93
Biology	9–12	28	1
Chemistry	9–12	33	1
Introductory Physics	9–12	34	1
Technology/Engineering	9–12	21	1

3.6.3 Equating

The purpose of equating is to ensure that scores obtained from different forms of a test are equivalent to one another. Equating may be used if multiple test forms are administered in the same year; or one year's forms may be equated 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 that taken by other students. See section 3.2 for more information about how the test development process supports successful equating.

The 2016 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 equating 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.

It is noteworthy that there were a few changes in the equating design and analyses in 2016 from the previous years. First, the ELA composition tests were eliminated in grades 4 and 7 in 2016. Secondly, the equating items changed from an external linking design to an internal linking design this year. There were fewer equating items in each test, but each of them was administered to all MCAS examinees. Lastly, the equating sample in 2016 was not representative of the state and not readily comparable to the equating samples in the previous years. In light of the changes in grades 4 and 7 ELA, two additional equating analyses were performed. One was pre-equating (i.e., equating using the existing item parameters). The other was using the 2015 sample with the writing prompt data removed. Results from the additional analyses confirmed the accuracy of the initial equating solutions and were provided in the *2015–2016 Equating Report*.

The groups of students who take equating items on the MCAS ELA reading comprehension and mathematics tests are never strictly equivalent to the groups who took the tests in the reference years. Although the nonequivalence was larger for 2016 given that more school districts chose to employ the PARCC assessments than the previous year, IRT is particularly useful for equating scenarios that involve nonequivalent groups (Allen & Yen, 1979). Equating for the MCAS uses the anchor test–nonequivalent groups design described by Petersen, Kolen, and Hoover (1989). In this equating design, no assumption is made about the equivalence of the examinee groups taking different test forms (i.e., naturally occurring groups are assumed). Comparability is instead evaluated by using a set of anchor items (also called equating items), assuming they perform in the same way in both groups and can, thus, accurately measure the differences in the two groups. Note, however, that as the differences between the two groups increase, the more difficult it is for the equating items to accurately measure the differences. Thus, given that the differences between 2015 and 2016 were greater than past years, greater caution should be used in interpreting the equating results for 2016.

Item parameter estimates for 2016 were placed on the 2015 scale by using the Fixed Common Item Parameter method (FCIP-2; Kim, 2006), which is based on the IRT principle of item parameter invariance. According to this principle, the equating items for both the 2015 and 2016 MCAS tests should have the same item parameters. Thus, prior to implementing this method, various evaluations of the equating items were conducted to check the equating items for parameter drift. These evaluations included delta analysis, rescore analysis, and IRT-based analysis. Items that were flagged as a result of these evaluations are listed in Table 3-30 at the beginning of this section. Each of these items was scrutinized, and a decision was made whether to include each item as an equating item or to discard it.

Appendix J presents the results from the delta analysis. This procedure was used to evaluate the adequacy of equating items; the discard status presented in the appendix indicates whether the item was flagged as potentially inappropriate for use in equating.

Also presented in Appendix J are the results from the rescore analysis of constructed-response items. In this analysis, 200 random papers from the previous year were interspersed with this year's papers to evaluate scorer consistency from one year to the next. An effect size—comparing the difference between last year's score and this year's score using the same set of student responses with a new set of raters—was calculated. All effect sizes were well below the criterion of 0.50.

The third and final statistical evaluation of the equating items is an IRT-based analysis. In this analysis, the item parameters for each 2016 test are first freely estimated (using PARSCALE; Muraki & Bock, 2003). The resulting item parameter estimates for the equating items are analyzed. These analyses result in *a*-plots and *b*-plots, which show the IRT parameters for the previous administrations plotted against the values for 2016. These results 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.

The equating items that successfully survived these meticulous evaluation procedures were then employed in the FCIP-2 method to place the item parameters for the nonequating items onto the previous year's scale. This method is performed by fixing the parameters of the equating items to their previously obtained on-scale values and then calibrating the remaining items using PARSCALE to place them on scale.

It is important to note that while post-equating is used for most of the MCAS tests, pre-equating is used with the high school biology, chemistry, introductory physics, and technology/engineering tests. The basic difference between post-equating and pre-equating is that every operational item on the test is treated as an equating item in pre-equating. Thus, in pre-equating the item parameters for all the operational items are estimated in a previous administration and are fixed to values estimated in a previous administration. Hence, there are no operational nonequating items that are re-estimated. These known item parameters are then used for estimating student performance. Since student performance and reported scores are based on the pre-equated item parameters, all the operational items on a pre-equated test undergo the meticulous evaluation described above for the equating items.

To provide scale validation evidence, Measured Progress performed a post-equating check for the four high school science tests. The primary purpose of the check is to ensure there was no significant drift in the parameters of the equating items and to exclude the adverse effect of parameter drift on the stability and health of the item bank. To perform the post-equating check, all the pre-equating items were re-estimated using the current year students' response data. The stability of their pre-equated item parameters were checked against their re-estimated values through *b*-*b* and delta analyses. Any item detected with a parameter drift was removed as an equating item and its item parameter was updated as needed in the item bank.

3.6.4 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 3-32. The operational θ-metric cut scores will remain fixed throughout the assessment program unless standards are reset.

Also shown in the table are the cutpoints on the reporting score scale (*2007 Standard Setting Report*).

Table 3-32. 2016 MCAS: Cut Scores on the Theta Metric and Reporting Scale by Content Area and Grade

Content Area	Grade	Theta			Scaled Score				
		Cut 1	*Cut 2*	*Cut 3*	*Min*	*Cut 1*	*Cut 2*	*Cut 3*	*Max*
ELA	3	-1.692	-0.238	1.128	200	220	240	260	280
	4	-1.126	0.067	1.572	200	220	240	260	280
	5	-1.535	-0.248	1.152	200	220	240	260	280
	6	-1.380	-0.279	1.392	200	220	240	260	280
	7	-1.529	-0.390	1.460	200	220	240	260	280
	8	-1.666	-0.637	1.189	200	220	240	260	280
	10*	-2.752	-1.495	0.153	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*	-1.555	-0.778	0.009	200	220	240	260	280
STE	5	-1.130	0.090	1.090	200	220	240	260	280
	8	-0.500	0.540	1.880	200	220	240	260	280
Biology	9–12	-0.962	-0.129	1.043	200	220	240	260	280
Chemistry	9–12	-0.134	0.425	1.150	200	220	240	260	280
Introductory Physics	9–12	-0.714	0.108	1.133	200	220	240	260	280
Technology/ Engineering	9–12	-0.366	0.201	1.300	200	220	240	260	280

* The theta cuts for grade 10 mathematics and ELA differ from those reported in technical reports prior to 2014. This is because a rescaling of these tests was conducted in summer 2013 that shifted the mean and standard deviation of the theta distribution. To maintain the same measurement scale, this required a corresponding shift in the cut scores, as well as a shift in the theta-to-scaled score transformation constants.

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

3.6.5 Reported Scaled Scores

Because the θ scale used in IRT calibrations is not understood by most stakeholders, reporting scales were developed for the MCAS. The reporting scales are linear transformations of the underlying θ scale within each performance level. Student scores on the MCAS tests are reported in even-integer values from 200 to 280. Because there are four separate transformations (one for each achievement level, shown in Table 3-33), a 2-point difference between scaled scores in the *Warning/Failing* level does not mean the same thing as a 2-point difference in the *Needs Improvement* level. Because the scales differ across achievement levels, it is not appropriate to calculate means and standard deviations with scaled scores.

By providing information that is more specific about the position of a student's results, scaled scores supplement achievement level scores. Students' raw scores (i.e., total number of points) on the 2016 MCAS tests were translated to scaled scores using a data analysis process called scaling. Scaling

simply converts from one scale to another. In the same way that a given temperature can be expressed on either the Fahrenheit or Celsius scale, or the same distance can be expressed in either miles or kilometers, student scores on the 2016 MCAS tests can be expressed in raw or scaled scores.

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

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

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

where
m is the slope and
b is the intercept.

A separate linear transformation is used for each grade and content area combination and for each achievement level. Table 3-33 shows the slope and intercept terms used to calculate the scaled scores for each grade, content area, and achievement level. Note that the values in Table 3-33 will not change unless the standards are reset.

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

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

Table 3-33. 2016 MCAS: Scaled Score Slopes and Intercepts by Content Area and Grade

Content Area	Grade	Line 1		Line 2		Line 3		Line 4	
		Slope	*Intercept*	*Slope*	*Intercept*	*Slope*	*Intercept*	*Slope*	*Intercept*
ELA	3	5.7117	229.6643	13.7552	243.2737	14.6413	243.4846	10.6838	247.9487
	4	5.6534	226.3657	16.7645	238.8768	13.2890	239.1096	14.0056	237.9832
	5	5.9667	229.1589	15.5400	243.8539	14.2857	243.5429	10.8225	247.5325
	6	5.6440	227.7888	18.1653	245.0681	11.9689	243.3393	12.4378	242.6866
	7	6.8230	230.4324	17.5593	246.8481	10.8108	244.2162	12.9870	241.0390
	8	5.8526	229.7505	19.4363	252.3810	10.9529	246.9770	11.0436	246.8691
	10	7.6084	240.9383	15.9109	263.7868	12.1359	258.1432	7.0249	258.9252
Mathematics	3	5.5287	225.5896	21.6450	241.8831	17.8891	241.5564	10.1574	249.5277
	4	5.4473	224.6792	15.2905	233.1346	23.2829	229.5460	11.8203	244.5390
	5	5.1980	223.7114	22.6244	236.1538	22.7531	236.1320	10.2512	249.2465
	6	5.3492	222.7281	26.9542	233.7466	22.7273	234.7273	10.5932	248.2203
	7	5.1900	222.5171	26.7023	232.9506	21.5983	234.2981	11.0497	246.8508
	8	4.7620	221.5143	27.1739	228.6413	22.7273	230.5000	11.7509	244.7474
	10	7.6953	231.9661	25.7400	260.0257	25.4130	259.7713	6.6867	259.9398
STE	5	5.3795	226.0788	16.3934	238.5246	20.0000	238.2000	10.4712	248.5864
	8	4.8814	222.4407	19.2308	229.6154	14.9254	231.9403	17.8571	226.4286
Biology	9–12	4.6952	224.5168	24.0096	243.0972	17.0648	242.2014	10.2197	249.3408
Chemistry	9–12	4.5317	220.6072	35.7782	224.7943	27.5862	228.2759	10.8108	247.5676
Introductory Physics	9–12	4.6776	223.3398	24.3309	237.3723	19.5122	237.8927	10.7124	247.8629
Technology/ Engineering	9–12	7.3699	222.6974	35.2734	232.9101	18.1984	236.3421	11.7647	244.7059

3.7 MCAS Reliability

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

There are a number of ways to estimate an assessment's reliability. The approach that was implemented to assess the reliability of the 2016 MCAS tests is the α coefficient of Cronbach (1951). This approach is most easily understood as an extension of a related procedure, the split-half reliability. In the split-half approach a test is split in half, and students' scores on the two half-tests are correlated. To estimate the correlation between two full-length tests, the Spearman-Brown correction (Spearman, 1910; Brown, 1910) is applied. If the correlation is high, this is evidence that the items complement one another and function well as a group, suggesting that measurement error is minimal. The split-half method requires psychometricians to select items that contribute to each half-test score. This decision may have an impact on the resulting correlation, since each different possible split of the test into halves will result in a different correlation. Cronbach's α eliminates the item selection by comparing individual item variances to total test variance, and it has been shown to be the average of all possible split-half correlations. Along with the split-half reliability, Cronbach's α is referred to as a coefficient of internal consistency. The term "internal" indicates that the index is measured internal to each test of interest, using data that come only from the test itself (Anastasi & Urbina, 1997). The formula for Cronbach's α is given as follows:

$$a = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} \sigma^2_{(Y_i)}}{\sigma^2_x}\right],$$

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

3.7.1 Reliability and Standard Errors of Measurement

Table 3-34 presents descriptive statistics, Cronbach's α coefficient, and raw score SEMs for each content area and grade. (Statistics are based on common items only.) The reliability estimates range from 0.89 to 0.93, which generally are in acceptable ranges, and are consistent with results obtained for previous administrations of the tests.[3]

[3] Note that the number of students who were included in the calculations for the mathematics and ELA tests at grades 3–8 was less than one-third the total number of state examinees because many school districts chose to administer the PARCC mathematics and ELA tests instead of the MCAS tests. Thus, comparisons to previous years should be made with caution at those grade levels.

