MASSACHUSETTS DEPARTMENT OF
ELEMENTARY AND SECONDARY
EDUCATION

2017 Legacy MCAS Technical Report

measured progress

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
Jeffrey C. Riley
Commissioner

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

TABLE OF CONTENTS

Chapter 1 Overview

1.1 Purposes of the MCAS

The Massachusetts Education Reform Mandate

The Massachusetts Comprehensive Assessment System (MCAS) was developed in response to provisions in the Massachusetts Education Reform Act of 1993, which established greater and more equitable funding to schools, accountability for student learning, and statewide standards and assessments for students, educators, schools, and districts. 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 learner (EL) students;
- measure performance based on the learning standards in the Massachusetts curriculum frameworks (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.

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

- measure student, school, and district performance in meeting the state's learning standards as detailed in the Massachusetts curriculum frameworks;
- provide measures of student achievement that will lead to improvements in student outcomes; 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 *2017 Legacy MCAS Technical Report* is to document the technical quality and characteristics of the legacy MCAS operational tests that were administered in 2017: the grade 10 ELA and mathematics tests, and the science and technology/engineering (STE) tests in grade 5, grade 8, and high school. The report presents evidence of the validity and reliability of test score interpretations, and describes modifications made to the MCAS program in 2017. A companion document, the *2017 Next-Generation MCAS and MCAS-Alt Technical Report*, provides information regarding the next-generation MCAS tests administered in 2017 in grades 3–8 ELA and mathematics.

Technical reports for previous testing years are available on the ESE website at www.doe.mass.edu/mcas/tech/?section=techreports. The 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 2017 legacy 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 2017 legacy 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 2017. Chapter 2 explains the guiding philosophy, purpose, uses, components, and validity of MCAS. Chapter 3 covers the test design and development, test administration, scoring, and analysis and reporting of results for the MCAS assessment. This chapter includes 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. The appendices, which appear after Chapter 3, are referenced throughout the report.

1.4 Current Year Updates

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 include innovative items developed by the Partnership for Assessment of Readiness for College and Careers (PARCC).

The 2017 MCAS assessments marked the beginning of the transition from the legacy MCAS tests (administered from 1998 to 2016) to the next-generation MCAS tests. Next-generation tests were administered for the first time in ELA and mathematics at grades 3–8. Because of this transition, ESE has published two separate technical reports for 2017. This document focuses on

the legacy MCAS assessments administered in grade 10 ELA and mathematics, grades 5 and 8 STE, and high school STE.

Technical information about the next-generation MCAS assessments is documented in the *2017 Next-Generation MCAS and MCAS-Alt Technical Report*. Additional information on the next-generation MCAS assessments is available at www.doe.mass.edu/mcas/nextgen/resources.html.

Chapter 2 The State Assessment System: Legacy MCAS

2.1 Guiding Philosophy

The MCAS program plays 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 with the introduction of next-generation tests.

2.2 Alignment to the Massachusetts Curriculum Frameworks

All items included on the MCAS tests are developed to measure the standards contained in the Massachusetts curriculum frameworks. Each test item correlates and is aligned to at least one standard in a curriculum framework. All learning standards defined in the frameworks are addressed by and incorporated into the local curriculum and instruction, whether or not they are assessed on MCAS.

2.3 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

2.4 Validity of MCAS

Validity information for the MCAS is provided throughout this technical report. Validity evidence includes information on test design and development; administration; scoring; technical evidence of test quality (classical item statistics, differential item functioning, item response theory statistics, reliability, dimensionality, decision accuracy and consistency); and reporting. Validity information is described in detail in section 3.9 of this report.

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 Legacy Test Design and Development

The 2017 legacy MCAS test administration included operational tests in the following grades and content areas:

- grade 10 ELA, including a reading comprehension component and a composition component
- grade 10 mathematics
- grades 5 and 8 STE
- high school STE end-of-course tests in biology, chemistry, introductory physics, and technology/engineering

The 2017 MCAS administration also included retest opportunities in ELA and mathematics in November 2016 and March 2017 for students beyond grade 10 who had not yet passed the standard grade 10 tests. A February 2017 biology test was also administered. This test could be taken as a retest or as a first experience of MCAS STE for transfer students or students in block-scheduled science classes who completed their biology class in January.

3.2.1 Test Specifications

3.2.1.1 Criterion-Referenced Test

Items used on the MCAS are developed specifically for Massachusetts and are aligned 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 skills described in the Massachusetts curriculum frameworks. In 2011, Massachusetts adopted new curriculum standards in mathematics and ELA. In 2012–2017, all legacy items were double-coded to the 2000 standards and the 2011 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 item types used in the legacy MCAS tests. The types of items and their functions are described below.

- **Multiple-choice** items are used to provide breadth of coverage within a content area. 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). 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.
- **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. 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.

3.2.1.3 Description of Test Design

The MCAS assessments 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.

Grades 5 and 8 and High School STE Tests

The matrix portions of the STE tests are composed of both equating and field-test items that do not count toward student scores. Equating items are used to link one year's results to those of previous years. Field-test items are also included in the matrix portion of the tests. An item is field tested to determine how it performs to help determine if it should be used as a future common item. The number of test forms varies by test between 1 and 15 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 1,500 students per item, with the exception of the high school technology/engineering test) is obtained to produce reliable data that can be used to inform item selection for future tests. The technology/engineering test sample size is approximately 500.

Grade 10 ELA and Mathematics Tests

The matrix portions of the ELA and mathematics grade 10 tests are composed of equating items that are used to link one year's results to those of previous years. Typically, the matrix items are composed of both field-test and equating items; however due to the transition to the next-generation MCAS tests, no field-test items are part of the matrix portion of the grade 10 ELA and mathematics tests. There are three forms for the ELA test and three forms for the mathematics test.

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 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.

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, 2017, the MCAS ELA assessments were aligned to both the 2001/2004 standards and the 2011 standards listed above. 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 2016 and March 2017 ELA retests were aligned to both the 2001/2004 and 2011 Massachusetts ELA standards.

