MASSACHUSETTS DEPARTMENT OF
ELEMENTARY AND SECONDARY
EDUCATION

2017 Next-Generation MCAS and MCAS-Alt 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 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.

To fulfill the requirements of the Act, the MCAS is designed to

- measure student, school, and district performance in meeting the state's learning standards as detailed in the Massachusetts curriculum frameworks; and
- provide measures of student achievement that will lead to improvements in student outcomes.

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 report is to document the technical quality and characteristics of the 2017 next-generation MCAS ELA and mathematics tests in grades 3–8 and of the 2017 MCAS-Alt, in order to present evidence of the validity and reliability of test score interpretations and to describe modifications made to the program in 2017. A companion document, the *2017 Legacy MCAS Technical Report*, provides information regarding the legacy tests administered in 2017 (the MCAS high school ELA and mathematics tests, and Science and Technology/Engineering [STE] tests in grades 5, 8 and high school).

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 next-generation 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 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, purposes, uses, components, and validity of MCAS. The next two chapters cover the test design and development, test administration, scoring, and analysis and reporting of results for the standard MCAS assessments (Chapter 3) and the MCAS Alternate Assessment (Chapter 4). These two chapters include information about the characteristics of test items, how scores were calculated, the reliability of scores, how scores were reported, and the validity of results. Numerous appendices, which appear after Chapter 4, are referenced throughout the report.

1.4 Current Year Updates

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

Computer-based administrations were made available for the ELA and mathematics tests in grades 3–8 in 2017. Computer-based administration was mandatory at grades 4 and 8, and optional for grades 3, 5, 6, and 7. Paper-based tests (PBT) were available as a test accommodation at grades 4 and 8. Because of the transition from legacy MCAS to next-generation MCAS tests, the presentation of psychometric results for 2017 does not include any comparisons with previous years.

1.4.1 Next-Generation MCAS Assessments

On November 17, 2015, the Massachusetts Board of Elementary and Secondary Education (the Board) 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). Next-generation MCAS assessments include the following elements:

- high-quality test items aligned to the Massachusetts learning standards
- new item types that more deeply assess both skills and knowledge; for example:
  - *writing to text in ELA*
  - *solving complex problems in mathematics*
- achievement levels that send clear signals to students, parents, and educators about readiness for work at the next level
- a full range of student accessibility features and accommodations
- for the 2017 administration: both computer-based and paper test administrations, with a goal of phasing in computer-based testing as the primary testing method statewide in 2019

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

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

1.4.2 Background on the Transition to Next-Generation Assessments

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

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

Chapter 2 The State Assessment System: MCAS

2.1 Guiding Philosophy

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

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

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

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

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 the curriculum framework for its content area.

The 2017 next-generation MCAS tests were aligned to the 2011 Massachusetts curriculum frameworks. Tests given in 2018 and beyond will align to the revised curriculum framework standards adopted in March 2017.

All learning standards defined in the frameworks are addressed by and incorporated into 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

- providing information to support program evaluation at the school and district levels
- providing diagnostic information to help all students reach higher levels of performance

2.4 Validity of MCAS and MCAS-Alt

Validity information for the MCAS and MCAS-Alt assessments is provided throughout this technical report. Although validity is considered a unified construct, the various types of validity evidence contained in this report 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.

Tables 2-1 and 2-2 summarize validity information for MCAS and MCAS-Alt provided in specific sections of this report. Note that some of these sections will point the reader to additional validity evidence located in the appendices of the report.

Table 2-1. 2017 Next-Generation MCAS: Summary of Validity Evidence for the Next-Generation MCAS Tests

Type of Validity Evidence	*Section*	*Description of Information Provided*
Reliability and classical item analyses; scoring consistency and classification consistency by achievement level	3.4	Scoring consistency, interrater agreement, and scoring accuracy
	3.5	Classical item analyses
	3.7	Overall reliability and standard error of measurement by test; reliability by student subgroups
	3.7.5	Decision accuracy and consistency (DAC): estimates of accuracy for student classification by achievement level and for each achievement level cut score
Content-related validity evidence	3.2 and 3.9.1	Test blueprints; item alignment to test blueprints and standards
Construct-related and structural validity evidence	3.9.2	Response process validity evidence
	3.5 to 3.7	Item response theory modeling; dimensionality; scaling; linking online to paper results; differential item functioning
Consequential validity	3.8	MCAS Reporting
	3.9.5	Supporting the valid use of MCAS data

Because MCAS-Alt assessment results are both aggregated with and disaggregated from MCAS results, validity information provided for MCAS with respect to reliability and content-related validity also pertains, to some extent, to the MCAS-Alt (see sections 3.4, 3.5, 3.7, 3.7.4, and 3.2 as noted in Table 2-1 above). In addition, MCAS-Alt, which is a portfolio-based assessment, also

includes reliability and dimensionality characteristics specific to the portfolio assessment, as described below in Table 2-2.