Table 3-34. 2016 MCAS: Raw Score Descriptive Statistics, Cronbach's Alpha, and SEMs by Content Area and Grade

Content Area	Grade	Number of Students	Raw Score			Alpha	SEM
			Maximum	*Mean*	*Standard Deviation*		
ELA	3	19,682	48	37.28	7.63	0.90	2.36
	4	19,556	52	36.68	7.78	0.89	2.63
	5	19,793	52	36.03	8.48	0.89	2.79
	6	20,604	52	38.02	8.52	0.89	2.79
	7	20,031	52	39.45	7.92	0.89	2.65
	8	20,046	52	38.34	8.67	0.90	2.75
	10	69,028	72	52.84	10.18	0.89	3.34
Mathematics	3	19,700	40	30.04	7.42	0.89	2.48
	4	19,579	54	38.76	10.22	0.91	3.14
	5	19,826	54	38.41	11.18	0.91	3.36
	6	20,594	54	38.53	11.09	0.91	3.29
	7	20,018	54	36.04	12.48	0.93	3.36
	8	20,030	54	36.42	12.04	0.93	3.25
	10	69,033	60	40.01	12.55	0.92	3.48
STE	5	68,495	54	35.38	10.05	0.89	3.40
	8	69,572	54	33.93	10.34	0.89	3.48
Biology	9–12	51,147	60	39.94	11.63	0.91	3.48
Chemistry	9–12	840	60	37.26	12.71	0.92	3.58
Introductory Physics	9–12	15,317	60	37.82	11.67	0.92	3.31
Technology/ Engineering	9–12	2,713	60	35.95	10.29	0.89	3.48

Because of the dependency of the alpha coefficients on the sample, it is inappropriate to make inferences about the quality of one test by comparing its reliability to that of another test from a different grade or content area. To elaborate, reliability coefficients are highly influenced by sample characteristics such as the range of individual differences in the group (i.e., variability of the sample), average ability level of the sample that took the exams, test designs, test difficulty, test length, ceiling or floor effect, and influence of guessing. Hence, "the reported reliability coefficient is only applicable to samples similar to that on which it was computed" (Anastasi & Urbina, 1997, p. 107).

3.7.2 Subgroup Reliability

The reliability coefficients discussed in the previous section were based on the overall population of students who took the 2016 MCAS tests. Appendix P presents reliabilities for various subgroups of interest. Cronbach's α coefficients were calculated using the formula defined above based only on the members of the subgroup in question in the computations; values are calculated only for subgroups with 10 or more students. The reliability coefficients for subgroups range from 0.78 to 0.96 across the tests, with a median of 0.90 and a standard deviation of 0.02, indicating that reliabilities are generally within a reasonable range.

For several reasons, the subgroup reliability results should be interpreted with caution. First, inherent differences between grades and content areas preclude valid inferences about the reliability of a test

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

3.7.3 Reporting Subcategory Reliability

Reliabilities were calculated for the reporting subcategories within MCAS content areas, which are described in section 3.2. Cronbach's α coefficients for subcategories were calculated via the same formula defined previously using just the items of a given subcategory in the computations. Results are presented in Appendix P. The reliability coefficients for the reporting subcategories range from 0.36 to 0.89, with a median of 0.68 and a standard deviation of 0.11. Because they are based on a subset of items rather than the full test, subcategory reliabilities were typically lower than were overall test score reliabilities, approximately to the degree expected based on classical test theory, and interpretations should take this into account. Qualitative differences between grades and content areas once again preclude valid inferences about the reliability of the full test score based on statistical comparisons among subtests.

3.7.4 Reliability of Achievement Level Categorization

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

Accuracy refers to the extent to which achievement classifications based on test scores match the classifications that would have been assigned if the scores did not contain any measurement error. Accuracy must be estimated, because errorless test scores do not exist. Consistency measures the extent to which classifications based on test scores match the classifications based on scores from a second, parallel form of the same test. Consistency can be evaluated directly from actual responses to test items if two complete and parallel forms of the test are administered to the same group of students. In operational testing programs, however, such a design is usually impractical. Instead, techniques have been developed to estimate both the accuracy and consistency of classifications based on a single administration of a test. The Livingston and Lewis (1995) technique was used for the 2016 MCAS tests because it is easily adaptable to all types of testing formats, including mixed formats.

The DAC estimates reported in Tables 3-35 and 3-36 make use of "true scores" in the classical test theory sense. A true score is the score that would be obtained if a test had no measurement error. True scores cannot be observed and so must be estimated. In the Livingston and Lewis (1995) method, estimated true scores are used to categorize students into their "true" classifications.

For the 2016 MCAS tests, after various technical adjustments (described in Livingston & Lewis, 1995), a four-by-four contingency table of accuracy was created for each content area and grade,

where cell [*i,j*] represented the estimated proportion of students whose true score fell into classification *i* (where *i* = 1 to 4) and observed score fell 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 & 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.

Measured Progress also measured consistency on the 2016 MCAS tests using Cohen's (1960) coefficient κ (kappa), which assesses the proportion of consistent classifications after removing the proportion of consistent classifications that would be expected by chance. It is calculated using the following formula:

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

where

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

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

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

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

3.7.5 Decision Accuracy and Consistency Results

Results of the DAC analyses described above are provided in Table 3-35. The table includes overall accuracy indices with consistency indices displayed in parentheses next to the accuracy values, as well as overall kappa values. Overall ranges for accuracy (0.75–0.85), consistency (0.66–0.79), and kappa (0.52–0.64) indicate that the vast majority of students were classified accurately and consistently with respect to measurement error and chance. Accuracy and consistency values conditional on achievement level are also given. For these calculations, the denominator is the proportion of students associated with a given achievement level. For example, the conditional accuracy value is 0.69 for *Needs Improvement* for grade 3 mathematics. This figure indicates that among the students whose true scores placed them in this classification, 69% would be expected to be in this classification when categorized according to their observed scores. Similarly, a consistency value of 0.58 indicates that 58% of students with observed scores in the *Needs Improvement* level would be expected to score in this classification again if a second, parallel test form were taken.

For some testing situations, the greatest concern may be decisions around achievement level thresholds. For example, for tests associated with NCLB, the primary concern is distinguishing between students who are proficient and those who are not yet proficient. In this case, accuracy at

the *Needs Improvement/Proficient* threshold is critically important, which summarizes the percentage of students who are correctly classified either above or below the particular cutpoint. Table 3-36 provides accuracy and consistency estimates for the 2016 MCAS tests at each cutpoint, as well as false positive and false negative decision rates. (A false positive is the proportion of students whose observed scores were above the cut and whose true scores were below the cut. A false negative is the proportion of students whose observed scores were below the cut and whose true scores were above the cut.)

The accuracy and consistency indices at the *Needs Improvement/Proficient* threshold range from 0.90–0.96 and 0.85–0.95. The false positive and false negative decision rates at the *Needs Improvement/Proficient* threshold range from 1–5% and 2–5%. These results indicate that nearly all students were correctly classified with respect to being above or below the *Needs Improvement/Proficient* cutpoints.

Table 3-35. 2016 MCAS: Summary of Decision Accuracy (and Consistency) Results by Content Area and Grade—Overall and Conditional on Achievement Level

Content Area	Grade	Overall	Kappa	Conditional on Achievement Level			
				*Warning**	*Needs Improvement*	*Proficient*	*Advanced*
ELA	3	0.80 (0.72)	0.58	0.81 (0.67)	0.83 (0.77)	0.75 (0.68)	0.83 (0.70)
	4	0.78 (0.69)	0.55	0.82 (0.70)	0.77 (0.69)	0.76 (0.68)	0.82 (0.70)
	5	0.80 (0.71)	0.58	0.79 (0.62)	0.79 (0.70)	0.77 (0.71)	0.85 (0.76)
	6	0.79 (0.70)	0.56	0.80 (0.65)	0.75 (0.66)	0.77 (0.70)	0.85 (0.75)
	7	0.83 (0.76)	0.59	0.79 (0.61)	0.78 (0.68)	0.85 (0.82)	0.82 (0.68)
	8	0.83 (0.76)	0.6	0.79 (0.62)	0.73 (0.62)	0.85 (0.82)	0.85 (0.75)
	10	0.85 (0.79)	0.64	0.74 (0.47)	0.76 (0.62)	0.84 (0.78)	0.89 (0.83)
Mathematics	3	0.78 (0.69)	0.55	0.82 (0.68)	0.69 (0.58)	0.74 (0.66)	0.87 (0.79)
	4	0.79 (0.71)	0.59	0.83 (0.71)	0.80 (0.74)	0.70 (0.60)	0.86 (0.78)
	5	0.78 (0.69)	0.57	0.85 (0.75)	0.73 (0.63)	0.74 (0.65)	0.84 (0.77)
	6	0.77 (0.69)	0.57	0.86 (0.77)	0.68 (0.56)	0.74 (0.66)	0.84 (0.76)
	7	0.80 (0.72)	0.62	0.88 (0.81)	0.71 (0.61)	0.77 (0.69)	0.86 (0.78)
	8	0.79 (0.72)	0.62	0.88 (0.82)	0.66 (0.55)	0.72 (0.63)	0.88 (0.83)
	10	0.83 (0.76)	0.62	0.83 (0.72)	0.67 (0.56)	0.72 (0.63)	0.92 (0.89)
STE	5	0.75 (0.66)	0.52	0.85 (0.75)	0.77 (0.69)	0.68 (0.59)	0.75 (0.62)
	8	0.79 (0.71)	0.57	0.86 (0.78)	0.75 (0.67)	0.79 (0.73)	0.61 (0.35)
Biology	9–12	0.81 (0.73)	0.6	0.82 (0.69)	0.68 (0.57)	0.81 (0.75)	0.87 (0.80)
Chemistry	9–12	0.78 (0.70)	0.59	0.86 (0.78)	0.64 (0.53)	0.74 (0.65)	0.88 (0.81)
Introductory Physics	9–12	0.81 (0.73)	0.62	0.84 (0.72)	0.76 (0.67)	0.79 (0.72)	0.87 (0.80)
Technology/Engineering	9–12	0.79 (0.70)	0.56	0.83 (0.72)	0.73 (0.64)	0.82 (0.76)	0.79 (0.61)

* Failing on all high school tests.

Table 3-36. 2016 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	*False*		*Accuracy*	*False*		*Accuracy*	*False*	
		(consistency)	*Positive*	*Negative*	*(consistency)*	*Positive*	*Negative*	*(consistency)*	*Positive*	*Negative*
ELA	3	0.97 (0.96)	0.01	0.02	0.90 (0.86)	0.05	0.05	0.92 (0.89)	0.05	0.03
	4	0.95 (0.93)	0.02	0.03	0.90 (0.85)	0.05	0.05	0.93 (0.90)	0.04	0.03
	5	0.97 (0.96)	0.01	0.02	0.91 (0.87)	0.04	0.05	0.91 (0.88)	0.05	0.04
	6	0.97 (0.96)	0.01	0.02	0.91 (0.87)	0.04	0.05	0.91 (0.87)	0.05	0.04
	7	0.98 (0.97)	0.01	0.01	0.92 (0.89)	0.04	0.04	0.93 (0.90)	0.05	0.02
	8	0.98 (0.97)	0.01	0.01	0.94 (0.91)	0.03	0.04	0.92 (0.88)	0.05	0.03
	10	1.00 (0.99)	0	0	0.96 (0.95)	0.01	0.02	0.89 (0.85)	0.06	0.05
Mathematics	3	0.96 (0.94)	0.01	0.02	0.92 (0.88)	0.04	0.05	0.90 (0.86)	0.06	0.04
	4	0.96 (0.95)	0.01	0.02	0.91 (0.87)	0.05	0.05	0.92 (0.88)	0.05	0.03
	5	0.96 (0.94)	0.02	0.03	0.92 (0.89)	0.04	0.04	0.90 (0.86)	0.05	0.05
	6	0.95 (0.93)	0.02	0.03	0.92 (0.89)	0.04	0.04	0.90 (0.86)	0.05	0.04
	7	0.95 (0.93)	0.02	0.03	0.93 (0.90)	0.03	0.04	0.92 (0.89)	0.04	0.04
	8	0.95 (0.92)	0.02	0.03	0.93 (0.90)	0.04	0.04	0.92 (0.89)	0.04	0.04
	10	0.97 (0.96)	0.01	0.02	0.94 (0.92)	0.03	0.03	0.92 (0.88)	0.04	0.04
STE	5	0.94 (0.92)	0.02	0.03	0.90 (0.86)	0.05	0.05	0.91 (0.87)	0.05	0.04
	8	0.92 (0.89)	0.03	0.04	0.90 (0.86)	0.05	0.05	0.97 (0.95)	0.03	0.01
Biology	9–12	0.97 (0.95)	0.01	0.02	0.93 (0.91)	0.03	0.04	0.91 (0.87)	0.05	0.04
Chemistry	9–12	0.94 (0.91)	0.03	0.04	0.92 (0.88)	0.04	0.04	0.92 (0.89)	0.04	0.03
Introductory Physics	9–12	0.96 (0.94)	0.01	0.02	0.92 (0.89)	0.04	0.04	0.92 (0.89)	0.04	0.03
Technology/Engineering	9–12	0.93 (0.91)	0.03	0.04	0.90 (0.85)	0.05	0.05	0.96 (0.94)	0.03	0.01

* Failing on all high school tests.