3.2.2.2 Item Types

The reading comprehension portion of the grade 10 ELA test uses a mix of multiple-choice and open-response items. Additionally, grade 10 students take a composition test as part of their ELA test administration.

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. 2017 Legacy MCAS: ELA Item Types and Score Points

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

3.2.2.3 Test Design

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

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 2017 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 3-2 lists the distribution of ELA common and matrix items in the spring grade 10 test and in the retests.

Table 3-2. 2017 Legacy MCAS: Distribution of Grade 10 ELA Common Items by Item Type

Grade and Test			Items per Form						Total Matrix Positions Across Forms					
			Common			Matrix			Equating Positions			Field-Test Positions		
Grade	*Test*	*# of Forms*	*MC*	*OR*	*WP*	*MC*	*OR*	*WP*	*MC*	*OR*	*WP*	*MC*	*OR*	*WP*
10	Reading Comprehension	3	36	4		12	2		36[b]	6[b]		0[c]	0[c]	
	Composition	2[a]			1									
Retest[d]	Reading Comprehension	1	36	4										
	Composition	1			1									
	Reading Comprehension	1	36	4										
	Composition	1			1									

[a] The ELA composition is field-tested out of state.
[b] The grade 10 ELA test is pre-equated; however, in 2017, because of the 2015 rescaling of ELA items, equating items were added to the test for the fourth year in a row.
[c] Items were not field-tested in 2017.
[d] ELA retests consist of common items only.
Key: MC = Multiple Choice, OR = Open Response, WP = Writing Prompt

3.2.2.4 Blueprints

Table 3-3 shows the percentage of common item points by reporting category. The reporting categories are aligned to the Massachusetts ELA curriculum framework strands.

Table 3-3. 2017 Legacy MCAS: Target (and Actual) Percentage of ELA Item Points by Reporting Category for Grade 10 ELA

Reporting Category	% of Points
Language	8 (6)
Reading	64 (66)
Writing	28 (28)
Total	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 learner (EL) 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 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 2017 ELA composition writing prompts are not associated with a specific reading passage.

3.2.3 Mathematics Test Specifications

3.2.3.1 Standards

The items on the 2017 grade 10 mathematics assessment were aligned to 2011 standards that matched content in the 2000/2004 standards.

The 2011 standards are grouped by conceptual categories at the high school level.

High School Conceptual Categories

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

3.2.3.2 Item Types

The grade 10 mathematics test 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. Each type of item is worth a specific number of points in the student's total mathematics score, as shown in Table 3-4.

Table 3-4. 2017 Legacy MCAS: Mathematics Item Types and Score Points

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

3.2.3.3 Test Design

The mathematics tests typically comprise common and matrix items. The matrix slots in each test form are used to equate the current year's test to that of previous years by using previously administered items. Table 3-5 presents the distributions of Mathematics common and matrix items by grade and item type for the 2017 legacy MCAS assessment.

Table 3-5. 2017 Legacy MCAS: Distribution of Grade 10 Mathematics Common and Matrix Items by Item Type

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*[a]	*MC*	*SA*	*OR*
10	3	32	4	6	7	1	2	21	3	6	21[a]	3[a]	4[a]	0	0	0
Retest[b]	1	32	4	6												
	1	32	4	6												

[a] Not all equating items are unique. The grade 10 mathematics test is pre-equated. However, in 2017, because of the 2015 rescaling of items, equating items were added to the test.

[b]Mathematics retests consist of common items only.

Key: MC = Multiple Choice, SR = Selected Response, OR = Open Response

3.2.3.4 Blueprints

Table 3-6 shows the distribution of common item points in the grade 10 mathematics test across the strands of the *2000 Massachusetts Mathematics Framework*. Table 3-7 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 3-6. 2017 Legacy MCAS: Mathematics Common Point Distribution by *2000 Mathematics Framework* Strand, Grade 10

Reporting Category	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 3-7. 2017 Legacy MCAS: Target (and Actual) Mathematics Common Point Distribution by Reporting Category, Grade 10*

Reporting Category	Percent of Raw Score Points
Number and Quantity	20 (22)
Algebra and Functions	30 (30)
Geometry	30 (28)
Statistics and Probability	20 (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 grade 10 mathematics test 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 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 grade 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.

3.2.4 Science and Technology/Engineering 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 3-11 through 3-14.

3.2.4.2 Item Types

The STE tests include multiple-choice and open-response items. Each type of item is worth a specific number of points in the student's total test score, as shown in Table 3-8.

Table 3-8. 2017 Legacy 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 common items, which are taken by all students, and a set of matrix items. Table 3-9 lists the distribution of common and matrix items in each STE test.

Table 3-9. 2017 Legacy 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	19	38	4	3	1	19	2	38	17
8	STE	19	38	4	3	1	19	2	38	17
HS	Biology	10	40[b]	5[b]	12	2	NA[c]	NA[c]	120[d]	20[d]
	Chemistry	1	40	5	20	2	NA[c]	NA[c]	20	2
	Introductory Physics	5	40	5	12	2	NA[c]	NA[c]	60	10
	Technology/Engineering	5	40	5	20	2	NA[c]	NA[c]	60	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.
Key: MC = Multiple Choice, OR = Open Response

3.2.4.4 Blueprints

Grades 5 and 8

Table 3-10 shows the distribution of common items across the four strands of the *2006 Massachusetts Science and Technology/Engineering Curriculum Framework*.

Table 3-10. 2017 Legacy MCAS: Target (and Actual) STE Common Point Distribution by Reporting Category and Grade

Reporting Category	% for Grade 5	% for Grade 8
Earth and Space Science	30 (30)	25 (26)
Life Science	30 (30)	25 (26)
Physical Sciences	25 (25)	25 (24)
Technology/Engineering	15 (15)	25 (24)
Total	100	100

High School

Tables 3-11 through 3-14 show the distribution of common items across the reporting categories for the MCAS high school STE tests. All numbers listed are both target and actual percentages.