Table 2-2. 2017 MCAS-Alt: Summary of Validity Evidence for MCAS-Alt

Type of Validity Evidence	*Section*	*Description of Information Provided*
Content-related validity evidence	4.2.1	Test blueprints are aligned to MCAS blueprints with modifications made for the range and complexity of standards. Includes primary evidence and supporting documentation sufficient for three dimensions of scoring for the evidence submitted.
Reliability and subgroup statistics and scoring consistency	4.4, 4.7.3, and 4.8	Procedures to ensure consistent scoring; interrater scoring statistics
	4.5	Classical item statistics
	4.7.1 and 4.7.2	Overall and subgroup reliability statistics
Construct-related and structural validity evidence	4.5.3	Interrelations among scoring dimensions
	4.6	Item bias review and procedures

2.5 Next-Generation MCAS Achievement-Level Descriptors

The achievement-level descriptors (ALDs) used to define expectations on the next-generation MCAS assessments were established to identify students who are academically prepared for academic work at the next grade level. In so defining the ALDs, Massachusetts's "Meeting Expectations" level is also aligned to the level of academic work a student must perform to eventually be prepared for college-level work upon completion of high school. The general ALDs for the next-generation MCAS tests at grades 3–8 are as follows:

Exceeding Expectations
A student who performed at this level exceeded grade-level expectations by demonstrating mastery of the subject matter.

Meeting Expectations
A student who performed at this level met grade-level expectations and is academically on track to succeed in the current grade in this subject.

Partially Meeting Expectations
A student who performed at this level partially met grade-level expectations in this subject. The school, in consultation with the student's parent/guardian, should consider whether the student needs additional academic assistance to succeed in this subject.

Not Meeting Expectations
A student who performed at this level did not meet grade-level expectations in this subject. The school, in consultation with the student's parent/guardian, should determine the coordinated academic assistance and/or additional instruction the student needs to succeed in this subject.

2.5.1 Grade-Specific Achievement-Level Descriptors

The grade-specific achievement level descriptors provided in Appendix A illustrate the knowledge and skills students at each grade are expected to demonstrate on MCAS at each achievement level. Knowledge and skills are cumulative at each level. No descriptors are provided for the *Not Meeting Expectations* achievement level because a student's work at this level, by definition, does not meet the criteria of the *Partially Meeting Expectations* level.

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 now assessed in both ELA and mathematics.

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

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

3.2 Test Design and Development

In 2017, the MCAS next-generation operational tests were administered for the first time at grades 3–8 in both ELA and mathematics. The assessments included newly developed next-generation MCAS items as well as PARCC items. Some students took these tests online, while others took the paper version.

3.2.1 Test Specifications

3.2.1.1 Criterion-Referenced Test

Items used on the next-generation MCAS are either developed specifically for Massachusetts or are PARCC items. Both sets of items are directly linked to Massachusetts curriculum framework content standards. These content standards are the basis for the reporting categories developed for each content area and are used to help guide the development of test items. The MCAS assesses only the content and processes described in the Massachusetts curriculum frameworks. Items on the 2017 next-generation MCAS tests were coded to the standards in the 2011 Massachusetts curriculum frameworks in ELA and mathematics.

3.2.1.2 Item Types

The types of items and their functions (by content area) are described below. For all items, blank responses, although considered wrong responses and assigned scores of zero, are disaggregated from incorrect responses in reports of student results.

ELA

- **Selected-response items (SR)** are worth one or two points**.**
  - **One-point selected-response items** (online/paper) make efficient use of limited testing time and allow for coverage of a wide range of knowledge and skills within a content area. Each one-point selected-response item requires students to select the single best answer from four response options. Each item is aligned to one primary standard. Items are machine-scored; correct responses are worth one score point, and incorrect and blank responses are assigned zero score points.
  - **Two-point selected-response items** (online/paper) have two parts. In the first part, students select the single best answer from four response options. In the second part, students select, from four response options, the evidence from the stimulus that supports the answer from the first part. The items are machine-scored; correct responses are worth two points, partially correct answers are worth one point, and incorrect and blank responses are assigned zero points. Students who answer the first part incorrectly receive a score of zero; students must answer the first part correctly in order to receive one or two points.
  - **Two-point technology-enhanced items** (online only) use computer-based interactions such as "drag-and-drop" and "hot spot." The items are machine-scored; correct responses are worth two points, partially correct answers are worth one point, and incorrect and blank responses are assigned zero points. In 2017, only one technology-enhanced item was used on the grade 4 test. In the future, more technology-enhanced items will be used on all tests.
- **Constructed-response (CR) items** (online/paper) are worth three points and are used only on the grades 3 and 4 tests. Students are expected to generate approximately one paragraph of text in response to a passage-driven question. Student responses are hand-scored, with a range of possible score points from zero to three. Students earn three points if their responses are completely correct and zero points if their responses are completely incorrect.
- **Essays (ES)** (online/paper) are administered to all students in grades 3–8, and include both narrative and text-based essays. Students are required to write an essay in response to text they have read. Each essay is hand-scored by scorers trained in the specific requirements of each question scored, with a range of possible score points from zero to eight, depending on the grade and type of essay.

Mathematics

- **Selected-response/multiple-select (SR) items** (online/paper) make efficient use of limited testing time and allow for coverage of a wide range of knowledge and skills within a content area. Each item is aligned to one primary standard and is worth one point. Selected-response items require students to select the single best answer from four response options. Multiple-select items require students to select one to three correct answers from a set of answer options. Selected-response and multiple-select items are machine-scored; correct responses are assigned one point for each correct selection, and incorrect and blank responses are assigned zero points.