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

As with other methods of evaluating reliability, DAC statistics that are calculated based on small groups can be expected to be lower than those calculated based on larger groups. For this reason, the values presented in Tables 3-35 and 3-36 should be interpreted with caution. In addition, it is important to remember that it is inappropriate to compare DAC statistics across grades and content areas.

3.8 Reporting of Results

The MCAS tests are designed to measure student achievement in the Massachusetts content standards. Consistent with this purpose, results on the MCAS were reported in terms of achievement levels, which describe student achievement in relation to these established state standards. There are four achievement levels: *Warning* (at grades 3–8) or *Failing* (at high school), *Needs Improvement*, *Proficient*, and *Advanced*. Students receive a separate achievement level classification in each content area. Reports are generated at the student level. *Parent/Guardian Reports* and student results labels are printed and mailed to districts for distribution to schools. Students who attended a school that participated in the PARCC assessment did not receive MCAS reports. In grades 5 and 8, where MCAS science testing was required of all students, students who attended schools that participated in PARCC received reports with only science results.

The details of the reports are presented in the sections that follow. See Appendix Q for a sample *Parent/Guardian Report*.

The Department also provides numerous reports to districts, schools, and teachers through its Edwin Analytics reporting system. Section 3.9.5 provides more information about the Edwin Analytics system, along with examples of commonly used reports.

3.8.1 *Parent/Guardian Report*

The *Parent/Guardian Report* is a standalone single page (11" x 17") with a center fold, and it is generated for each student eligible to take the MCAS tests. Two black-and-white copies of each student's report are printed: one for the parent and one for the school. A sample report is provided in Appendix Q. The report is designed to present parents/guardians with a detailed summary of their child's MCAS performance and to enable comparisons with other students at the school, district, and state levels. The ESE has revised the report's design several times to make the data displays more user-friendly and to add additional information, such as student growth data. The most recent revisions, in 2009 and 2010, were undertaken with input from the MCAS Technical Advisory Committee and from parent focus groups. These focus groups were held in several towns across the state, with participants from various backgrounds.

The front cover of the *Parent/Guardian Report* provides student identification information, including student name, grade, birth date, ID (SASID), school, and district. The cover also presents a commissioner's letter to parents, general information about the test, and website information for parent resources. The inside portion contains the achievement level, scaled score, and standard error of the scaled score for each content area tested. If the student does not receive a scaled score, the reason is displayed under the heading "Achievement Level." The student's historical scaled scores are reported where appropriate and available. An achievement level summary of school, district, and state results for each content area is reported. For 2016, grades 3–8 ELA and mathematics, state-level results are not 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, and an explanation of the growth percentile is provided. On the back cover, the student's performance on individual test questions is reported, along with a subcontent area summary for each tested content area. In addition, for grades 3–8 ELA and mathematics, the student's performance on the PARCC-based individual test questions is reported, along with a subcontent area summary for each tested content area. PARCC items are displayed for informational purposes and do not affect the student's scaled score and achievement level.

A note is printed on the report, in the area where the scaled score and achievement level are reported, if the student took the ELA or mathematics test with one of the following nonstandard accommodations:

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

At the high school level, there is an additional note stating whether a student has met the graduation requirement for each content area, as well as whether the student is required to fulfill an Educational Proficiency Plan (EPP) to meet the graduation requirement. EPPs are applicable to ELA and mathematics only.

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

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

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

3.8.2 Decision Rules

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

3.8.3 Quality Assurance

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

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

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

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

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

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

3.9 MCAS Validity

One purpose of this report is to describe the technical and reporting aspects of the MCAS program that support valid score interpretations. According to the *Standards for Educational and Psychological Testing* (AERA et al., 2014), considerations regarding establishing intended uses and interpretations of test results and conforming to these uses are of paramount importance in regard to valid score interpretations. These considerations are addressed in this section.

Many sections of this technical report provide evidence of validity, including sections on test design and development, test administration, scoring, scaling and equating, item analysis, reliability, and score reporting. Taken together, the technical document provides a comprehensive presentation of validity evidence associated with the MCAS program.

3.9.1 Test Content Validity Evidence

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

3.9.2 Response Process Validity Evidence

Response process validity evidence pertains to information regarding the cognitive processes used by examinees as they respond to items on an assessment. The basic question posed is: Are examinees responding to the test items as intended? This type of validity evidence is explicitly specified in the *Standards for Educational and Psychological Testing* (AERA et al., 2014; Standard 1.12).

Response process validity evidence can be gathered via cognitive interviews and/or focus groups with examinees. It is particularly important to collect this type of information prior to introducing a new test or test format, or when introducing new item types to examinees.

The ESE will ensure that evidence of response process validity is collected and reported for all new MCAS item types developed for future next-generation assessments. In particular, learning labs will be conducted for all new item types on the online test administrations to ensure that these items function as intended.

3.9.3 Internal Structure Validity Evidence

Evidence of test validity based on internal structure is presented in great detail in the discussions of item analyses, reliability, and scaling and equating in sections 3.5 through 3.7. Technical characteristics of the internal structure of the assessments are presented in terms of classical item

statistics (item difficulty, item-test correlation), DIF analyses, dimensionality analyses, reliability, SEM, and IRT parameters and procedures. Each test is equated to the previous year's test in that grade and content area to preserve the meaning of scores over time. In general, item difficulty and discrimination indices were within acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that most items were assessing consistent constructs, and students who performed well on individual items tended to perform well overall. See the individual sections for more complete results of the different analyses.

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

3.9.4 Validity Evidence in Relationships to Other Variables

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

On the next-generation MCAS assessments, the ESE will focus on collecting evidence to evaluate the extent to which the new assessments measure "student readiness for the next level" of schooling, such as readiness for the next grade level, or readiness for postsecondary education. The ESE will also collect validity evidence on new item types, such as technology-enhanced items.

3.9.5 Efforts to Support the Valid Use of MCAS Data

The ESE takes many steps to support the intended uses of MCAS data. (The intended uses are listed in section 2.4 of this report.) This section will examine some of the reporting systems and policies designed to address each use.

1. Determining school and district progress toward the goals set by the state and federal accountability systems

MCAS results and student growth percentiles are used as two categories of information in the ESE's accountability formulas for schools and districts.[4] The accountability formulas also consider the following variables when making accountability determinations for schools and districts: the rate of assessment participation, graduation rates (for high schools and districts), and student demographic group. Information on the state's accountability system is available on the ESE website at www.mass.gov/edu/government/departments-and-boards/ese/programs/accountability/reports/.

[4] Accountability for educators is addressed in the ESE's Educator Evaluation Framework documents, available here: www.doe.mass.edu/edeval/.

As documented on the accountability Web page above, the ESE carefully weighs all available evidence prior to rendering accountability decisions for schools and districts. No school, for instance, is placed in Level 4 or 5 without an agency-wide review of data, including (but not limited to) four years of assessment data. Assignment to a lower accountability level comes with increased involvement between the ESE and the local education agencies (LEAs). The different levels of engagement are explained in the State's System of Support, presented here: www.mass.gov/edu/government/departments-and-boards/ese/programs/accountability/tools-and-resources/massachusetts-tiered-system-of-support/. Among the supports, districts with schools in Level 3 get assistance with data analysis from one of the six regional District and School Assistance Centers (DSACs). The supports for LEAs in Levels 4 and 5 and documented outcomes associated with these supports are available here: www.mass.gov/edu/government/departments-and-boards/ese/programs/accountability/support-for-level-3-4-and-5-districts-and-schools/school-and-district-turnaround/.

2. 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

No student can be reported as a high school graduate in Massachusetts without first earning a CD. The typical path to earning a CD is to pass three MCAS high school exams—an ELA exam, a mathematics exam, and one of four STE exams. Most examinees in the state (around 90%, in a typical year) score *Needs Improvement* or higher on all three exams on their first try.[5] Examinees who have not earned a CD are given many opportunities to retake the exams during the retest and spring test administrations, with no limit to reexaminations. Examinees who are not awarded a CD may also appeal the decision. The ESE has instituted a rigorous appeals process that can afford some examinees the opportunity to demonstrate their competency on the state standards through the successful completion of high school course work. (Additional information on the appeals process can be found at www.doe.mass.edu/mcasappeals/.) Finally, students with significant disabilities who are unable to take the MCAS exams can participate in the MCAS-Alt program, which allows students to submit a portfolio of work that demonstrates their proficiency on the state standards. Technical information on the MCAS-Alt program is presented in Chapter 4 of this report.

3. Helping to determine the recipients of scholarships, including the John and Abigail Adams Scholarship

The same initial grade 10 test scores used to enforce the CD requirement are also used to award approximately 18,000 tuition waivers each year that can be used at Massachusetts public colleges (www.doe.mass.edu/mcas/adams.html). The tuition waivers, which do not cover school fees, are granted to the top 25% of students in each district based on their MCAS scores. Students with *Advanced* MCAS scores may also apply for the Stanley Z. Koplik Certificate of Mastery with Distinction award (www.doe.mass.edu/FamComm/Student/mastery.html).

4. Providing information to support program evaluation at the school and district levels, and

[5] To earn a CD, students must either score *Proficient* or higher on the grade 10 MCAS ELA and mathematics tests or score *Needs Improvement* on these tests and fulfill the requirements of an EPP. Students must also score *Needs Improvement* or higher on one of the four high school STE tests. Approximately 70 percent of examinees earn their CD by scoring *Proficient* or higher on the ELA and mathematics exams and *Needs Improvement* or higher on an STE exam.

5. Providing diagnostic information to help all students reach higher levels of performance

Each year, student-level data from each test administration are shared with parents/guardians and school and district stakeholders in personalized *Parent/Guardian Reports*. The current versions of these reports (see the sample provided in Appendix Q) were designed with input from groups of parents. These reports contain scaled scores and achievement levels, as well as norm-referenced student growth percentiles. They also contain item-level data broken down by standard. The reports include links that allow parents and guardians to access the released test items on the ESE website.

The ESE's secure data warehouse, Edwin Analytics, provides users with more than 150 customizable reports that feature achievement data and student demographics, geared toward educators at the classroom, school, and district levels. All reports can be filtered by year, grade, subject, and student demographic group. In addition, Edwin Analytics gives users the capacity to generate their own reports with user-selected variables and statistics. Edwin Analytics provides educators the capacity to use state-level data for programmatic and diagnostic purposes. These reports can help educators review patterns in the schools and classrooms that students attended in the past, or make plans for the schools and classrooms the students are assigned to in the coming year. The ESE monitors trends in report usage in Edwin Analytics. Between June and November (the peak reporting season for MCAS), over one million reports are run in Edwin Analytics, with approximately 400,000 reports generated in August when schools review their preliminary assessment results in preparation for the return to school.