Table 3-11. 2017 Legacy MCAS: High School Biology Common Point Distribution by Reporting Category

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 3-12. 2017 Legacy MCAS: High School Chemistry Common Point Distribution by Reporting Category

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 3-13. 2017 Legacy MCAS: High School Introductory Physics Common Point Distribution by Reporting Category

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 3-14. 2017 Legacy MCAS: High School Technology/Engineering Common Point Distribution by Reporting Category

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 3-15 describes the cognitive skills used for the STE test items.

Table 3-15. 2017 Legacy 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. Calculators, formula sheets, and rulers are not allowed or used on the STE tests in grades 5 and 8.

3.2.5 Item and Test Development Process

Table 3-16 provides a high-level view of the item and test development process in chronological order.

Table 3-16. 2017 Legacy MCAS: Overview of Test Development Process

Development Step	Detail of the Process
Select reading passages (for ELA only)	Contractor's content specialists find potential passages and present them to ESE 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 content specialists 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 content specialists for review. 2. ESE content specialists review and edit items prior to presenting the items to ADCs. 3. ADCs review items and make recommendations. 4. Bias and Sensitivity Committee reviews items and makes recommendations. 5. ESE content specialists 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 content specialists 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 content specialists 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 content specialists. The common set of items is delivered with proposed cut scores, including Test Characteristic Curves (TCCs) and Test Information Functions (TIFs). ESE content specialists and editorial staff review and edit proposed sets of items. Contractor and ESE content specialists 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	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. Content

specialists 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. Content specialists 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 specialists 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 content specialists, in consultation with Measured Progress content specialists, 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 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 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 ESE and 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 1,500 students (except where noted), 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 1,500 student responses are scored per constructed-response field-tested item. As with the multiple-choice items, 1,500 student responses are sufficient to provide reliable performance data. 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 content specialists review 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
- Differential Item Functioning (DIF)
- distribution of scores across answer options and score points
- distribution of answer options and score points across quartiles

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
- EL and former EL who have been transitioned out of EL for fewer than two years/native English speakers and former EL who have been transitioned from EL for two or more years (for mathematics and STE only)

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 Review 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.
- **"Clueing" items.** Items are reviewed for any information that might "clue" 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 content specialists 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 must have a disability that is documented in an individualized education program (IEP) or a 504 plan. All 2017 MCAS 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 indicates bias toward students with visual disabilities (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 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 ESE 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** – Test versions are offered on a CD for students with disabilities who require a read-aloud function using locally installed Kurzweil-3000 software. This edition contains only the common items found in the operational test. The items are not modified and are read aloud to the student as they appear in the standard test booklet. For items or passages that include graphics, the captions and words in the graphics are read aloud verbatim to the student. Students typically use headphones with this format, but may also be tested individually in a separate setting to minimize distractions to other students from reading aloud through a speaker.

- **American Sign Language DVD edition** – 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 A 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 those 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 EL 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 2017:

- March–April
  - Grade 10 ELA
- April–May
  - Grades 5 and 8 STE

- May
  - Grade 10 mathematics
- June
  - High school end-of-course STE
    - biology
    - chemistry
    - introductory physics
    - technology/engineering

The 2017 MCAS administration also included retest opportunities in ELA and mathematics for students in grades 11 and 12 and former students who exited high school and who did not previously pass one or both grade 10 tests. Retests were offered in November 2016 and March 2017.

An additional high school biology test was administered in February 2017. Table 3-17 shows the complete 2016–2017 MCAS test administration schedule. Former students were also eligible to participate in the February biology administration, as well as in one of the four tests administered in June.

Table 3-17. 2017 Legacy MCAS: Test Administration Schedule

Grade and Content Area	Test Administration Date(s)	Deadline for Return of Materials to Contractor
Retest Administration Windows		
November 2–10, 2016		
ELA Composition Retest	November 2	November 17
ELA Reading Comprehension Retest Sessions 1 and 2 Session 3	 November 3 November 4	
Mathematics Retest Session 1 Session 2	 November 9 November 10	
March 1–7, 2017		
ELA Composition Retest	March 1	March 10
ELA Reading Comprehension Retest Sessions 1 and 2 Session 3	 March 2 March 3	
Mathematics Retest Session 1 Session 2	 March 6 March 7	
March–April 2017 Test Administration Window		
Grade 10 ELA Composition	March 21	April 5
Grade 10 ELA Reading Comprehension Sessions 1 and 2 Session 3	 March 22 March 23	
Grade 10 ELA Composition Make-Up	March 30	
May 2017 Test Administration Window		
Grades 5 and 8 STE	April 5–May 26	May 31
Grade 10 Mathematics Session 1 Session 2	 May 16 May 17	May 25
High School End-of-Course STE Test Administration Windows		
February 6–7, 2017		
Biology Session 1 Session 2	 February 6 February 7	February 13
June 5–6, 2017		
STE (Biology, Chemistry, Introductory Physics, Technology/Engineering) Session 1 Session 2	 June 5 June 6	June 12

3.3.2 Security Requirements

Principals were responsible for ensuring that all test administrators complied with the requirements and instructions contained in the *Test Administrator's Manuals*. In addition, other administrators, educators, and staff within the school were 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 *Principal's Administration Manual*: *High School Spring 2017,* the *Principal's Administration Manual: Grades 3–8 Spring 2017,* the *Fall 2016/Winter 2017 Principal's Administration Manual*, and all *Test Administrator's Manuals*.

3.3.3 Participation Requirements

In spring 2017, students educated with Massachusetts public funds were required by state and federal laws to participate in MCAS testing. The 1993 Massachusetts Education Reform Act mandates that **all** students in the tested grades who are educated with Massachusetts public funds participate in the MCAS, including the following groups of students:

- students enrolled in public schools
- students enrolled in charter schools
- students enrolled in 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
- English learner (EL) 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 was the responsibility of the principal to ensure that all enrolled students participated in testing as mandated by state and federal laws. To certify that **all** students participated in testing as required, principals were required to complete the online Principal's Certification of Proper Test Administration (PCPA) following each test administration. See Appendix B for a summary of participation rates.