- **Short-answer/fill-in-the-blank (SA) items** (online/paper) are worth one point each, and are used to assess students' skills and abilities to work with brief, well-structured problems that have one solution, 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. These items are machine-scored; correct responses are assigned one point, and incorrect and blank responses are assigned zero points.
- **Technology-enhanced (TE) items** (online only) use interactions such as "drag-and-drop" or "hot spot" that require the student to choose from a range of options presented.
  - **Two-point technology-enhanced items** include two question parts, which are machine-scored separately; the sum of the scores is the score earned on the item. Student responses are assigned one, two, or (for incorrect or blank responses) zero points.
  - **Four-point technology-enhanced items** (online only) are used only in grades 6 through 8. Students respond to questions using interactions such as "drag-and-drop," "hot spot," or choosing from a drop-down menu. Each part of these multi-part items is machine-scored separately and the sum of the scores is the score earned on the item. Each student response is assigned a score of one to three based on correct responses; incorrect and blank responses are assigned a score of zero.
- **Constructed-response (CR) items** (online/paper) require students to generate written responses to prompts. Student responses are hand-scored and the number of possible score points depends on item type.
  - **Two-** and **three-point constructed-response items** are used only on the grade 3 math test. Students are expected to generate one or two sentences of text in response to a word problem. Student responses are assigned score points ranging from zero to three, depending on item type.
    - Completely correct student responses are assigned a score of either two points for two-point items or three points for three-point items.
    - Partially correct student responses are assigned a score of either one point for two-point items, or one or two points for three-point items.
    - Completely incorrect or blank responses are assigned a score of zero.
  - **Three-** and **four-point constructed-response items**, administered in grades 4–8, typically require students to use higher-order thinking skills—such as evaluation, analysis, and summarization—to construct satisfactory responses. Student responses are assigned score points ranging from zero to four, depending on the item type.
    - Completely correct student responses are assigned a score of either three points for three-point items, or four points for four-point items.
    - Partially correct student responses are assigned a score of either one or two points for three-point items, or one, two, or three points for four-point items.
    - Completely incorrect or blank responses are assigned a score of zero.

3.2.1.3 Description of Test Design

The MCAS assessment instruments are structured using both common and matrix items. Identical common items are administered to all students in a given grade. Student scores are based on student performance on common items only. Matrix items are either new items included on the test for field-test purposes or equating items used to link one year's results to those of previous years. Equating

and field-test items are divided among the multiple forms of the test for each grade and content area. The number of test forms varies by grade and content area but ranges between 10 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) is obtained to produce reliable data that can be used to inform item selection for future tests.

In 2017, two common forms were developed for the grades 3–8 ELA and mathematics assessments: one form designated as the computer-based (CB) common form and one form designated as the paper-based (PB) common form. To create the PB common form, technology-based items that appeared on the CB form were replaced in the paper form by items that could be administered on paper. The replacement items were worth the same value and aligned to the same or a similar content standard as the online-only items.

3.2.2 ELA Test Specifications

3.2.2.1 Standards

The 2017 MCAS grades 3–8 ELA tests, including all field-test items, were aligned to, and measured the following learning standards from the *2011 Massachusetts Curriculum Framework for English Language Arts and Literacy*.

- **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)*

- **Anchor Standards for Writing**
  - *Text Types and Purposes (Standards 1–3[1])*
  - *Production and Distribution of Writing (Standard 4)*

For grade-level articulation of these standards, please refer to the *2011 Massachusetts Curriculum Framework for English Language Arts and Literacy*.

3.2.2.2 Item Types

The next-generation grades 3–8 ELA tests used a mix of selected-response, multiple-select, technology-enhanced, and essay items. The grades 3 and 4 tests also included short-response items.

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.

[1] Standard 1 was not assessed on the 2017 tests, but is an assessable standard.

Table 3-1. 2017 Next-Generation MCAS: ELA Item Types and Score Points

Item Type	Possible Raw Score Points	Grade Levels
Selected-response (SR)	0, 1, or 2	3–8
Evidence-based multiple-select (SR)	0, 1, or 2	3–8
Technology-enhanced (SR)	0, 1, or 2	3–8
Short-response (CR)	0, 1, 2, or 3	3–4
Essay – narrative (ES)	0 to 6 0 to 7	3–5 6–8
Essay – text-based (ES)	0 to 7 0 to 8	3–5 6–8

3.2.2.3 Passage Types

Passages range in length from approximately 600 to 2500 words per passage set. Word counts are slightly reduced at lower grades. Most passage sets consist of either a single passage or paired passages. Passages were selected from published works; no passages were specifically written for the MCAS 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. Some informational passages are more narrative or essayistic in nature, and yet provide sufficient information to be assessed via the informational ELA standards. Informational passages are drawn from a variety of sources, including magazines, newspapers, and books.

In grades 3–8, the common form of the 2017 next-generation ELA test included three passage sets, with some forms containing two literary passage sets and one informational passage set, and other forms containing one literary passage set and two informational passage sets.

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. Another important consideration is that passages be of interest to the age group being tested.

The main difference among the passages used for grades 3–8 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 at each grade level. 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 relevant prior knowledge must be connected to the text 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 ELA framework language standards use the passage as a stimulus. There are no stand-alone items on the next-generation MCAS ELA assessments; all vocabulary, grammar, and mechanics questions on the MCAS ELA tests are associated with a passage.

3.2.2.4 Test Design

In 2017, the next-generation ELA tests at grades 3–8 were comprised of items embedded within passage sets. Most passage sets consisted of a single passage or paired passages, followed by selected-response items, technology-enhanced items (online only), short-response items (grades 3 and 4 only), and essay items.

Only common items are used to determine student scores. The matrix slots in each test form are used to field-test items or to equate the current year's test to that of previous years by using previously administered items. In 2017, since it was the first operational year for the next-generation tests, only field-test items were included in the matrix slots.