Examples of two of the most popular reports are provided on the following pages.

The *MCAS School Results by Standards* report, shown in Figure 3-1, indicates the percentage of examinees in the school, the district, and the state with correct responses on MCAS items, grouped by the standard strand/topic. The reporting of total possible points provides educators with a sense of how reliable the statistics are, based on the number of test items/test points. The School/State Diff column allows educators to compare their school or district results to the state results. Filters provide educators with the capacity to compare student results across nine demographic categories, which include gender, race/ethnicity, economically disadvantaged status, and special education status.

Figure 3-1. MCAS School Results by Standards Report

Massachusetts Department of ELEMENTARY & SECONDARY EDUCATION

Spring 2016 MCAS School Results by Standards
Mathematics
All Students

District: Sunny Vale
School: Sunnyside Elementary
Grade: 06

All Students (61)

Standards: MA 2011 Standards

	Possible Points	School % Possible Points	District % Possible Points	State % Possible Points	School/ State Diff
Mathematics					
All items	54	81%	76%	71%	10
Question Type					
Multiple Choice	32	82%	79%	75%	7
Open Response	16	83%	75%	69%	14
Short Answer	6	68%	67%	59%	9
Strand / Topic					
Expressions and Equations	**16**	**82%**	**77%**	**72%**	**10**
Apply and extend previous understandings of arithmetic to algebraic expressions.	6	84%	80%	72%	12
Reason about and solve one-variable equations and inequalities.	8	82%	75%	74%	8
Represent and analyze quantitative relationships between dependent and independent variables.	2	78%	76%	68%	10
Geometry	**8**	**73%**	**69%**	**58%**	**15**
Solve real-world and mathematical problems involving area, surface area, and volume.	8	73%	69%	58%	15
Ratios and Proportional Relationships	**11**	**92%**	**89%**	**85%**	**7**
Understand ratio concepts and use ratio reasoning to solve problems.	11	92%	89%	85%	7
Statistics and Probability	**9**	**71%**	**63%**	**59%**	**12**
Summarize and describe distributions.	9	71%	63%	59%	12
The Number System	**10**	**82%**	**80%**	**77%**	**5**
Apply and extend previous understandings of multiplication and division to divide fractions by fractions.	1	77%	71%	71%	6
Apply and extend previous understandings of numbers to the system of rational numbers.	5	78%	79%	75%	3
Compute fluently with multi-digit numbers and find common factors and multiples.	4	88%	82%	80%	9

The *MCAS Growth Distribution* report, shown in Figure 3-2, presents the distribution of students by student growth percentile band across years, alongside the median student growth percentile and percentage of students scoring *Proficient* or *Advanced* on MCAS exams for each year. Teachers, schools, and districts use this report to monitor student growth from year to year. As in the report above, all demographic filters can be applied to examine results within student groups.

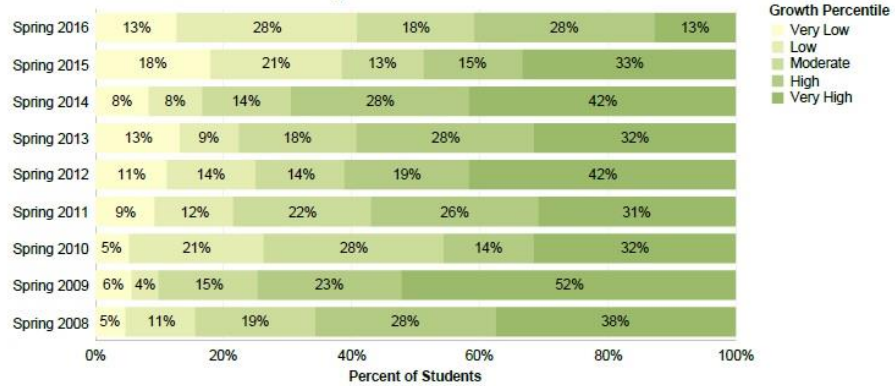
Figure 3-2. MCAS Growth Distribution Report

Massachusetts Department of ELEMENTARY & SECONDARY EDUCATION

Spring 2016 MCAS School Growth Distribution
English Language Arts Grade 06

District: Sunny Vale
Subject: English Language Arts

Growth Distribution by Year
Sunnyside Middle - All Students



Vertical lines at 20%, 40%, 60%, 80% and 100% represent the Statewide distribution for very low, low, moderate, high and very high growth.

Test	Very Low	Low	Moderate	High	Very High	Median SGP	N Students (SGP)	% Proficient or Higher	N Students (Ach. Level)
Spring 2016	9	20	13	20	9	52.0	71	81	74
Spring 2015	7	8	5	6	13	59.0	39	83	76
Spring 2014	3	3	5	10	15	76.0	36	88	40
Spring 2013	10	7	14	21	24	68.0	76	80	82
Spring 2012	8	10	10	14	30	73.5	72	88	77
Spring 2011	6	8	14	17	20	66.0	65	85	72
Spring 2010	3	12	16	8	18	54.0	57	97	59

The assessment data in Edwin Analytics are also available on the ESE public website through the school and district profiles (profiles.doe.mass.edu). In both locations, stakeholders can click on links to view released assessment items, the educational standards they assess, and the rubrics and model student work at each score point. The public is also able to view each school's progress toward the performance goals set by the state and federal accountability system.

The high-level summary provided in this section documents the ESE's efforts to promote uses of state data that enhance student, educator, and LEA outcomes while reducing less-beneficial unintended uses of the data. Collectively, this evidence documents the ESE's efforts to use MCAS results for the purposes of program and instructional improvement and as a valid component of school accountability.

Chapter 4 MCAS-Alt

4.1 Overview

4.1.1 Background

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

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

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

4.1.2 Purposes of the Assessment System

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

The purposes of the MCAS-Alt are to

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

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

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

4.1.3 Format

The MCAS-Alt consists of a portfolio containing a structured set of "evidence" based on instructional activities that is collected in each subject required for assessment during the school year. The portfolio documents the student's achievement and progress in the skills, knowledge, and concepts outlined in the state's curriculum frameworks. The portfolio also includes the student's demographic information and weekly schedule, parent/guardian verification and signoff, and a school calendar, which is submitted together with the student's "evidence" to the state each spring. Preliminary results are reported to parents/guardians, schools, and the public in June, with final results provided in August.

The Department's *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities (Incorporating the Common Core State Standards)* contains the 2006 science and technology/engineering (STE) and 2011 English language arts (ELA) and mathematics standards, and describes the content to be assessed by the MCAS-Alt. It also provides strategies for adapting and using the state's learning standards to instruct and assess students taking the MCAS-Alt. The fall 2016 *Resource Guide* is intended to ensure that all students receive instruction in the Common Core State Standards in ELA and mathematics, as well as the science and technology/engineering standards, at levels that are challenging and attainable for each student. It is also intended to serve as a guide for teachers who work with students with more significant disabilities who are participating in the MCAS-Alt. For the MCAS-Alt, students are expected to achieve the same standards as their nondisabled peers. However, they may need to learn the necessary knowledge and skills differently, such as through presentation of the knowledge/skills at lower levels of complexity, in smaller segments, and at a slower pace.

4.2 Test Design and Development

4.2.1 Test Content

MCAS-Alt assessments are required for all grades and content areas in which standard MCAS tests are administered, although the range and level of complexity (but not the *essence*) of the standards being assessed have been modified. Specific MCAS-Alt content areas and strands/domains required for students in each grade level are listed in Table 4-1.

New in 2016 is the introduction of an ELA–Writing assessment in grades 3–8 and 10, a change from past years when only students in grades 4, 7, and 10 were assessed in ELA–Composition.

Table 4-1. 2016 MCAS-Alt: Requirements

Grade	ELA Strands Required	Mathematics Strands Required	STE Strands Required
3	▪ Language ▪ Reading ▪ Writing	▪ Operations and Algebraic Thinking ▪ Measurement and Data	
4	▪ Language ▪ Reading ▪ Writing	▪ Operations and Algebraic Thinking ▪ Numbers and Operations – Fractions	
5	▪ Language ▪ Reading ▪ Writing	▪ Number and Operations in Base Ten ▪ Number and Operations – Fractions	Any three of the four STE strands*
6	▪ Language ▪ Reading ▪ Writing	▪ Ratios and Proportional Relationship ▪ The Number System	
7	▪ Language ▪ Reading ▪ Writing	▪ Ratios and Proportional Relationships ▪ Geometry	
8	▪ Language ▪ Reading ▪ Writing	▪ Expressions and Equations ▪ Geometry	Any three of the four STE strands*
10	▪ Language ▪ Reading ▪ Writing	Any three of the five mathematics *conceptual categories*: ▪ Functions ▪ Geometry ▪ Statistics and Probability ▪ Number and Quantity ▪ Algebra	Any three learning standards in one of the following strands: ▪ Biology ▪ Chemistry ▪ Introductory Physics or ▪ Technology/Engineering

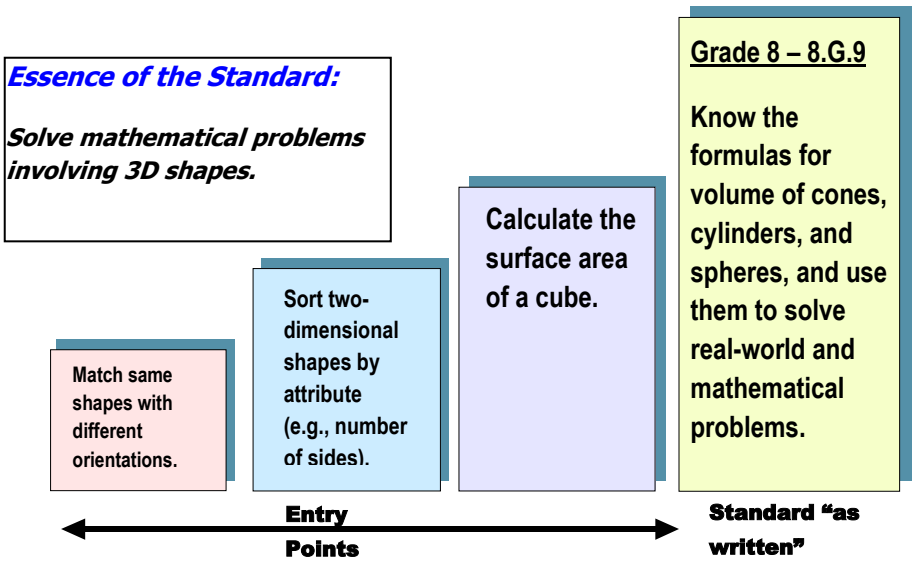
* Earth and Space Science, Life Science, Physical Sciences, Technology/Engineering

4.2.1.1 Access to the Grade-Level Curriculum

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

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

Figure 4-1. 2016 MCAS-Alt: Access to the General Curriculum (Mathematics Example) Through Entry Points That Address the Essence of the Standard



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 pieces must be a data chart (e.g., field data chart, line graph, or bar graph) that includes

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

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

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

Each piece of primary evidence must be labeled with

- the student's name,
- the date of the activity,

- a brief description of how the task or activity was conducted and what the student was asked to do,
- the percentage of accuracy for the performance, and
- the percentage of independence for the performance.

The data chart and at least two additional pieces of primary evidence compose the "core set of evidence" required in each portfolio strand, with the exception of the ELA–Writing strand, which consists only of three samples of the student's expressive communication.