3.3.3.1 Students Not Tested on Standard Tests

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

- EL students in their first year of enrollment in U.S. schools, who were not required to participate in ELA testing
- students with significant disabilities who instead participated in the MCAS-Alt. See the *2017 Next-Generation MCAS and MCAS-Alt Technical Report* for details.
- students with a medically documented absence who were 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 could, in lieu of having the student participate in MCAS retests, submit to the ESE alternative evidence or information that demonstrated that the student has met the CD graduation standard in each required content area)

More details about test administration policies and student participation requirements (including requirements for students with disabilities, EL students, and students educated in alternate settings), can be found in the *Principal's Administration Manual: High School Spring 2017*, the *Principal's Administration Manual: Grades 3–8 Spring 2017*, and the *Fall 2016/Winter 2017 Principal's Administration Manual*.

3.3.4 Administration Procedures

It was the principal's responsibility to coordinate the school's 2017 MCAS test administration. This coordination responsibility included the following responsibilites:

- 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 2017 testing can be found in Appendix C. For a list of test accommodations, see Appendix A.)
- 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
- reading the Student Assessment Update emails throughout the year for important information
- providing the ESE with correct contact information to receive important notices 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 *Principal's Administration Manual: High School Spring 2017*, the

Principal's Administration Manual: Grades 3–8 Spring 2017, and the *Fall 2016/Winter 2017 Principal's Administration Manual*.

The MCAS program is supported by the MCAS Service Center, which includes a toll-free telephone line and email 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 Time), Monday through Friday.

3.4 Scoring

For paper-based tests, 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. For computer-based tests, images of the student answers were transferred to iScore from the test administration platform and sorted at the item level.

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 by scorers within each group. 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 can be scored in various locations. The location used to score the 2017 legacy MCAS tests is shown in Table 3-18.

Table 3-18. 2017 Legacy MCAS: Summary of Scoring Locations and Scoring Shifts

Measured Progress Scoring Center, Content Area	Grade(s)	Shift	Hours
Menands, NY			
ELA Composition	10	Day	8:00 a.m. – 4:30 p.m.
STE	5	Night	5:30 p.m. – 10:00 p.m.
STE	8	Day	8:00 a.m. – 4:30 p.m.

The following staff members were involved with scoring the 2017 MCAS responses:

- The **Scoring Project Manager (SPM)** was located in Dover, New Hampshire, and oversaw communication and coordination of MCAS scoring across all scoring sites, scheduling of activities, and oversight of contractual work.
- 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 the satellite scoring location providing logistical coordination.
- 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. 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 sample responses; 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 recruited by 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. Additionally, scorers assigned to high school items had to have either a degree related to the content area being scored, or two classes related to the content area being scored with demonstrated experience in scoring the content area.

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 3-19 summarizes the scorers' backgrounds across all scoring shifts at all scoring locations.

Table 3-19. 2017 Legacy MCAS: Summary of Scorers' Backgrounds Across Scoring Shifts and Scoring Locations

Education	Scorers		Leadership	
	Number	*Percent*	*Number*	*Percent*
Less than 48 college credits	0	0.00	0	0.00
Associate's degree/more than 48 college credits	35	8.35	0	0.00
Bachelor's degree	235	56.09	28	47.46
Master's degree/doctorate	149	35.56	31	52.54
Teaching Experience				
No teaching certificate or experience	204	48.69	23	39.98
Teaching certificate or experience	177	42.24	28	47.45
College instructor	38	9.07	8	13.56
Scoring Experience				
No previous experience as scorer	111	26.49	0	0.00
1–3 years of experience	143	34.13	8	13.56
3+ years of experience	165	39.38	51	86.44

3.4.2.4 Methodology for Scoring Polytomous Items

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

Table 3-20 provides a sample 4-point mathematics open-response scoring guide. It was one of the many different item-specific MCAS scoring guides used in 2017. The task associated with this scoring guide required students to design four different gardens, each with a different shape.

Table 3-20. 2017 Legacy 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 grades 5 and 8 STE tests, were double-blind scored.

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 3-21 outlines the rules for instances when the two read-behind or two double-blind scores were not identical (i.e., adjacent or discrepant).

Table 3-21. 2017 Legacy MCAS: Read-Behind and Double-Blind Resolution Charts

Read-Behind Scoring*

Scorer #1	*Scorer #2*	*Scoring Leadership Resolution*	*Final*
4	--	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	1	2	2
0	1	--	1
2	4	3	3
1	2	--	2
2	0	2	2

* 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.

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 had 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 didn't know 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 3-22. 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 E. These interrater consistency statistics are the result of the processes implemented to ensure valid and reliable hand-scoring of items.

Table 3-22. 2017 Legacy 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	10	4	64,941	78.95	20.74	0.69	0.51
		5	267,497	64.15	34.68	0.76	1.16
		6	64,941	76.05	23.66	0.73	0.51
Mathematics	10	2	275,106	99.06	0.94	0.98	0.00
		5	411,190	82.39	16.04	0.93	1.56
STE	5	5	28,020	70.33	27.45	0.87	2.22
	8	5	29,151	66.77	30.05	0.85	3.25
Biology	HS	5	252,891	73.42	24.67	0.89	1.94
Chemistry	HS	5	3,305	69.59	27.20	0.85	3.21
Introductory Physics	HS	5	67,767	72.88	25.21	0.88	1.92
Technology/ Engineering	HS	5	11,927	68.57	28.80	0.84	2.64

*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 2017. 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 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 3-23. Note that the statistics are presented for all items as well as

by item type (multiple-choice and open-response). The mean difficulty (*p*-value) and discrimination values shown in the table are within generally acceptable and expected ranges and are consistent with results obtained in previous administrations.