Test Design by Grade

Grades 3–4

The common portion of each test at grades 3 and 4 included three passage sets. The first passage set typically included four 2-point selected-response items, one 6-point narrative essay item, and, in the case of grade 4, a technology item on the online test. The other two passage sets each included eight 1- or 2-point selected-response items and either a 7-point text-based essay item or one or two 3-point short-response items. Each test contained a total of 42 common points distributed across three testing sessions.

Grade 5

The common portion of each test included three passage sets. The first passage set included four 2-point selected-response items, and one 6-point narrative essay item. The other two passage sets each included eight 1- or 2-point selected-response items and a 7-point text-based essay item. The test contained a total of 46 common points distributed across three testing sessions.

Grades 6–8

The common portion of each test at grades 6, 7, and 8 included three passage sets. The first passage set included four 2-point selected-response items, and one 7-point narrative essay item. The other two passage sets each included eight 1- or 2-point selected-response items and an 8-point text-based essay item. The test contained a total of 49 common points distributed across two testing sessions.

Common and Matrix Item Distribution

Table 3-2 lists the distribution of common and matrix items in each 2017 next-generation ELA test, by grade.

Table 3-2. 2017 Next-Generation MCAS: Distribution of ELA Common and Matrix Items by Grade and Item Type – Online and Paper

Grade and Test			Items per Form							
			Common				Matrix			
Grade	*Test*	*# of Forms*	*SR (1 pt)*	*SR (2 pt)*	*CR*	*ES*	*SR (1 pt)*	*SR (2 pt)*	*CR*[1]	*ES*
3	ELA	10	18	4	1	2	6	2	2	1
4	ELA	10	18	4	1	2	6	2	2	1
5	ELA	10	18	4	0	3	6	2	0	1
6	ELA	10	18	4	0	3	6	2	0	1
7	ELA	10	18	4	0	3	6	2	0	1
8	ELA	10	18	4	0	3	6	2	0	1

[1] Each grade 3 and grade 4 matrix form contained either two constructed-response items or one essay item.

3.2.2.5 Blueprints

Table 3-3 shows the target and actual percentages of common item points by reporting category. Reporting categories are based on the Massachusetts curriculum framework strands.

Table 3-3. 2017 Next-Generation MCAS: Target (and Actual) Distribution of ELA Common Item Points by Reporting Category

Reporting Category	% of Points at Each Grade (+/-5%)					
	3	*4*	*5*	*6*	*7*	*8*
Language	25 (26)	25 (28)	25 (33)	25 (22)	25 (24)	25 (26)
Reading	55 (57)	55 (55)	45 (43)	45 (49)	45 (47)	45 (45)
Writing	20 (17)	20 (17)	30 (24)	30 (29)	30 (29)	30 (29)
Total	100	100	100	100	100	100

3.2.2.6 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 student recognize basic information presented in the text.
- **Level II (Infer/Analyze)** – Level II items require that the student understand a given text by making inferences and drawing conclusions related to the text.
- **Level III (Evaluate/Apply)** – Level III items require that the student 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 ELA test.

3.2.2.7 Recommended Testing Times

Table 3-4 shows the recommended testing times for the 2017 next-generation ELA tests at grades 3–8. MCAS tests are untimed; therefore, times shown in the table are approximate.

Table 3-4. 2017 Next-Generation MCAS: ELA Recommended Testing Times, Grades 3–8

Grade	*Session 1 recommended testing time (min)*	*Session 2 recommended testing time (min)*	*Session 3 recommended testing time (min)*	*Total recommended testing time (min)*
3	60	60	45	165
4	60	60	45	165
5	60	75	45	180
6	110	100	NA	210
7	110	100	NA	210
8	110	100	NA	210

3.2.2.8 Reference Materials

The use of bilingual word-to-word dictionaries was allowed during both ELA tests only for current and former English learner (EL) students. No other reference materials were allowed during the ELA tests.

3.2.3 Mathematics Test Specifications

3.2.3.1 Standards

All items on the 2017 next-generation MCAS mathematics assessments at grades 3–8 were aligned to the *2011 Massachusetts Curriculum Framework for Mathematics*.

The 2011 standards are grouped by domains at grades 3–8.

- **Domains for Grades 3–5**
  - *Operations and Algebraic Thinking*
  - *Number and Operations in Base Ten*
  - *Number and Operations—Fractions*
  - *Measurement and Data*
  - *Geometry*
- **Domains for Grades 6 and 7**
  - *Ratios and Proportional Relationships*
  - *The Number System*
  - *Expressions and Equations*
  - *Geometry*
  - *Statistics and Probability*

- **Domains for Grade 8**
  - *The Number System*
  - *Expressions and Equations*
  - *Functions*
  - *Geometry*
  - *Statistics and Probability*

3.2.3.2 Item Types

The 2017 next-generation MCAS mathematics tests in grades 3–8 included selected-response, multiple-select, short-answer, technology-enhanced, and constructed-response items, as well as PARCC items. Each type of item is worth a specific number of points in the student's total mathematics score, as shown in Table 3-5.

Table 3-5. 2017 Next-Generation MCAS: Mathematics Item Types and Score Points

Item Type	Possible Raw Score Points	Grade Levels
Selected-response/multiple-select	0 or 1	3–8
Short-answer/fill-in-the-blank	0 or 1	3–8
Technology-enhanced	0, 1, or 2	3–5
	0, 1, 2, 3, or 4	6–8
Constructed-response	0, 1, or 2	3
	0, 1, 2, or 3	4–5
	0, 1, 2, 3, or 4	4–8

3.2.3.3 Test Design

The 2017 next-generation MCAS mathematics tests in grades 3–8 were comprised of common and matrix items. The matrix slots in each test form are used to field-test potential items and to equate the current year's test to that of previous years by using previously administered items. In 2017, since it was the first operational year for the next-generation tests, only field-test items were included in the matrix slots.