Supporting Documentation

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

- photographs of the student that show how the student engaged in the instructional activity (i.e., the context of the activity)
- tools, templates, graphic organizers, or models used by the student
- reflection sheet or other self-evaluation documenting the student's awareness, perceptions, choice, decision-making, and self-assessment of work he or she created, and the learning that occurred as a result. For example, a student may respond to questions such as:
  - What did I do? What did I learn?
  - What did I do well? What am I good at?
  - Did I correct my inaccurate responses?
  - How could I do better? Where do I need help?
  - What should I work on next? What would I like to learn?
- work sample description labels providing important information about the activity or work sample

4.2.1.3 Assessment Dimensions (Scoring Rubric Areas)

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

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

- **completeness** of portfolio materials
- **level of complexity** at which the student addressed learning standards in the Massachusetts curriculum frameworks in the content area being assessed
- **accuracy** of the student's responses or performance of specific tasks
- **independence** demonstrated by the student in responding to questions or performing tasks

- **self-evaluation** during or after each task or activity (e.g., reflection, self-correction, goal-setting)
- **generalized performance** of the skill in different instructional contexts, or using different materials or methods of presentation or response

4.2.1.4 MCAS-Alt Grade-Level and Competency Portfolios

All high school students, including students with disabilities, are required to meet the CD standard to be eligible to earn a high school diploma. Students must attain a score of *Proficient* or higher on the MCAS ELA and mathematics tests (or a score of *Needs Improvement*, plus fulfilling the requirements of an Educational Proficiency Plan [EPP]) and a minimum score of *Needs Improvement* on an MCAS high school STE test. Students with disabilities who take alternate assessments in Massachusetts can meet the graduation requirement by submitting a competency portfolio that demonstrates a level of performance equivalent to a student who has achieved these scores on the standard MCAS tests.

MCAS-Alt competency portfolios in ELA, mathematics, and STE include a collection of work samples that assess a broader range of standards than are assessed by the basic MCAS-Alt portfolio. Competency portfolios are evaluated by panels of content experts to ensure that they meet the appropriate standard of performance in that subject. Since students with significant cognitive disabilities comprise the majority of students taking alternate assessments, however, the proportion of students who achieve scores of *Needs Improvement* on the MCAS-Alt in ELA, mathematics, and STE will likely remain low in comparison to the number of students who meet the CD requirement by taking standard MCAS tests.

For students in grades 3–8, a grade-level portfolio may be submitted that assesses a broader range of standards than those assessed in the basic MCAS-Alt portfolio, if the student is working at or close to grade-level expectations and wishes to earn a score of *Needs Improvement* or higher on the assessment. A relatively small number of MCAS-Alt grade-level portfolios and competency portfolios (for high school students) are submitted each year for students who address learning standards at or near grade-level expectations but are unable to participate in standard MCAS testing, even with accommodations, due to a significant disability. Participation rates for 2016 are provided in section 4.3.3.3. The participation guidelines section of the *Educator's Manual for MCAS-Alt* (available at www.doe.mass.edu/mcas/alt/edmanual.pdf) describes the characteristics of the students who should be considered for the MCAS-Alt by their IEP team or 504 plan coordinator, as well as the characteristics of those students for whom it may also be appropriate to submit grade-level and competency portfolios.

For additional information on how the 2016 MCAS-Alt grade-level and competency portfolios were evaluated, see section 4.4 of this report.

4.2.2 Test Development

4.2.2.1 Rationale

Alternate assessment is the component of the state's assessment system that measures the academic performance of students with the most significant disabilities. Students with disabilities are required by federal and state laws to participate in the MCAS so that their performance of skills and

knowledge of content described in the state's curriculum frameworks can be assessed, and so they can be visible and accountable in reports of results for each school and district.

The federal Elementary and Secondary Education Act (ESEA) requires states to include an alternate assessment option for certain students with disabilities. This requirement ensures that students with significant disabilities receive academic instruction based in the state's learning standards, have an opportunity to "show what they know" on the state assessment, and are included in reporting and accountability. Alternate assessment results provide accurate and detailed feedback that can be used to identify challenging instructional goals for each student. When schools are held accountable for the performance of students with disabilities, these students are more likely to receive consideration when school resources are allocated.

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

4.2.2.2 Role of the Advisory Committee

An MCAS-Alt Advisory Committee meets periodically to receive updates and discuss policy issues related to the alternate assessment. This diverse group of stakeholders—including teachers, school administrators, special education directors, parents/guardians, advocates, approved private school and educational collaborative personnel, and representatives of institutions of higher education—has been crucial in assisting the Department to develop, implement, and continue the enhancement of the MCAS-Alt. A list of advisory committee members is provided in Appendix A.

4.3 Test Administration

4.3.1 Evidence Collection

Each portfolio strand (with the exception of ELA–Writing) must include a data chart documenting the student's performance (i.e., the percentage of accuracy and independence of the performance) and progress (whether the rates of accuracy and/or independence increase over time) in learning a new academic skill related to the standard(s) required for assessment. Data are collected on at least eight different dates to determine whether progress has been made and the degree to which the skill has been mastered. On each date, the data point must indicate the percentage of correct versus inaccurate responses given on that date and whether the student required cues or prompts to respond (i.e., the overall percentage of independent responses given by the student). Data are collected either during routine classroom instruction or during tasks and activities set up specifically for the purpose of assessing the student. All data charts include a brief description of the activity (or activities) conducted on each date, describing how the task relates to the measurable outcome being assessed. Data charts may include performance data either from a collection of work samples or from a series of responses to specific tasks summarized for each date.

In addition to the chart of instructional data, each portfolio strand must include at least two individual work samples (or photographs, if the student's work is large, fragile, or temporary in nature) that provide evidence of the percentage of accuracy and independence of the student's responses on a given date, based on the same measurable outcome that was documented in the instructional data chart.

The ELA–Writing strand requires that students submit **at least three writing samples** that demonstrate their expressive communication skills, based on *any combination* of the following text types found in the *2011 Massachusetts Curriculum Frameworks*:

1. Opinion (grades 3–5)/Argument (grades 6–8 and 10)
2. Informative/Explanatory text
3. Narrative
4. Poetry

In addition to the three writing samples, a **baseline sample** must be submitted with each final writing sample of the same text type. The baseline sample must be dated before the final sample in the same text type, and may include an outline, completed graphic organizer, or draft of the same assignment as the final sample. The baseline sample should provide information to inform additional instruction in that text type.

4.3.2 Construction of Portfolios

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

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

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

4.3.3 Participation Requirements

4.3.3.1 Identification of Students

All students educated with Massachusetts public funds, including students with disabilities educated inside or outside their home districts, must be engaged in an instructional program guided by the standards in the Massachusetts curriculum frameworks and must participate in assessments that correspond with the grades in which they are reported in the ESE's Student Information Management System (SIMS). Students with significant disabilities who are unable to take the standard MCAS tests, even with accommodations, must take the MCAS-Alt, as determined by the student's IEP team or designated in his or her 504 plan.

4.3.3.2 Participation Guidelines

A student's IEP team (or 504 coordinator, in consultation with other staff) 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 if the student is unable to take the standard test because of the severity of his or her disabilities. This information is documented in the student's IEP or 504 plan and must be revisited on an annual basis. A student may take the general assessment, with or without accommodations, in one subject and the alternate assessment in another subject.

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

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

A student's team must review the options provided in Figure 4-2.

Figure 4-2. 2016 MCAS-Alt: Participation Guidelines

OPTION 1

Characteristics of Student's Instructional Program and Local Assessment	Recommended Participation in MCAS
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 http://www.doe.mass.edu/mcas/accessibility/) and that are documented in an approved IEP or 504 plan prior to testing.

OPTION 2

Characteristics of Student's Instructional Program and Local Assessment	Recommended Participation in MCAS
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; ***or*** 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.

OPTION 3

Characteristics of Student's Instructional Program and Local Assessment	Recommended Participation in MCAS
If the student is	***Then***
a) working on learning standards at or near grade-level expectations; ***and*** is b) ***sometimes able*** to take a paper-and-pencil test, either without accommodations or with one or more accommodation(s); ***but*** c) has a complex and significant disability that does not allow the student to fully demonstrate knowledge and skills on a test of this format and duration, (Examples of complex and significant disabilities for which the student may require an alternate assessment are provided on the following page.)	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 (in grades 3–8) or "competency" portfolio (in high school) 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, the student's IEP team or 504 plan may determine that the student should take the MCAS-Alt, even though the student may be working at or near grade-level expectations.

In addition to the criteria outlined in Options 2 and 3, the following are examples of unique circumstances that would warrant use of the MCAS-Alt.

- 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, or the entire test during the testing window).

4.3.3.3 MCAS-Alt Participation Rates

Across all content areas, a total of 8,373 students, or 1.7% of the assessed population, participated in the 2016 MCAS-Alt in grades 3–10. A slightly higher relative proportion of students in grades 3–8 took the MCAS-Alt compared with students in grade 10, and slightly more students were alternately assessed in mathematics than in ELA. Additional information about MCAS-Alt participation rates by content area is provided in Appendix B, including the comparative rate of participation in each MCAS assessment format (i.e., routinely tested, tested with accommodations, or alternately assessed). The 2016 MCAS-Alt State Summary is available at http://www.doe.mass.edu/mcas/alt/2016statesum.docx

4.3.4 Educator Training

During October 2015, a total of 2,840 educators and administrators received training on conducting the 2016 MCAS-Alt. Attendees had the option of participating in one of three "tracks": an introduction to MCAS-Alt for educators new to the assessment, an update for those with previous MCAS-Alt experience, and an administrator's overview. Topics for the introduction session included the following:

- decision-making regarding which students should take the MCAS-Alt
- portfolio requirements in each grade and content area
- developing measurable outcomes using the *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities* (Fall 2006 and 2014)
- collecting data on student performance and progress based on measurable outcomes

Topics for the update session included the following:

- a summary of the statewide 2015 MCAS-Alt results
- changes to the MCAS-Alt requirements for 2016, including requirements of the ELA–Writing strand
- where to find information in the *2016 Educator's Manual for MCAS-Alt*
- avoiding mistakes that lead to scores of *Incomplete*
- reporting results
- using data charts to improve teaching and learning
- competency and grade-level portfolio requirements

- accessing the general curriculum and preparing alternate assessment portfolios for students with the most severe cognitive disabilities

Topics for the administrator's session included the following:

- purposes of MCAS-Alt
- who should take MCAS-Alt
- what MCAS-Alt assesses
- MCAS-Alt results
    - participation, performance, and trends over time
- principals' role in MCAS-Alt
- findings of the biannual MCAS-Alt teacher survey

During January 2016, a total of 1,437 educators attended training sessions in which they were able to review and discuss their students' portfolios and have their questions answered by MCAS-Alt Training Specialists (i.e., expert teachers). Some were new to the process and had not attended the introduction training in the fall; others wished to discuss their specific questions and concerns about their portfolios-in-progress with expert teachers.

These training sessions were repeated in February and March 2016, with an additional 1,126 educators in attendance.

4.3.5 Support for Educators

A total of 115 MCAS-Alt Training Specialists were trained by the ESE to provide assistance and support for teachers conducting the MCAS-Alt in their districts, as well as to assist the Department at eight Department-sponsored portfolio review training sessions in January, February, and March 2016. In addition, ESE staff provided assistance throughout the year via e-mail and telephone to educators with specific questions about their portfolios.

The MCAS Service Center provided toll-free telephone support to district and school staff regarding test administration, reporting, training, materials, and other relevant operations and logistics. The Measured Progress project management team provided extensive training to the MCAS Service Center staff on the logistical, programmatic, and content-specific aspects of the MCAS-Alt, including Web-based applications used by the districts and schools to order materials and schedule shipment pickups. Informative scripts were used by the Service Center coordinator and approved by the ESE to train Service Center staff in relevant areas 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 2016. The ESE and Measured Progress trained and closely monitored scorers to ensure that portfolio scores were accurate.

Evidence of the student's performance was evaluated and scored using research-based criteria for how students with significant disabilities learn and demonstrate knowledge and skills. The criteria included the application of a universal scoring rubric; verification that measurable outcomes were aligned with the standards required for assessment in the *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities*; and rigorous training and qualification of

scorers based on the *2016 Guidelines for Scoring MCAS-Alt Portfolios*. The *MCAS-Alt Rubric for Scoring Portfolio Strands* was developed with assistance from teachers and the statewide advisory committee. The criteria for scoring portfolios are listed and described in detail on the following pages.

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

4.4.1 Scoring Logistics

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

Security was maintained at the scoring site; access to unscored portfolios was restricted to ESE and Measured Progress staff. MCAS-Alt scoring leadership staff included several floor managers (FMs) and table leaders (TLs). The scoring room was monitored by as many as five FMs, Massachusetts educators who are current or were previous MCAS-Alt teacher consultants. Each FM managed a group of tables at the elementary, middle, or secondary level. Each TL managed a table with four to five scorers.