Table 3-23. 2017 Legacy 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	10	ALL	42	0.78	0.11	0.43	0.11
		MC	36	0.80	0.11	0.39	0.06
		OR	6	0.70	0.10	0.67	0.06
Mathematics	10	ALL	42	0.70	0.12	0.46	0.14
		MC	32	0.72	0.12	0.41	0.10
		OR	10	0.63	0.08	0.62	0.14
Science	5	ALL	42	0.70	0.12	0.37	0.10
		MC	38	0.72	0.10	0.34	0.07
		OR	4	0.49	0.07	0.61	0.06
	8	ALL	42	0.65	0.13	0.40	0.10
		MC	38	0.67	0.13	0.38	0.07
		OR	4	0.51	0.08	0.61	0.12
Biology	HS	ALL	45	0.73	0.13	0.43	0.10
		MC	40	0.77	0.09	0.40	0.07
		OR	5	0.48	0.13	0.66	0.04
Chemistry	HS	ALL	45	0.68	0.13	0.43	0.14
		MC	40	0.69	0.12	0.39	0.10
		OR	5	0.53	0.07	0.69	0.05
Introductory Physics	HS	ALL	45	0.68	0.11	0.44	0.12
		MC	40	0.70	0.10	0.41	0.07
		OR	5	0.53	0.08	0.72	0.03
Technology/ Engineering	HS	ALL	45	0.65	0.12	0.39	0.10
		MC	40	0.67	0.10	0.36	0.07
		OR	5	0.45	0.09	0.58	0.09

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 F 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 were 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 G; 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.[1]

For the 2017 MCAS administration, DIF analyses were conducted for the following subgroups:

- male/female
- white/black
- white/Hispanic
- not LEP-FLEP/LEP-FLEP

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

3.5.3 Dimensionality Analysis

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

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 high school ELA and mathematics, grade 5 and 8 science, and high school biology, chemistry, introductory physics, and technology/engineering tests administered during spring 2017. A total of eight tests were analyzed, and the results for these analyses are reported below, including a comparison with the results from 2015–16.

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.

[1] 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.

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 eight MCAS tests administered during spring 2017. The data for each grade were split into a training sample and a cross-validation sample. For high-school mathematics and ELA, there were over 69,000 students per test. For the science assessments, all the elementary and middle school administrations had over 69,000 students per test, while the high school administrations had over 50,500 for biology, over 14,000 for physics, over 2,500 for technology/engineering, and over 650 for chemistry. Because DIMTEST had an upper limit of 24,000 students, the training and cross-validation samples for the tests that had over 24,000 students were limited to 12,000 each, randomly sampled from the total sample. DETECT, on the other hand, had an upper limit of 500,000 students, so every training sample and cross-validation sample used all the available data. After randomly splitting the data into training and cross-validation samples, DIMTEST was applied to each dataset to see if the null hypothesis of unidimensionality would be rejected. DETECT was then applied to each dataset for which the DIMTEST null hypothesis was rejected in order to estimate the effect size of the multidimensionality.

3.5.3.1 DIMTEST Analyses

The results of the DIMTEST analyses indicated that the null hypothesis was rejected at a significance level of 0.01 for every dataset except for high school chemistry. The nonrejection for chemistry was likely due to the combined effects of the presence of weak multidimensionality (as evidenced in analyses from years prior to spring 2013) and small sample size (the sample size dropped from about 2,300 in spring 2008 to about 800 in spring 2016). Because strict unidimensionality is an idealization that almost never holds exactly for a given dataset, the statistical rejections in the DIMTEST results were not surprising. Indeed, because of the very large sample sizes (over 14,000) involved in six of the datasets, DIMTEST would be expected to be sensitive to even quite small violations of unidimensionality.

3.5.3.2 DETECT Analyses

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

Table 3-24. 2017 Legacy MCAS: Multidimensionality Effect Sizes by Grade and Content Area

Content Area	Grade	Multidimensionality Effect Size	
		2015–16	*2016–17*
STE	5	0.13	0.08
	8	0.13	0.08
	(Biology) HS	0.09	0.08
	(Chemistry) HS	0.09	0.07
	(Introductory Physics) HS	0.07	0.08
	(Technology/Engineering) HS	0.09	0.10
	Average	0.10	0.08
ELA	10	0.21	0.20
Mathematics	10	0.08	0.12

The DETECT values indicate very weak to weak multidimensionality for all the tests for 2016–17. The ELA and the mathematics test forms tended to show slightly greater multidimensionality than did the science test forms. Also shown in Table 3-24 are the values reported in last year's dimensionality analyses. Last year's results are 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, high school ELA 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 high school mathematics test and the six science tests, none of them showed evidence of consistent separation of multiple-choice and constructed-response.

In summary, for the 2016–17 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 2017 MCAS Tests Parent/Guardian Reports* (Measured Progress Psychometrics and Research Department, *2016–2017 MCAS Equating Report*, unpublished manuscript).

Table 3-25 lists items that required intervention either during item calibration or as a result of the evaluations of the equating items. For each flagged item, the table shows the reason it was flagged (e.g., the *c*-parameter could not be estimated; the delta analysis indicated that the item'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 sections 3.6.2 and 3.6.3 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 3-25. 2017 Legacy MCAS: Items That Required Intervention During IRT Calibration and Equating