Table 3-6 shows the distribution of common and matrix points on the 2017 next-generation MCAS mathematics tests, as well as recommended testing times, for grades 3–8. Since MCAS tests are untimed, the times shown are approximate.

Table 3-6. 2017 Next-Generation MCAS: Mathematics Test, Grades 3–8, Recommended Testing Times and Common/Matrix Points per Test

Grade	*# of Sessions*	*Session 1 Recommended Testing Time*	*Session 2 Recommended Testing Time*	*Total Recommended Testing Time*	*Common Points*	*Matrix Points*
3	2	65	65	130	48	5–9
4–5	2	70	70	140	54	6–10
6–8	2	70	70	140	54	12–24

The grades 3–8 next-generation mathematics tests were administered to some students in online forms and to other students in a paper form. Tables 3-7 (for the online form) and 3-8 (for the paper form) show the distribution of common and matrix item types on the 2017 next-generation MCAS mathematics tests.

Table 3-7. 2017 Next-Generation MCAS: Distribution of Mathematics Common and Matrix Items by Grade and Item Type – Online Form

Grade	# of Forms	Items per Form									
		Common								*Matrix*	
		MC/MS	*SA/FIB*	*TE*			*OR*			*MC/MS SA/FIB TE*	*OR*
		(1 pt)	*(1 pt)*	*(1 pt)*	*(2 pt)*	*(4 pt)*	*(2 pt)*	*(3 pt)*	*(4 pt)*	*(1, 2, or 4 pt)*	*(2, 3, or 4 pt)*
3	15	23	6	5	2	0	2	2	0	3	1
4	14	21	9	2	4	0	0	2	2	3	1
5	14	19	9	4	4	0	0	2	2	3	1
6	11	16	6	2	4	2	0	2	2	4	2
7	11	16	6	2	4	2	0	2	2	4	2
8	9	18	6	0	4	2	0	2	2	4	2

Table 3-8. 2017 Next-Generation MCAS: Distribution of Mathematics Common and Matrix Items by Grade and Item Type – Paper Form

Grade	# of Forms	Items per Form									
		Common								*Matrix*	
		MC/MS	*SA/FIB*	*Replacements for TE Items*			*OR*			*MC/MS SA/FIB*	*OR*
		(1 pt)	*(1 pt)*	*(1 pt)*	*(2 pt)*	*(4 pt)*	*(2 pt)*	*(3 pt)*	*(4 pt)*	*(1, 2, or 4 pt)*	*(2, 3, or 4 pt)*
3	1	27	7	0	2	0	2	2	0	3	1
4	1	24	8	0	4	0	0	2	2	3	1
5	1	24	8	0	4	0	0	2	2	3	1
6	1	18	6	0	4	2	0	2	2	4	2
7	1	18	6	0	4	2	0	2	2	4	2
8	1	18	6	0	4	2	0	2	2	4	2

3.2.3.4 Blueprints

Tables 3-9 through 3-11 show the target and actual percentages of common item points by reporting category. Reporting categories are based on the Massachusetts curriculum framework strands.

Table 3-9. 2017 Next-Generation MCAS: Target (and Actual) Distribution of Math Common Item Points by Reporting Category, Grades 3–5

Domain	% of Points at Each Grade (+/-5%)		
	3	*4*	*5*
Operations and Algebraic Thinking	33 (33)	25 (24)	20 (19)
Number and Operations in Base Ten	15 (15)	20 (20)	25 (24)
Number and Operations – Fractions	15 (17)	20 (23)	25 (26)
Geometry	12 (10)	15 (13)	10 (11)
Measurement and Data	25 (25)	20 (20)	20 (20)
Total	100	100	100

Table 3-10. 2017 Next-Generation MCAS: Target (and Actual) Distribution of Math Common Item Points by Reporting Category, Grades 6 and 7

Domain	% of Points at Each Grade (+/-5%)	
	6	*7*
Ratios and Proportional Relationships	19 (20)	20 (20)
The Number System	18 (18)	22 (23)
Expressions and Equations	30 (30)	20 (20)
Geometry	15 (15)	20 (20)
Statistics and Probability	18 (17)	18 (17)
Total	100	100

Table 3-11. 2017 Next-Generation MCAS: Target (and Actual) Distribution of Math Common Item Points by Reporting Category, Grade 8

Domain	% of Points at Each Grade (+/-5%)
The Number System	5 (5)
Expressions and Equations	30 (30)
Functions	25 (26)
Geometry	30 (30)
Statistics and Probability	10 (9)
Total	100

3.2.3.5 Cognitive Levels

Each item on the mathematics test is assigned a cognitive level according to the cognitive demand of the item. Cognitive levels are not synonymous with difficulty. The cognitive level 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 non-routine 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, non-routine, or less well-defined problems.

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

3.2.3.6 Reference Materials

Rulers are provided to students in grades 3–8. Paper rulers are provided to students taking the paper version of the mathematics test. Students taking the online mathematics test have access to two separate rulers: a centimeter ruler and a 1/8-inch ruler; students are not permitted to use hand-held rulers on the online test.

Reference sheets are provided to students at grades 5–8. These sheets contain information, such as formulas, that students may need to answer certain items. The reference sheets were updated for the 2017 MCAS administration, with the inclusion of the PARCC items.