Communication and coordination among scorers was 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 over the course of scoring in 2016 and throughout summer 2016 to accomplish the following:

- select sample portfolio strands to use for training, calibration, and qualification of scorers in 2016
- discuss issues to be addressed during fall 2016 training sessions and in the *2016 Educator's Manual for MCAS-Alt* and *2016 Guidelines for Scoring Student Portfolios*

All sample strands were scored using the 2016 guidelines, noting any scoring problems that arose during the review. All concerns were resolved by using the *2016 Educator's Manual for MCAS-Alt* or by following additional scoring rules agreed upon by the PLT and subsequently addressed in the final 2016 guidelines.

Of the portfolios reviewed, several sample strands were set aside in spring 2016 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 triple scores, only the ones in exact agreement in all five scoring dimensions—Level of Complexity, Demonstration of Skills and Concepts, Independence, Self-Evaluation, and Generalized Performance—were considered as possible exemplars.

Recruitment and Training of Scorers

Recruitment

Through Kelly Services, Measured Progress recruited prospective scorers and TLs for the MCAS-Alt Scoring Center. All TLs and many scorers had previously worked on scoring projects for other states' test or alternate assessment administrations, and all had four-year college degrees. Additionally, the PLT recruited 16 MCAS-Alt Training Specialists to assist the ESE and Measured Progress, many of whom had previously served as TLs or scorers.

Training

Scorers were rigorously trained in all rubric dimensions by reviewing scoring rules and "mock scoring" of numerous sample portfolio strands selected to illustrate examples of each rubric score point. Scorers were given detailed instructions on how to review data charts and other primary evidence to tally the rubric area scores using a strand organizer. Trainers facilitated discussions and review among scorers to clarify the rationale for each score point and describe special scoring scenarios and exceptions to the general scoring rules.

Scorer Qualification

Before scoring actual student portfolios, each scorer was required to take a qualifying assessment consisting of 21 questions and to score four sample portfolio strands (i.e., 20 scoring dimensions). To qualify as a scorer, the threshold score on the 21 questions was 85% (18 correct out of 21 total questions); and the threshold score on the portfolio strands was 85% exact agreement overall for the five scoring dimensions (i.e., exact agreement on 17 out of 20 scorable dimensions for the four strands).

Scorers who did not achieve the required percentages were retrained using another qualifying assessment. Those who achieved the required percentages 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

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

Sixteen MCAS-Alt Training Specialists were selected to participate in portfolio scoring and were designated as expert scorers who assisted in resolving scores of "M" (indicating that evidence was missing or insufficient to determine a score), and in the training/retraining of TLs.

4.4.3 Scoring Methodology

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

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

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

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

During spring 2016, scorers used an automated, customized scoring program called *AltScore* to score MCAS-Alt portfolios. Scorers were guided through the scoring process by answering a series of yes/no and fill-in-the-blank questions onscreen, which the program used to calculate the correct score. The computer-based scoring application made it possible for scorers to focus exclusively and sequentially on each portfolio product and record the necessary information, rather than keeping track of products they had previously viewed and calculating the score.

The MCAS-Alt 2016 score distributions for all scoring dimensions are provided in Appendix F.

A. Level of Complexity

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

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

Table 4-2. 2016 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 (one or two) at grade-level expectations in this strand.	Student addresses a broad range of curriculum framework learning standards (three or more) at grade-level expectations in this strand.

B. Demonstration of Skills and Concepts

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

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

- the student's name
- the date of performance
- a brief description of the activity
- the percentage of accuracy
- the percentage of independence

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

Brief descriptions of each activity on the data chart were also considered in determining the completeness of a data chart. Educators had been instructed during educator training workshops and in the *2016 Educator's Manual for MCAS-Alt* that "each data chart must include a brief description beneath each data point that clearly illustrates how the task or activity relates to the measurable outcome being assessed." One- or two-word descriptions were likely to be considered insufficient to document the relationship between the activity and the measurable outcome and therefore would result in the exclusion of those data points from being scored.

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

A score of M was also given if any of the following was true:

- The data chart listed the percentages of both accuracy and independence at or above 80% at the beginning of the data collection period, indicating that the student did not learn a challenging new skill in the strand and was instead addressing a skill he or she already had learned.
- The data chart did not document a single measurable outcome based on the required learning standard or strand on at least eight different dates, and/or did not indicate the student's accuracy and independence on each task or trial.
- Two additional pieces of primary evidence did not address the same measurable outcome as the data chart or were not labeled with all required information.

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

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

- all data points in the final 1/3 time frame of the data chart, and
- all other primary evidence in the strand produced during or after the final 1/3 time frame (provided the piece was not already included on the chart).

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

Table 4-3. 2016 MCAS-Alt: Scoring Rubric for Demonstration of Skills and Concepts

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

C. Independence

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

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

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

- all data points during the final 1/3 time frame of the data chart, and
- all other primary evidence in the strand produced during or after the final 1/3 time frame (provided the piece was not already included on the chart).

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

A score of M was given in both Demonstration of Skills and Concepts and Independence if at least two pieces of scorable primary evidence and a completed data chart documenting the student's performance of the same skill were not submitted.

A score of M was also given if any of the following was true:

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

Table 4-4. 2016 MCAS-Alt: Scoring Rubric for Independence

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

D. Self-Evaluation

The score for Self-Evaluation indicates the frequency of activities in the portfolio strand that involve self-correction, task-monitoring, goal-setting, reflection, and overall awareness by the student of his or her own learning. The 2016 MCAS-Alt score distributions for Self-Evaluation are provided in Appendix F.

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

Table 4-5. 2016 MCAS-Alt: Scoring Rubric for Self-Evaluation, Individual Strand Score

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

E. Generalized Performance

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

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

Table 4-6. 2016 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**.

4.4.3.1 ELA–Writing Scoring Methodology

Prior to submission, teachers were asked to score each of their student's three final writing samples using the state-provided rubrics shown in Appendix U. The four rubrics were each labeled according to the appropriate text type:

1. Opinions/Arguments
2. Informative/Explanatory texts
3. Narrative
4. Poetry

MCAS-Alt scorers verified the scores submitted by the teacher based solely on the responses generated by the *student*, rather than any text provided by the teacher. The rubric scores were lowered by scorers in cases where scores did not reflect the student's work.

Additional Information About ELA–Writing:

- Writing samples must be produced as independently as possible by the student. If teachers provide text for the student or apply their own revisions to the student's work, this is reflected in the score, particularly in the rubric area of Independence. Teachers are expected to explain how edits and revisions were made and indicate the student's contribution to the creation of the sample.
- Writing samples dictated to a scribe must be transcribed verbatim, with the scribe assuming capital letters and basic punctuation.
- Teachers are permitted to submit students' open-responses to **reading comprehension** questions as the basis of the writing samples, even if those responses are already part of the evidence compiled for the ELA–Reading strand.

4.4.4 Monitoring the Scoring Quality

The FM oversaw the general flow of work in the scoring room and monitored overall scoring consistency and accuracy, particularly among TLs. The TLs ensured that scorers at their table were consistent and accurate in their scoring.

Scoring consistency and accuracy were maintained using the following methods:

- double-scoring
- resolution (i.e., read-behind) scoring

Double-Scoring

Double-scoring meant that a portfolio was scored by two scorers at different tables, with neither scorer knowing the score assigned by the other.

For portfolios in all grades and subjects, at least one of the portfolios of each scorer was double-scored each morning and afternoon; or, at minimum, every fifth portfolio (i.e., 20 percent of the total scored) for each scorer was double-scored.

The required rate of scoring accuracy for double-scored portfolios was 80% exact agreement. The TL retrained any scorer whose interrater consistency fell below 80% agreement with the TL's resolution score. The TL reviewed discrepant scores with the responsible scorers and determined when they could resume scoring.

Table 4-10 in section 4.7.3 shows the percentages of interrater agreement for the 2016 MCAS-Alt.

Resolution Scoring

Resolution scoring refers to the rescoring of a portfolio by a TL and a comparison of the TL's score with the one assigned by the previous scorer. If there was exact score agreement, the first score was retained as the score of record. If the scores differed, the TL's score became the score of record.

Resolution scoring was conducted on all portfolios during the first full day of scoring. After that, a double-score was performed at least once each morning, once each afternoon, and on every fifth subsequent portfolio per scorer.

The required rate of agreement between a scorer and the score of record was 80% exact agreement. A double-score was performed on each subsequent portfolio for any scorer who fell below 80% interrater consistency and who was permitted to resume scoring after being retrained, until consistency with the TL's scores was established.

Tracking Scorer Performance

A real-time and cumulative scorer data record was maintained digitally for each scorer. The scorer data record showed the number of portfolio strands and portfolios scored by each scorer, plus his or her interrater consistency in each rubric dimension.

In addition to maintaining a record of scorers' accuracy and consistency over time, leadership also monitored scorers for output, with slower scorers remediated to increase their production. The overall ratings were used to enhance the efficiency, accuracy, and productivity of scorers.

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

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

Grade-Level Portfolios in Grades 3–8

Students in grades 3–8 who required an alternate assessment, but who were working at or close to grade-level expectations, submitted grade-level portfolios in one or more subjects required for assessment. Grade-level portfolios included an expanded array of work samples that demonstrated the student's attainment of a range of grade-equivalent skills, according to guidelines outlined in the *Educator's Manual for MCAS-Alt*.

Each grade-level portfolio was evaluated by a panel of content area experts to determine whether it achieved a score of *Needs Improvement* or higher. To receive an achievement level of *Needs Improvement* or higher, the portfolio must have demonstrated

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

Competency Portfolios in High School

Students in high school who required an alternate assessment, but who were working at or close to grade-level expectations, submitted competency portfolios in one or more subjects required for assessment. Competency portfolios included work samples that demonstrated the student's attainment of the skills and content assessed by the grade 10 MCAS test in that subject.

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

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

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

4.5 MCAS-Alt Classical Item Analyses

As noted in Brown (1983), "A test is only as good as the items it contains." A complete evaluation of a test's quality must therefore include an evaluation of each item. Both *Standards for Educational and Psychological Testing* (AERA et al., 2014) and the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004) include standards for identifying high-quality items. While the specific statistical criteria identified in these publications were developed primarily for general, rather than alternate, assessments, the principles and some of the techniques apply to the alternate assessment framework as well. Both qualitative and quantitative analyses are conducted to ensure that the MCAS-Alt meets these standards. Qualitative analyses are described in earlier sections of this chapter; this section focuses on quantitative evaluations.

Quantitative analyses presented here are based on the statewide administration of the 2016 MCAS-Alt including three of the five dimension scores on each task (Level of Complexity, Demonstration of Skills and Concepts, and Independence). Although the other two dimension scores (Self-Evaluation and Generalized Performance) are reported, they do not contribute to a student's overall achievement level; therefore, they are not included in quantitative analyses.

For each MCAS-Alt subject and strand, dimensions are scored polytomously across tasks according to scoring rubrics described previously. Specifically, a student can achieve a score of 1, 2, 3, 4, or 5 on the Level of Complexity dimension and a score of M, 1, 2, 3, or 4 for the Demonstration of Skills and Concepts and Independence dimensions. Dimensions within subjects and strands are subsequently treated as traditional test items, as they capture or represent student performance against the content of interest; dimension scores for each strand are, therefore, treated as item scores for the purpose of conducting quantitative analyses.

Statistical evaluations of MCAS-Alt include difficulty and discrimination indices, structural relationships (correlations among the dimensions), and bias and fairness. Item-level classical statistics—item difficulty and discrimination values—are provided in Appendix E. Item-level score distributions (i.e., the percentage of students who received each score point) are provided in Appendix F for each item. Note that the Self-Evaluation and Generalized Performance dimension scores are also included in Appendix F.

4.5.1 Difficulty

Following from the definition of dimensions and dimension scores as similar to traditional test items and scores, all items are evaluated in terms of difficulty according to standard classical test theory practices. Difficulty is traditionally described according to an item's *p*-value, which is calculated as the average proportion of points achieved on the item. Dimension scores achieved by each student are divided by the maximum possible score to return the proportion of points achieved on each item;

p-values are then calculated as the average of these proportions. Computing the difficulty index in this manner places items on a scale that ranges from 0.0 to 1.0. This statistic 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 comprised entirely of very easy or very hard items, all students would receive nearly the same scores, and the assessment would not be able to differentiate high-ability students from low-ability students.