Content Area	Grade	ItemID	Reason	Action	Source
ELA	10	299076	*c*-parameter	set *c* = 0	Initial
	10	304017	delta analysis	removed from equating	
	10	304291	b/b analysis	removed from equating	
	10	304291	delta analysis	removed from equating	
	10	309519	*c*-parameter	set *c* = 0	Initial
	10	309519	*c*-parameter	set *c* = 0	Final
	10	314460	*c*-parameter	set *c* = 0	Initial
	10	314460	*c*-parameter	set *c* = 0	Final
	10	316607	*c*-parameter	set *c* = 0	Initial
	10	316607	*c*-parameter	set *c* = 0	Final
	10	316621	*c*-parameter	set *c* = 0	Initial
	10	316621	*c*-parameter	set *c* = 0	Final
Mathematics	10	312338	*c*-parameter	set *c* = 0	Initial
	10	314930	*c*-parameter	set *c* = 0	Initial
	10	314930	*c*-parameter	set *c* = 0	Final
	10	314972	*c*-parameter	set *c* = 0	Initial
	10	314972	*c*-parameter	set *c* = 0	Final
	10	315083	*c*-parameter	set *c* = 0	Initial
	10	315083	*c*-parameter	set *c* = 0	Final
Science	5	273732	*c*-parameter	set *c* = 0	Initial
	5	273732	*c*-parameter	set *c* = 0	Final
	5	281800	*c*-parameter	set *c* = 0	Initial
	5	281800	*c*-parameter	set *c* = 0	Final
	5	289163	*c*-parameter	set *c* = 0	Initial
	5	289163	*c*-parameter	set *c* = 0	Final
	5	289487	*c*-parameter	set *c* = 0	Initial
	5	289487	*c*-parameter	set *c* = 0	Final
	5	291143	*c*-parameter	set *c* = 0	Initial
	5	291143	*c*-parameter	set *c* = 0	Final
	5	299421	*c*-parameter	set *c* = 0	Final
	5	301136	delta analysis	removed from equating	
	5	309745	*c*-parameter	set *c* = 0	Initial
	5	309745	*c*-parameter	set *c* = 0	Final
	5	313146	*c*-parameter	set *c* = 0	Initial
	5	314833	*c*-parameter	set *c* = 0	Initial
	5	314833	*c*-parameter	set *c* = 0	Final
	8	291915	*c*-parameter	set *c* = 0	Initial
	8	291915	*c*-parameter	set *c* = 0	Final
Biology	10	222249	delta analysis	retained for equating	
	10	299780	delta analysis	retained for equating	
	10	305791	*c*-parameter	set *c* = 0	Initial
	10	313376	*c*-parameter	set *c* = 0	Initial
	10	314832	*c*-parameter	set *c* = 0	Initial
Introductory Physics	10	299362	delta analysis	retained for equating	
	10	311044	*c*-parameter	set *c* = 0	Initial
	10	313696	delta 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 2017 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,
a represents item discrimination,
b represents item difficulty,
d represents threshold, and
D is a normalizing constant equal to 1.701.

After computing k ICTCs in the GRM, $k+1$ item category characteristic curves (ICCCs), 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 I give the IRT item parameters and associated standard errors of all operational scoring items on the 2017 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 J contains graphs of the TCCs and TIFs, which are defined below.

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 3-25 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 3-26. The number of cycles required fell within acceptable ranges (less than 150) for all tests.

Table 3-26. 2017 Legacy MCAS: Number of Newton Cycles Required for Convergence

Content Area	Grade	Cycles	
		Initial	*Equating*
ELA	10	57	13
Mathematics	10	36	24
Science	5	30	85
	8	34	78
Biology	HS	40	1
Chemistry	HS	28	1
Introductory Physics	HS	29	1
Technology/Engineering	HS	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 2017 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.

The groups of students who take equating items on the MCAS tests are never strictly equivalent to the groups who took the tests in the reference years. IRT is particularly useful for equating scenarios that involve nonequivalent groups (Allen & Yen, 1979). Equating for the MCAS uses the anchor test–nonequivalent groups design described by Petersen, Kolen, and Hoover (1989). In this equating design, no assumption is made about the equivalence of the examinee groups taking different test forms (i.e., naturally occurring groups are assumed). Comparability is instead evaluated by using a set of anchor items (also called equating items), assuming they perform in the same way in both groups and can, thus, accurately measure the differences in the two groups.

Item parameter estimates for 2017 were placed on the 2016 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 2016 and 2017 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 3-25 at the beginning of this section. Each of these items was scrutinized, and a decision was made whether to include each item as an equating item or to discard it.

Appendix K 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 K 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 2017 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 2017. 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 high school ELA and mathematics tests as well as science grade 5 and 8 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 2007, establishing the theta cuts used for reporting each year. These theta cuts are presented in Table 3-27. 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 3-27. 2017 Legacy MCAS: Cut Scores on the Theta Metric and Reporting Scale by Content Area and Grade

Content Area	Grade	Theta			Scale Score				
		Cut 1	*Cut 2*	*Cut 3*	*Min*	*Cut 1*	*Cut 2*	*Cut 3*	*Max*
ELA	10*	-2.752	-1.495	0.153	200	220	240	260	280
Mathematics	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-scale score transformation constants.

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

3.6.5 Reported Scale 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 3-27), a 2-point difference between scale 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 scale scores.

By providing information that is more specific about the position of a student's results, scale scores supplement achievement level scores. Students' raw scores (i.e., total number of points) on the 2017 MCAS tests were translated to scale 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 2017 MCAS tests can be expressed in raw or scale scores.

It is important to note that converting from raw scores to scale scores does not change students' achievement level classifications. Given the relative simplicity of raw scores, it is fair to question why scale scores for the MCAS are reported instead of raw scores. The answer is that scale 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 scale scores of 240. It is this uniformity across scale scores that facilitates the understanding of student performance. The psychometric advantage of scale scores over raw scores comes from their being linear transformations of θ. Since the θ scale is used for equating, scale scores are comparable from one year to the next. Raw scores are not.

The scale 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 scale score metric. Students' ability estimates are based on their raw scores and are found by mapping through the TCC. Scale scores are calculated using the linear equation

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

where
m is the slope and
b is the intercept.

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

Appendix N contains raw-score-to-scale-score look-up tables. The tables show the scale score equivalent of each raw score for this year and last year. Appendix O contains scale score distribution graphs for each grade and content area. These distributions were calculated using the sparse data matrix files that were used in the IRT calibrations.