The second session of the grades 7 and 8 mathematics tests is a calculator session. All items included in this session are either calculator-neutral (calculators are permitted but not required to answer the question) or calculator-active (students are expected to use a calculator to answer the question). Each grade 7 student had access to a five-function calculator during session 2 of the mathematics test. Each grade 8 student had access to a scientific calculator during session 2 of the mathematics test.

3.2.4 Item and Test Development Process

Table 3-12 provides a detailed view of the item and test development process, in chronological order.

Table 3-12. 2017 Next-Generation MCAS: Overview of Item and Test Development Process

Development Step	*Detail of the Process*
Select reading passages (for ELA only)	Contractor's content specialists find potential ELA 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 passages have been reviewed by an ADC and Bias. ADC and Bias make recommendations, and ESE makes the final determination of which passages will be used.
Develop items	Contractor's content specialists develop items in ELA, mathematics, aligned to specific Massachusetts standards.

Development Step	*Detail of the Process*
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 reviews items and makes recommendations. 5. ESE test developers edit and 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 constructed-response items and compositions	ESE and contractor content specialists meet to determine appropriate benchmark papers for training of scorers of field-tested constructed-response items and compositions. Scoring rubrics and notes are reviewed and edited during benchmarking meetings. During the scoring of field-tested items, the contractor contacts ESE content specialists with any unforeseen issues.
Item statistics meeting	ADCs review field-test statistics and recommend items for 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 items to become common-eligible or to be rejected.
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. Each 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	Approved common-eligible items become part of the common item set, and are used to determine individual student scores.
Released common items	Some common items in grades 3–8 are released to the public, and the remaining items return to the common-eligible pools to be used on future MCAS tests.

3.2.4.1 ELA Passage Selection

Passages used in the ELA tests are authentic published passages selected for the MCAS. Section 3.2.2.3 provides a detailed description of passage types and lengths. Test developers, including ESE content specialists, review numerous texts to find passages that possess the characteristics required for use in ELA tests. Passages must

- be of interest to and appropriate for students in the grade being addressed;
- have a clear beginning, middle, and end;
- contain appropriate content;
- support the development of unique, and a sufficient number of, assessment items; and
- be free of bias and sensitivity issues

Passages that are approved by the ESE are presented to the ADCs as well as the Bias and Sensitivity Review 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.

3.2.4.2 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 sections 3.2.2 and 3.2.3 for specific content standard alignment. Content not found in the curriculum frameworks is not subject to the statewide assessment.

Before items are field-tested, they go through several review steps:

- Initial ESE item review
- ADC review
- Bias review
- External content expert review
- Editing of recommended items

Initial ESE Item Review

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

- **Alignment**: Are the items aligned to the standards? Is there a better standard to which the item could be aligned?
- **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 selected-response items been chosen based on plausible content errors?
- **Mechanics**: How well are the items written? Do they follow the conventions of item writing?
- **Missed opportunities (for ELA only)**: Were there items that should have been written based on the passage, but were not?

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

Assessment Development Committee Review

ADCs are composed of 10 to 12 Massachusetts educators from across the state. Each ADC is facilitated by test development experts from Measured Progress and ESE. There is an ADC committee for each content area and grade (e.g., ELA grade 3).

ADC Passage Review (ELA Only)

ELA ADCs review passages before any corresponding items are written. Committee members consider all the elements listed on the previous page (i.e., grade-level and content appropriateness,

richness of content, etc.) as well as familiarity to students. If a passage is well known 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)
- 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.

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

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

All ADC committee recommendations remain with the item.

In the cycle of test development, ADCs first work to review new items and item passages for item accuracy, accessibility, and content alignment. After testing, they review item statistics to determine if students are responding to items as expected, and to identify items that are performing poorly or have potential bias issues.

Bias and Sensitivity Committee Review

After items have been developed and subsequently approved by the ADCs, they also undergo review by the Bias and Sensitivity Review Committee. (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 an issue(s) and suggests edits.)
- reject (The committee describes why the item should be rejected.)

All Bias and Sensitivity Review Committee comments are kept with the item.

After the ADC and Bias and Sensitivity reviews, ESE-approved items become “field-test eligible” and move to the next step in the development process.

External Content Expert 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. The External Content Experts choose one of the following recommendations regarding each item:

- accept
- reject (The expert describes the problem with the item and why rejecting the item is recommended.)

Expert reviewers' comments are included with the items.

Editing of Recommended Items

ESE content specialists review the recommendations of the ADC and Bias committees and expert reviewers, and determine whether 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.

3.2.4.3 Field-Testing of Items

Items that pass the reviews listed above are approved to be field-tested. Field-tested items appear in the matrix portions of the tests. Each matrix item is answered by a minimum of 1,500 students, resulting in enough responses to yield reliable performance data.

3.2.4.4 Scoring of Field-Tested Items

Each field-tested selected-response, multiple-select, short-answer/fill-in-the-blank, and technology-enhanced item is machine-scored.

Each field-tested constructed-response item and essay is hand-scored. To train scorers, the ESE works closely with the scoring staff to refine rubrics and scoring notes and to select benchmark papers that exemplify the score points and variations within each score point. Approximately 1,500 student responses are scored per field-tested constructed-response item/essay. As with the machine-scored items, 1,500 student responses are sufficient to provide reliable results. See section 3.4 for additional information on scorers and scoring.