It is worth mentioning that using norm-referenced criteria such as *p*-values to evaluate test items is somewhat contradictory to the purpose of a criterion-referenced assessment like the MCAS-Alt. Criterion-referenced assessments are primarily intended to provide evidence of student progress relative to a standard rather than provide a comparison with other students. In addition, the MCAS-Alt makes use of teacher-designed instructional activities which then serve as items to measure performance. For these reasons, the generally accepted criteria regarding classical item statistics should be cautiously applied to the MCAS-Alt.

A summary of item difficulty for each grade and content area is presented in Table 4-7. The mean difficulty values shown in the table indicate that, overall, students performed well on the items on the MCAS-Alt. In assessments designed for the general population, difficulty values 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 4-7 should not be interpreted to mean that the students performed better on the MCAS-Alt than the students who took general assessments performed on those tests.

4.5.2 Discrimination

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. 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. A desirable feature of an item is that the higher-ability students perform better on the item than lower-ability students or that the item demonstrates strong, positive item-test correlation.

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. For example, in grade 3 ELA, total test score corresponds to the sum of scores received on the three dimensions included in quantitative analyses (i.e., Level of Complexity, Demonstration of Skills and Concepts, and Independence) across both Language and Reading strands.

The discrimination index used to evaluate MCAS-Alt items was the Pearson product-moment correlation, which has a theoretical range of -1.0 to 1.0. A summary of the item discrimination statistics for each grade and content area is also presented in Table 4-7. Because the nature of the MCAS-Alt is different from that of a general assessment, and because very few guidelines exist as to criteria for interpreting these values for alternate assessments, the statistics presented in Table 4-7 should be interpreted with caution.

Table 4-7. 2016 MCAS-Alt: Summary of Item Difficulty and Discrimination Statistics by Content Area and Grade

Content Area	Grade	Number of Items	*p*-Value		Discrimination	
			Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
ELA	3	9	0.79	0.20	0.44	0.08
	4	9	0.79	0.20	0.39	0.05
	5	9	0.79	0.20	0.40	0.09
	6	9	0.79	0.19	0.35	0.10
	7	9	0.80	0.19	0.39	0.06
	8	9	0.79	0.19	0.42	0.07
	HS	9	0.78	0.19	0.39	0.06
Mathematics	3	9	0.85	0.20	0.64	0.09
	4	9	0.85	0.20	0.59	0.08
	5	9	0.85	0.20	0.63	0.10
	6	9	0.85	0.20	0.62	0.07
	7	9	0.85	0.20	0.62	0.06
	8	9	0.84	0.20	0.62	0.05
	HS	15	0.84	0.19	0.45	0.05
STE	5	12	0.85	0.19	0.45	0.10
	8	12	0.85	0.19	0.41	0.06
Biology	HS	12	0.84	0.19	0.42	0.05
Chemistry	HS	9	0.81	0.17	0.32	0.20
Introductory Physics	HS	9	0.81	0.17	0.60	0.11
Technology/ Engineering	HS	9	0.82	0.19	0.56	0.09

4.5.3 Structural Relationships Between Dimensions

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

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

Table 4-8. 2016 MCAS-Alt: Average Correlations Among the Three Dimensions by Content Area and Grade

Content Area	Grade	Number of Items Per Dimension	Average Correlation Between:*			Correlation Standard Deviation*		
			Comp/ Ind	*Comp/ Sk*	*Ind/ Sk*	*Comp/ Ind*	*Comp/ Sk*	*Ind/ Sk*
ELA	3	3	0.16	0.18	0.19	0.09	0.13	0.06
	4	3	0.18	0.21	0.17	0.07	0.10	0.06
	5	3	0.20	0.18	0.21	0.07	0.14	0.06
	6	3	0.10	0.17	0.13	0.02	0.11	0.05
	7	3	0.17	0.20	0.17	0.02	0.08	0.04
	8	3	0.23	0.28	0.23	0.05	0.06	0.01
	HS	3	0.21	0.28	0.15	0.05	0.08	0.09
Mathematics	3	2	0.25	0.12	0.32	0.01	0.00	0.02
	4	2	0.18	0.15	0.14	0.03	0.06	0.02
	5	2	0.22	0.15	0.26	0.03	0.02	0.01
	6	2	0.22	0.23	0.13	0.02	0.01	0.01
	7	2	0.20	0.17	0.18	0.02	0.03	0.01
	8	2	0.21	0.20	0.19	0.02	0.04	0.01
	HS	5	0.20	0.24	0.19	0.06	0.05	0.09
STE	5	4	0.23	0.15	0.23	0.03	0.08	0.14
	8	4	0.28	0.20	0.23	0.08	0.08	0.02
Biology	HS	4	0.22	0.22	0.12	0.01	0.04	0.02
Chemistry	HS	3	-0.10	-0.04	-0.01	0.01	0.09	0.05
Introductory Physics	HS	3	0.18	0.54	0.01	0.07	0.03	0.08
Technology/ Engineering	HS	3	0.42	0.20	0.21	0.08	0.15	0.14

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

The average correlations range from very weak (0.00 to 0.20) to weak (0.20 to 0.40) among the Level of Complexity and Independence dimensions; from very weak to moderate (0.40 to 0.60) among the Level of Complexity and Demonstration of Skills and Concepts dimensions; and from

very weak to weak among the Independence and Demonstration of Skills and Concepts dimensions. It is important to remember in interpreting the information in Table 4-8 that the correlations are based on small numbers of item scores and small numbers of students and should, therefore, be used with caution.

4.5.4 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., 2014) includes similar guidelines.

When appropriate, the standardization differential item functioning (DIF) procedure (Dorans & Kulick, 1986) is 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. 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. Conducting DIF analyses using groups of fewer than 200 students would result in inflated type I error rates.

4.6 Bias/Fairness

Fairness is addressed through the portfolio development and assembly processes, and in the development of the standards themselves, which have been thoroughly vetted for bias and sensitivity. The *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities* provides instructional and assessment strategies for teaching students with disabilities the same learning standards (by grade level) as general education students. The *Resource Guide* is intended to promote access to the general curriculum, as required by law, and to assist educators in planning instruction and assessment for students with significant cognitive disabilities. It was developed by panels of education experts in each content area, including ESE staff, testing contractor staff, higher education faculty, MCAS Assessment Development Committee members, curriculum framework writers, and regular and special educators. Each section was written, reviewed, and validated by these panels to ensure that each modified standard (entry point) embodied the essence of the grade-level learning standard on which it was based and that entry points at varying levels of complexity were aligned with grade-level content standards.

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

Issues of fairness are also addressed in the portfolio scoring procedures. Rigorous scoring procedures hold scorers to high standards of accuracy and consistency using monitoring methods that include frequent double-scoring, monitoring, and recalibrating to verify and validate portfolio scores. These procedures, along with the ESE's review of each year's MCAS-Alt results, indicate that the MCAS-Alt is being successfully used for the purposes for which it was intended. Section 4.4 describes in

greater detail the scoring rubrics used, selection and training of scorers, and scoring quality-control procedures. These processes ensure that bias due to differences in how individual scorers award scores is minimized.

4.7 Characterizing Errors Associated With Test Scores

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

4.7.1 MCAS-Alt Reliability

In the previous section, individual item characteristics of the 2016 MCAS-Alt were presented. Although individual item performance is an important focus for evaluation, a complete evaluation of an assessment must also address the way in which items function together and complement one another. Any assessment includes some amount of measurement error; that is, no measurement is perfect. This is true of all academic assessments—some students will receive scores that underestimate their true ability, and others will receive scores that overestimate their true ability. When tests have a high amount of measurement error, student scores are very unstable. Students with high ability may get low scores and vice versa. Consequently, one cannot reliably measure a student's true level of ability with such a test. Assessments that have less measurement error (i.e., errors are small on average, and therefore students' scores on such tests will consistently represent their ability) are described as reliable.

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

The split-half method requires psychometricians to select items that contribute to each half-test score. This decision may have an impact on the resulting correlation, since each different possible split of the test into halves will result in a different correlation. Another problem with the split-half method of calculating reliability is that it underestimates reliability, because test length is cut in half. All else being equal, a shorter test is less reliable than a longer test. Cronbach (1951) provided a statistic, alpha (α), that 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 2016 MCAS-Alt. The formula is as follows:

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

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

Table 4-9 presents Cronbach's α coefficient and raw score standard errors of measurement (SEMs) for each content area and grade.

Table 4-9. 2016 MCAS-Alt: Cronbach's Alpha and SEMs by Content Area and Grade

Content Area	Grade	Number of Students	Raw Score			Alpha	SEM
			Maximum Score	*Mean*	*Standard Deviation*		
ELA	3	1131	39	29.38	3.23	0.63	1.96
	4	1217	39	29.53	3	0.6	1.9
	5	1203	39	29.44	3.31	0.64	1.98
	6	1097	39	29.33	3.44	0.58	2.24
	7	1119	39	29.26	3.67	0.64	2.2
	8	1043	39	29.22	3.56	0.68	2.01
	HS	828	39	28.61	3.71	0.69	2.07
Mathematics	3	1064	26	21.47	1.31	0.72	0.7
	4	1154	26	21.49	1.11	0.62	0.68
	5	1132	26	21.51	1.23	0.7	0.68
	6	1080	26	21.53	1.12	0.66	0.65
	7	1067	26	21.58	1.13	0.66	0.66
	8	1020	26	21.43	1.27	0.68	0.72
	HS	810	39	31.07	3.3	0.88	1.16
STE	5	1103	39	31.89	2.6	0.83	1.08
	8	1008	39	31.63	2.68	0.83	1.1
Biology	HS	664	39	31.35	2.93	0.77	1.4
Chemistry	HS	34	39	30.56	2.5	0.55	1.67
Introductory Physics	HS	51	39	30.16	4.33	0.85	1.67
Technology/ Engineering	HS	86	39	30.52	3.53	0.84	1.42

An alpha coefficient toward the high end (greater than 0.50) is taken to mean that the items are likely measuring very similar knowledge or skills; that is, they complement one another and suggest a reliable assessment.

4.7.2 Subgroup Reliability

The reliability coefficients discussed in the previous section were based on the overall population of students who participated in the 2016 MCAS-Alt. Appendix P presents reliabilities for various subgroups of interest. Subgroup Cronbach's α coefficients were calculated using the formula defined on the previous page based only on the members of the subgroup in question in the computations; values are calculated only for subgroups with 10 or more students.

For several reasons, the results documented in this section should be interpreted with caution. First, inherent differences between grades and content areas preclude making valid inferences about the quality of a test based on statistical comparisons with other tests. Second, reliabilities are dependent not only on the measurement properties of a test but also on the statistical distribution of the studied subgroup. For example, it can be readily seen in Appendix P that subgroup sample sizes may vary considerably, which results in natural variation in reliability coefficients. Moreover 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.

4.7.3 Interrater Consistency

Section 4.4 of this chapter describes the processes that were implemented to monitor the quality of the hand-scoring of student responses. One of these processes was double-blind scoring of at least 20% of student responses in all portfolio strands. Results of the double-blind scoring, used during the scoring process to identify scorers who required retraining or other intervention, are presented here as evidence of the reliability of the MCAS-Alt. A third score was required for any score category in which there was not an exact agreement between scorer one and scorer two. A third score was also required as a confirmation score when either scorer one and/or scorer two provided a score of M for Demonstration of Skills and Concepts and Independence or a score of 1 for Level of Complexity. A summary of the interrater consistency results is presented in Table 4-10. Results in the table are aggregated across the tasks by content area, grade, and number of score categories (five for Level of Complexity and four for Demonstration of Skills and Concepts and Independence). The table shows the number of items, number of included scores, percent exact agreement, percent adjacent agreement, correlation between the first two sets of scores, and the percent of responses that required a third score. This information is also provided at the item level in Appendix O.