Table 3-28. 2017 Legacy MCAS: Scale Score Slopes and Intercepts by Content Area and Grade

Content Area	Grade	Cut Score Index	Theta Cut	Scale Score	Slope	Intercept
ELA	10	1	-4.000	200	0.862	200.000
		2	-3.000	218	6.694	238.424
		3	-2.752	220	15.910	263.786
		4	-1.495	240	12.135	258.143
		5	0.153	260	7.024	258.925
Mathematics	10	1	-4.000	200	1.055	200.000
		2	-3.000	210	6.767	230.523
		3	-1.555	220	25.740	260.025
		4	-0.778	240	25.412	259.771
		5	0.009	260	6.686	259.939
Science	5	1	-4.000	200	0.785	200.000
		2	-3.000	209	5.607	226.336
		3	-1.130	220	16.393	238.524
		4	0.090	240	20.000	238.200
		5	1.090	260	10.471	248.586
	8	1	-4.000	200	0.772	200.000
		2	-3.000	208	4.472	222.236
		3	-0.500	220	19.230	229.615
		4	0.540	240	14.925	231.940
		5	1.880	260	17.857	226.428
Biology	HS	1	-4.000	200	0.935	200.000
		2	-3.000	210	4.713	224.534
		3	-0.962	220	24.009	243.097
		4	-0.129	240	17.064	242.201
		5	1.043	260	10.219	249.340
Chemistry	HS	1	-4.000	200	0.744	200.000
		2	-3.000	206	4.638	220.621
		3	-0.134	220	35.778	224.794
		4	0.425	240	27.586	228.275
		5	1.150	260	10.810	247.567
Introductory Physics	HS	1	-4.000	200	0.925	200.000
		2	-3.000	210	4.024	222.873
		3	-0.714	220	24.330	237.372
		4	0.108	240	19.512	237.892
		5	1.133	260	10.712	247.862
Technology/Engineering	HS	1	-4.000	200	0.823	200.000
		2	-3.000	200	7.365	222.695
		3	-0.366	220	35.273	232.910
		4	0.201	240	18.198	236.342
		5	1.300	260	11.764	244.705

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 2017 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 3-29 presents descriptive statistics, Cronbach's α coefficient, and raw score SEMs for each content area and grade. (Statistics are based on common items only.) The reliability estimates range from 0.88 to 0.92, which generally are in acceptable ranges, and are consistent with results obtained for previous administrations of the tests.

Table 3-29. 2017 Legacy 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	10	69,365	72	53.48	10.67	0.90	3.32
Mathematics	10	69,429	60	39.71	12.56	0.92	3.54
Science	5	69,125	54	35.21	9.53	0.88	3.29
	8	69,971	54	33.64	10.28	0.90	3.33
Biology	9–12	52,728	60	40.14	11.30	0.91	3.35
Chemistry	9–12	705	60	38.40	12.00	0.91	3.54
Introductory Physics	9–12	14,126	60	38.68	12.21	0.92	3.46
Technology/ Engineering	9–12	2,519	60	35.90	10.83	0.90	3.49

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 2017 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 0.85 to 0.93 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 0.46 to 0.89, with a median of 0.68 and a standard deviation of 0.09. 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 2017 MCAS tests because it is easily adaptable to all types of testing formats, including mixed formats.

The DAC estimates reported in Tables 3-30 and 3-31 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 2017 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 (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 2017 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 3-30. 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 (0.74–0.86), consistency (0.64–0.80), and kappa (0.50–0.65) 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 0.75 for *Needs Improvement* for grade 10 ELA. This figure indicates that among the students whose true scores placed them in this classification, 75% would be expected to be in this classification when categorized according to their observed scores. Similarly, a consistency value of 0.62 indicates that 62% 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 3-31 provides accuracy and consistency estimates for the 2017 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 0.89–0.96 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%, respectively. These results indicate

that nearly all students were correctly classified with respect to being above or below the *Needs Improvement/Proficient* cutpoints.

Table 3-30. 2017 Legacy 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	10	0.86 (0.80)	0.65	0.76 (0.53)	0.75 (0.62)	0.84 (0.79)	0.89 (0.84)
Mathematics	10	0.82 (0.76)	0.61	0.83 (0.71)	0.68 (0.56)	0.70 (0.60)	0.92 (0.89)
Science	5	0.74 (0.64)	0.50	0.84 (0.75)	0.76 (0.67)	0.69 (0.60)	0.72 (0.55)
	8	0.81 (0.73)	0.60	0.87 (0.80)	0.77 (0.69)	0.81 (0.75)	0.54 (0.26)
Biology	9–12	0.82 (0.74)	0.62	0.82 (0.70)	0.76 (0.67)	0.79 (0.72)	0.87 (0.81)
Chemistry	9–12	0.77 (0.68)	0.57	0.85 (0.76)	0.64 (0.53)	0.72 (0.63)	0.87 (0.80)
Introductory Physics	9–12	0.82 (0.74)	0.63	0.83 (0.70)	0.75 (0.65)	0.80 (0.74)	0.88 (0.82)
Technology/Engineering	9–12	0.80 (0.71)	0.58	0.84 (0.74)	0.73 (0.64)	0.82 (0.77)	0.81 (0.65)

* Failing on all high school tests

Table 3-31. 2017 Legacy MCAS: Summary of Decision Accuracy (and Consistency) Results by Content Area and Grade—Conditional on Cutpoint

Content Area	Grade	Warning* / Needs Improvement			Needs Improvement / Proficient			Proficient / Advanced		
		Accuracy (consistency)	*False*		*Accuracy (consistency)*	*False*		*Accuracy (consistency)*	*False*	
			Positive	*Negative*		*Positive*	*Negative*		*Positive*	*Negative*
ELA	10	0.99 (0.99)	0.00	0.00	0.96 (0.95)	0.01	0.02	0.90 (0.86)	0.05	0.05
Mathematics	10	0.97 (0.95)	0.01	0.02	0.94 (0.91)	0.03	0.03	0.92 (0.88)	0.04	0.04
STE	5	0.94 (0.91)	0.03	0.04	0.89 (0.85)	0.05	0.05	0.91 (0.88)	0.06	0.03
	8	0.93 (0.90)	0.03	0.04	0.90 (0.87)	0.05	0.04	0.97 (0.96)	0.02	0.00
Biology	9–12	0.98 (0.97)	0.01	0.02	0.94 (0.91)	0.03	0.04	0.90 (0.87)	0.05	0.05
Chemistry	9–12	0.94 (0.92)	0.02	0.03	0.91 (0.88)	0.04	0.05	0.91 (0.88)	0.05	0.04
Introductory Physics	9–12	0.97 (0.95)	0.01	0.02	0.93 (0.90)	0.03	0.04	0.92 (0.89)	0.04	0.03
Technology/Engineering	9–12	0.94 (0.91)	0.02	0.04	0.90 (0.86)	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 3-30 and 3-31 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: *Failing*, *Needs Improvement*, *Proficient*, and *Advanced*. Students receive a separate achievement level classification in each content area. Reports are generated at the student level. *Parent/Guardian Reports* and student results labels are printed and mailed to districts for distribution to schools. The details of the reports are presented in the sections that follow. See Appendix Q for a sample *Parent/Guardian Report*.