3.2.4.5 Data Review of Field-Tested Items

Data Review by the ESE

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

Data Review by ADCs

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

- item difficulty (or mean score for polytomous items)
- item discrimination
- DIF
- distribution of scores across answer options and score points
- distribution of answer options and score points across quartiles
- distribution of unique student responses (for some items)

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

- accept
- edit and field-test again (This recommendation is made for mathematics items only; since ELA items are passage-based, individual items cannot be field-tested again. To address this matter in ELA, additional items are field-tested to ensure there are enough items to populate the operational test.)
- reject

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

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.

3.2.4.6 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 student answer another item.

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**. For paper-based tests, item placement is modified to ensure the best fit and arrangement of items on any given page.
- **Facing-page issues**. On paper-based tests, for selected-response items associated with a stimulus (ELA reading passages) and selected-response items with large graphics, consideration is given to whether those items need to begin on a left- or right-hand page, as well as to the nature and amount of material that needs to be placed on facing pages, in an effort to minimize the amount of page-flipping required of students.
- **Relationships among forms**. For paper-based tests, 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**. For paper-based tests, 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.

3.2.4.7 Operational Test Draft Review

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

3.2.4.8 Special Edition Test Forms

Students with Disabilities

MCAS is accessible to students with disabilities through the universal design of test items, provision of special edition test forms, and the availability of a range of accommodations and accessibility features 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 either in an individualized education plan (IEP) or in a 504 plan. All MCAS 2017 next-generation 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 and Department staff, and only when they do not provide clues or assistance to the student, or change what the item is measuring. When successful modification of an item or graphic is not possible, all or part of the item is omitted, and may be replaced with a similar item.
- **Screen reader** – This accommodation is available only for a student who is blind or has a visual disability. Students who use a screen reader also receive a separate hard-copy Braille edition test in order to provide the student with the appropriate Braille graphics. All answers are entered onscreen, either by the student using a Braille writing device, or by the test administrator.
- **Text-to-Speech** – This functionality was embedded in the grades 3–8 computer-based tests (CBT). 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.

Appendix B details other accommodations that do not require a special edition test form and also lists accessibility features that are available to all students, such as screen magnification and highlighting. Students who have an IEP or 504 plan are eligible to take the MCAS standard operational tests with accommodations. After testing is completed, the ESE receives a list that includes the number of students who participated in MCAS with each accommodation, based on information compiled in the Personal Needs Profile in PearsonAccess^{Next}.

3.3 Test Administration

3.3.1 Test Administration Schedule

The standard grades 3–8 next-generation MCAS tests were administered during two overlapping periods in spring 2017 as shown in Table 3-14 below:

Table 3-13. 2017 Next-Generation MCAS: Grades 3–8 ELA and Mathematics Test Administration Schedule

Content Area	Complete the Student Registration/ Personal Needs Profile (SR/PNP) Process	Receive Test Administration Materials	Test Administration Windows	Deadline to Complete the Principal's Certification of Proper Test Administration, Update Students' Accommodations, and Mark CBT Tests Complete	Deadline for Return of Materials to Contractor (for PBT Only)
ELA	January 23–February 10	March 20	April 3–May 3	May 4	May 5
Mathematics	January 23–February 10	March 20	April 4–May 26	May 30	May 31

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 violated the test security requirements were 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.

If test content is breached, quick identification and resolution of the breach are critical to the integrity of a testing program. In addition to reports of breaches in the field, the MCAS program used the services of Caveon Test Security, a nationally recognized test security organization, to perform web monitoring. Caveon Web Patrol leverages technology tools and human expertise to identify, prioritize, and monitor sites where sensitive test information may be disclosed. Caveon used the following strategies:

- systematically patrolled the Internet, websites, blogs, discussion forums, video archives, social media, document archives, brain dumps, auction sites, and media outlets
- identified and verified threats to MCAS test security and notified Pearson (who notified the Department and Measured Progress, as required)
- worked systematically through the steps necessary to have infringing content removed, if a threat was verified
- provided summary reporting that included overall and specific threat analysis

Full security requirements, including details about responsibilities of principals and test administrators, examples of testing irregularities, guidance for establishing and following a document tracking system, and lists of approved and unapproved resource materials, can be found in the *Spring 2017 Principal's Administration Manual, Grades 3–8* (PAM) and the 2017 *Test Administrator's Manuals* (TAMs). In spring 2017, there was one TAM for computer-based testing, and two TAMs for paper-based testing (one for grade 3, and one for grades 4–8).

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, including students with temporary disabilities such as a broken arm
- 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 C 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 are not required to participate in ELA testing
- students with significant disabilities who were unable to take the standard MCAS tests and instead participated in the MCAS-Alt (see Chapter 4 for more information)

- 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

More details about test administration policies and participation requirements for non-disabled students, for students with disabilities, for EL students, and for students educated in alternate settings can be found in the PAM.

3.3.4 Administration Procedures

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

- understanding and enforcing test security requirements and test administration protocols
- reviewing plans for maintaining test security with the superintendent
- ensuring that all enrolled students participated in testing at their grade level
- coordinating the school's test administration schedule and ensuring that tests were administered in the correct order and during the prescribed testing windows
- ensuring that test accommodations were properly provided and that transcriptions, if required for any accommodation, were done appropriately (Accommodation frequencies during 2017 testing can be found in Appendix D; for a list of test accommodations, see Appendix B. The overall number of accommodations has increased in the Next-Generation MCAS administration due to the inclusion of CBT-specific accommodations such as Text to Speech.)
- 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 PAM.