Table 4-10. 2016 MCAS-Alt: Summary of Interrater Consistency Statistics Aggregated Across Items by Content Area and Grade

Content Area	Grade	Number of			Percent		Correlation	Percent of Third Scores
		Items	*Score Categories*	*Included Scores*	*Exact*	*Adjacent*		
ELA	3	6	4	968	97.62	1.65	0.97	5.27
		3	5	549	96.90	1.28	0.66	8.38
	4	6	4	1,766	96.94	2.27	0.96	5.55
		3	5	974	98.67	0.72	0.81	7.19
	5	6	4	1,900	97.42	2.11	0.97	5.89
		3	5	1,101	96.91	1.54	0.64	8.90
	6	6	4	884	97.74	1.70	0.96	4.64
		3	5	520	97.12	1.54	0.64	7.31
	7	6	4	1,424	97.54	1.62	0.96	4.92
		3	5	851	97.88	0.82	0.69	6.11
	8	6	4	1,470	98.16	1.56	0.98	5.31
		3	5	850	97.41	1.06	0.65	8.24
	HS	6	4	1,814	96.47	2.76	0.96	7.77
		3	5	1,127	96.81	0.80	0.66	12.24
Mathematics	3	4	4	630	99.37	0.63	0.98	2.38
		2	5	364	97.80	0.27	0.57	2.47
	4	4	4	1,176	99.23	0.77	0.97	1.36
		2	5	636	99.37	0.63	0.94	0.79
	5	4	4	1,302	99.16	0.84	0.96	1.23
		2	5	758	98.42	0.66	0.66	1.98
	6	4	4	640	99.84	0.16	0.99	0.16
		2	5	354	98.59	1.13	0.85	1.69
	7	4	4	960	99.38	0.63	0.96	1.77
		2	5	577	98.96	0.35	0.77	1.91

continued

Content Area	Grade	Number of			Percent		Correlation	Percent of Third Scores
		Items	*Score Categories*	*Included Scores*	*Exact*	*Adjacent*		
Mathematics	8	4	4	1,052	99.81	0.19	0.99	0.76
		2	5	589	98.13	1.02	0.74	3.40
	HS	10	4	1,894	99.68	0.32	0.99	1.43
		5	5	1,122	98.66	0.45	0.82	4.01
STE	5	8	4	1,858	99.41	0.59	0.97	0.91
		4	5	1,018	99.41	0.59	0.94	2.06
	8	8	4	1,466	99.25	0.75	0.96	1.50
		4	5	816	98.90	0.49	0.74	1.59
Biology	HS	8	4	1,408	98.65	1.35	0.96	2.34
		4	5	829	98.07	0.97	0.67	3.86
Chemistry	HS	6	4	80	92.50	7.50	0.85	15.00
		3	5	48	97.92	0.00		6.25
Introductory Physics	HS	6	4	106	99.06	0.94	0.99	1.89
		3	5	60	100.00	0.00	1.00	0.00
Technology/ Engineering	HS	6	4	184	100.00	0.00	1.00	0.00
		3	5	106	97.17	1.89	0.82	6.60

4.8 MCAS-Alt Comparability Across Years

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

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

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

Table 4-11. 2016 MCAS-Alt Achievement Level Descriptions

Achievement Level	Description
Incomplete (1)	**Insufficient evidence** and information were 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 4-12. 2016 MCAS-Alt: Strand Achievement Level Look-Up

Level of Complexity	Demonstration of Skills	Independence	Achievement Level
2	1	1	1
2	1	2	1
2	1	3	1
2	1	4	1
2	2	1	1
2	2	2	1
2	2	3	1
2	2	4	1
2	3	1	1
2	3	2	1
2	3	3	2
2	3	4	2
2	4	1	1
2	4	2	1
2	4	3	2
2	4	4	2
3	1	1	1
3	1	2	1
3	1	3	1

continued

Level of Complexity	Demonstration of Skills	Independence	Achievement Level
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

4.9 Reporting of Results

4.9.1 Primary Reports

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

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

4.9.1.1 Portfolio Feedback Forms

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

4.9.1.2 Parent/Guardian Report

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

Two printed copies of each report are provided: one for the parent/guardian and one to be kept in the student's temporary record. A sample report is provided in Appendix S.

The *Parent/Guardian Report* was redesigned in 2012, with input from parents in two focus groups, to include information that had previously been published in a separate interpretive guide, which is no longer produced.

4.9.2 Decision Rules

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

4.9.3 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 data perform quality-control checks of their respective computer programs. Moreover, when data are handed off to different units within the Data and Reporting Services (DRS) Department, the sending unit verifies that the data are accurate before handoff. Additionally, when a unit receives a data set, the first step is to verify the accuracy of the data.

Quality assurance is also practiced through parallel processing. One data analyst is responsible for writing all programs required to populate the student and aggregate reporting tables for the

administration. Each reporting table is assigned to another data analyst who uses the decision rules to independently program the reporting table. The production and quality-assurance tables are compared; if there is 100% agreement, the tables are released for report generation.

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

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

4.10 MCAS-Alt Validity

One purpose of the *2016 MCAS and MCAS-Alt Technical Report* is to describe the technical aspects of the MCAS-Alt that contribute validity evidence in support of MCAS-Alt score interpretations. According to the *Standards for Educational and Psychological Testing* (AERA et al., 2014), considerations regarding establishing intended uses and interpretations of test results and conforming to these uses are of paramount importance in regard to valid score interpretations. These considerations are addressed in this section.

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

4.10.1 Test Content Validity Evidence

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

4.10.2 Internal Structure Validity Evidence

Evidence based on internal structure is presented in detail in the discussions of item analyses and reliability in sections 4.5 and 4.7. Technical characteristics of the internal structure of the assessment are presented in terms of classical item statistics (item difficulty and item-test correlation), correlations among the dimensions (Level of Complexity; Demonstration of Skills and Concepts; and Independence), fairness/bias, and reliability, including alpha coefficients and interrater consistency.

4.10.3 Response Process Validity Evidence

Response process validity evidence pertains to information regarding the cognitive processes used by examinees as they respond to items on an assessment. The basic question posed is: Are examinees responding to the test items as intended?

The MCAS-Alt directs educators to identify measurable outcomes for students based on the state's curriculum frameworks, and to collect data and work samples that document the extent to which the student engaged in the intended cognitive process(es) to meet the intended goal. The portfolio scoring process is intended to confirm the student's participation in instructional activities that were focused on meeting the measurable outcome, and to provide detailed feedback on whether the instructional activities were sufficient in duration and intensity for the student to meet the intended goal.

4.10.4 Efforts to Support the Valid Reporting and Use of MCAS-Alt Data

The assessment results of students who participate in the MCAS-Alt are included in all public reporting of MCAS results and in the state's accountability system.

In an effort to ensure that all students were provided access to the Massachusetts curriculum frameworks, the Department, and federal and state law, require that all students in grades 3–8 and 10 are assessed each year on their academic achievement and that all students appear in the reports provided to parents, guardians, teachers, and the public. The alternate assessment portfolio ensures that students with the most intensive disabilities have an opportunity to "show what they know" and receive instruction at a level that is challenging and attainable based on the state's academic learning standards. Annual state summaries of the participation and achievement of students on the MCAS-Alt are available at www.doe.mass.edu/mcas/alt/results.html.

In the past, it was not always possible to determine what had been taught and whether special education had been successful for students with disabilities; nor was it possible to compare outcomes among students and across programs, schools, and districts. One important reason to include students with significant disabilities in standards-based instruction is to explore their capacity to learn standards-based knowledge and skills. While "daily living skills" are critical for these students to function as independently as possible, academic skills are extremely important. Standards in the Massachusetts curriculum frameworks are defined as "valued outcomes for all students." Evidence indicates that students with significant disabilities learn more than anticipated when given opportunities to engage in challenging instruction with the necessary support.

As a result of taking the MCAS-Alt, students with significant disabilities have become more "visible" in their schools, and have a greater chance of being considered when decisions are made to allocate staff and resources to improve their academic achievement.

For state and federal accountability reporting, alternately assessed students can receive the full range of Composite Performance Index (CPI) points under the Massachusetts accountability plan approved by the U.S. Department of Education. According to the plan, up to 1% of students with significant cognitive disabilities may be counted "as if *Proficient*" and receive up to 100 CPI points toward their school's and district's total number of accountability points. Table 4-13 indicates how CPI points are awarded to each school and district in ELA and mathematics based on MCAS and MCAS-Alt scores for students with and without disabilities.

Table 4-13. 2016 MCAS-Alt: Composite Proficiency Index

Students taking standard MCAS tests			Students with significant disabilities taking MCAS-Alt	
MCAS Scaled Score	*Achievement Level*	*CPI Points Awarded*	*MCAS-Alt Achievement Level*	*CPI Points Awarded*
240–280	Proficient and Advanced	100	Progressing (for certain disability types)[1]	100
230–238	Needs Improvement – High	75	Progressing (for certain disability types)[2] and Emerging	75
220–228	Needs Improvement – Low	50	Awareness	50
210–218	Warning/Failing – High	25	Portfolio Incomplete	25
200–208	Warning/Failing – Low	0	Portfolio Not Submitted	0

[1] The ESE will assign 100 CPI points only to students scoring *Progressing* who have been identified through the Student Information Management System as having a High Level of Need and one of the following primary disabilities: Intellectual, Multiple Disabilities, Autism, Developmental Delay, or Sensory/Deaf and Blind.

[2] The ESE will assign 75 CPI points to students scoring *Progressing* if they are either reported in the above disability categories, but with lower levels of need, or reported as having one of the following primary disabilities: Sensory/Hard of Hearing or Deaf, Communication, Sensory/Vision Impairment or Blind, Emotional, Physical, Health, Specific Learning Disabilities, or Neurological.

Although students taking the MCAS-Alt can receive the full range of CPI points for accountability reporting, students with significant cognitive disabilities are included in the *Warning/Failing* achievement level for the reporting of MCAS school and district results. This is because the students are working on learning standards that have been modified below grade level, and because the students have not yet attained grade-level proficiency.

Typically, students who participate in the MCAS-Alt do not meet the state's graduation requirement. However, a small number of students who are working on learning standards at grade level may eventually attain scores of *Needs Improvement*, *Proficient*, or *Advanced* (and, in high school, may earn a CD) if the portfolio includes evidence that is comparable to the level of work attained by students who have earned a score of *Needs Improvement* or higher on the standard MCAS test in the content area.

Appendix S shows the report provided to parents and guardians for students assessed on the MCAS-Alt. The achievement level descriptors on the first page of that report describe whether the student's portfolio was based on grade-level standards or standards that were modified below grade level.

4.10.5 Summary

The evidence for validity and reliability presented in this chapter supports the use of the assessment to make inferences about student achievement of the skills and content described in the Massachusetts curriculum frameworks for ELA, mathematics, and STE. As such, this evidence

supports the use of MCAS-Alt results for the purposes of programmatic and instructional improvement and as a component of school accountability.

REFERENCES

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Baker, F. B. (1992). *Item Response Theory: Parameter Estimation Techniques.* New York, NY: Marcel Dekker, Inc.

Baker, F. B., & Kim, S. H. (2004). *Item Response Theory: Parameter Estimation Techniques* (2nd ed.). New York, NY: Marcel Dekker, Inc.

Brown, F. G. (1983). *Principles of Educational and Psychological Testing* (3rd ed.). Fort Worth, TX: Holt, Rinehart and Winston.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology 3*, 296–322.

Chicago Manual of Style (16th ed.). (2003). Chicago: University of Chicago Press.

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

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

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

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

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

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

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston, MA: 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, NY: Springer-Verlag.

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

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

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement 43*(4), 355–381.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: 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.

Massachusetts Department of Elementary and Secondary Education. (2016). *Representative Samples and PARCC to MCAS Concordance Studies*.

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

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

Nering, M., & Ostini, R. (2010). *Handbook of Polytomous Item Response Theory Models*. New York, NY: 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, NY: Macmillan Publishing Company.

Roussos, L. A., & Ozbek, O. Y. (2006). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. *Journal of Educational Measurement 43,* 215–243.

Spearman, C. C. (1910). Correlation calculated from faulty data. *British Journal of Psychology 3,* 271–295.

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, NY: 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