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") report 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. 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. Please note, for the 2016–2017 academic year, only high school students that took the existing MCAS tests received a *Parent/Guardian Report* in the legacy MCAS report format. Students in grades 3–8 participated in the new MCAS ELA and mathematics tests, as well as the

existing science test for students in grades 5 and 8, and received a newly designed *Parent/Guardian Report*.

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 the Commissioner's letter to parents/guardians, general information about the test, and website information for parent/guardian resources. The inside portion contains the achievement level, scale score, and standard error of the scale score for each content area tested. If the student does not receive a scale score, the reason is displayed under the heading "Achievement Level." The student's historical scale scores are reported where appropriate and available. An achievement level summary of school, district, and state results for each content area is reported. The student's growth percentiles in ELA and mathematics are reported if sufficient data exist to calculate growth percentiles. The median growth percentiles for the school and district are also reported, 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.

A note is printed on the report, in the area where the scale 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 scale 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 dataset, 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
- multi-school district
- private school
- special school (e.g., a charter school)
- small school that does not have enough students to report aggregations
- school with excluded (not tested) students

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

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.[2] 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: http://www.doe.mass.edu/accountability/.

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: http://www.doe.mass.edu/sfss/presentations-pubs/. 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: http://www.doe.mass.edu/turnaround/howitworks/. Determining whether high school students have demonstrated the knowledge and skills required to

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

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.[3] 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.

2. 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).

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

4. 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 scale 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

[3] 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% 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.

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 mean percentage of possible points earned by students in the school, the district, and the state on MCAS items assessing particular standards/topics. 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. 2017 Legacy MCAS: School Results by Standards Report

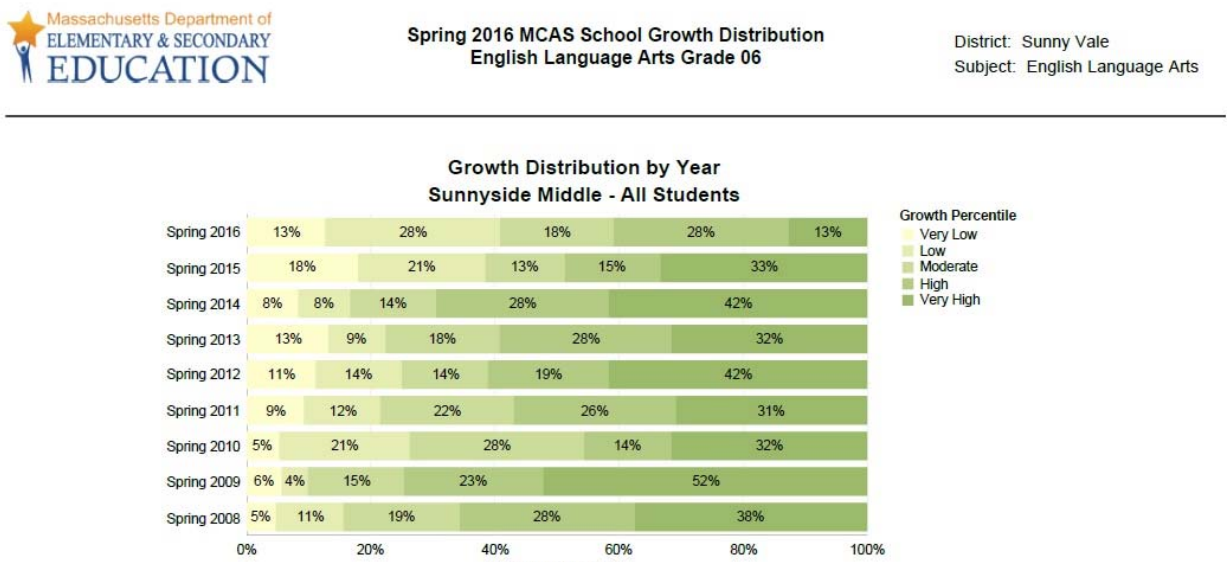
All Students Students (104)

Standards: MA 2011 Standards **Mode:** Online

	Possible Points	School % Possible Points	District % Possible Points	State % Possible Points	School/ State Diff
Mathematics					
All items	54	63%	47%	55%	8
Question Type					
Constructed Response	14	45%	29%	37%	8
Short Answer	8	58%	41%	48%	10
Selected Response	32	71%	57%	64%	8
Strand / Topic					
Expressions and Equations	**16**	**49%**	**36%**	**44%**	**6**
Analyze and solve linear equations and pairs of simultaneous linear equations.	4	36%	26%	32%	4
Understand the connections between proportional relationships, lines, and linear equations.	6	50%	39%	49%	2
Work with radicals and integer exponents.	6	57%	39%	47%	11
Functions	**14**	**64%**	**47%**	**53%**	**12**
Define, evaluate, and compare functions.	4	67%	43%	45%	22
Use functions to model relationships between quantities.	10	63%	48%	56%	7
Geometry	**16**	**67%**	**52%**	**59%**	**8**
Solve real-world and mathematical problems involving volume of cylinders, cones and spheres.	4	61%	46%	55%	6
Understand and apply the Pythagorean Theorem.	2	67%	52%	59%	8
Understand congruence and similarity using physical models, transparencies, or geometry software.	10	70%	54%	61%	8
Statistics and Probability	**5**	**84%**	**69%**	**75%**	**9**
Investigate patterns of association in bivariate data.	5	84%	69%	75%	9
The Number System	**3**	**63%**	**52%**	**59%**	**4**
Know that there are numbers that are not rational, and approximate them by rational numbers.	3	63%	52%	59%	4

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. 2017 Legacy MCAS: Growth Distribution Report



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.

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., & 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