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

Scoring of the 2017 next-generation MCAS tests was conducted by both Measured Progress and Pearson. For paper-based test takers, Measured Progress scanned each MCAS student answer booklet. Images for field-test items were loaded into iScore, Measured Progress's secure scoring engine. Images for operational items were transferred via FTP site to Pearson for uploading into the ePEN scoring engine. Computer-based test takers had images of their answers uploaded into the

appropriate scoring engine so that all scoring was conducted in a similar manner, regardless of the method of test administration.

Student identification information, demographic information, school contact information, and student answers to selected-response 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.

A set of quality-control procedures were enacted for scanning paper test forms. These are provided in Appendix E and included

- checks of the answer booklet codes against the grade level, to ensure that the correct answer booklets were scanned in each batch;
- counting checks, to ensure that all booklets were accounted for; and
- spot checks, in which the scanned results were checked against randomly selected answer booklets to ensure the scanners were working as intended.

For computer-based test takers, the Department had previously reviewed all items in the online item bank (ABBI) and approved all selected-response answer keys during test construction. The item scoring specifications (in Question and Test Interoperability [QTI]) were configured using the test maps and keys provided for the tests. Once the scoring system was configured, a quality-assurance group verified that the selected responses entered by the student for an item as shown in the uploaded image corresponded to the response recorded in the database, for both the pre-score and the scored student data files.

Scoring for selected-response items was verified against the specific ESE requirements for the item; the requirement of the test map, which includes the QTI response; and the keys and validations made for an individual student's derived scores per level of the test. This process included review of all score-value-related fields—such as raw scores, object scores (part one and part two of multi-part items), strand scores, performance levels, pass/fail indicators, attempt rules, and scale scores—against the tables provided by Pearson psychometrics.

Scoring consistency across scoring departments on all item types was established by conducting the following activities:

- Measured Progress provided annotated training materials for all existing items to Pearson for review in advance of scoring. Content specialists at Pearson and Measured Progress spoke with each other to address any questions and ensure clarity of training materials.
- Measured Progress facilitated benchmarking meetings at its Dover, New Hampshire, offices. Pearson scoring staff were in attendance, either virtually or in person, to observe the meetings and to facilitate the eventual transition of items to operational status.
- For operational ELA items that needed to be benchmarked again due to modifications, content specialists at Measured Progress, Pearson, and ESE collaborated on the establishment of final scoring decisions.
- Weekly meetings between the scoring departments were held to address any issues and questions before and during scoring.

To ensure consistency in scoring constructed-response and essay item types, the Measured Progress scoring project manager traveled to Pearson scoring centers in Columbus, Ohio, and San Antonio,

Texas, to observe leadership training, scorer training, and operational scoring for both ELA and Mathematics. Measured Progress's assistant director of scoring content also visited a Pearson scoring center in Virginia Beach, Virginia. In addition to ensuring scoring consistency, these trips enabled all parties to monitor quality-control processes.

3.4.1 Benchmarking Meetings

Samples of student responses to field-test items, along with some operational ELA items that needed to be re-benchmarked due to modifications, 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. To help ensure consistency between field-test scoring and eventual operational scoring, content specialists from Pearson were also in attendance at benchmarking meetings, either in person or virtually. All decisions were recorded and considered final upon ESE signoff.

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

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

In addition to standard benchmarking meetings for field-test items, some unreleased operational ELA items were modified from their previous use to align with the new test design. These modifications changed the scoring of the items. A simplified benchmarking approach was used for these items, during which scoring staff created revised Anchor, Practice, and Qualification papers aligned to the new rubrics. Phone meetings were then conducted between Measured Progress, Pearson, and the ESE to review the final training materials before ESE signoff.

3.4.2 Machine-Scored Items

Student responses to selected-response, multiple-select, and technology-enhanced items were machine-scored by ePEN Scoring. On paper-based tests, student responses with multiple marks and blank responses were assigned zero points.

3.4.3 Hand-Scored Items

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

3.4.3.1 Scoring Location and Staff

Hand-scoring of MCAS item responses occurred in various locations, as summarized in Table 3-14.

Table 3-14. 2017 Next-Generation MCAS: Summary of Scoring Locations and Scoring Shifts

Pearson Scoring Sites; Content	Grade	Shift	Hours
Distributed Scoring			
PARCC ELA Items	3–8	2–8 hrs/day	7:00 a.m.–11:00 p.m.
PARCC Math Items	3–8	2–8 hrs/day	7:00 a.m.– 1:00 p.m.
Austin, TX			
PARCC ELA Call Center	6–8	Day Night	7:00 a.m.–3:00 p.m. 3:00 pm–11:00 p.m.
PARCC Math Call Center	3–5	Day Night	7:00 a.m.–3:00 p.m. 3:00 p.m.–11:00 p.m.
Charlotte, NC			
ELA	7–8	Day	8:00 a.m.–4:30 p.m.
Columbus, OH			
ELA	5–6	Day	8:00 a.m.–4:30 p.m.
Iowa City, IA			
PARCC ELA Call Center	8	Day	7:00 a.m.–3:00 p.m.
Mesa, AZ			
Math	3–4	Day	8:00 a.m.–4:30 p.m.
ELA	3	Day	8:00 a.m.–4:30 p.m.
Naperville, IL			
PARCC Math Call Center	6–7	Day Night	7:00 a.m.–3:00 p.m. 3:00 pm–11:00 p.m.
San Antonio, TX			
Math	5–8	Day	8:00 a.m.–4:30 p.m.
PARCC ELA Call Center	6–8	Day Night	7:00 a.m.–3:00 p.m. 3:00 p.m.–11:00 p.m.
PARCC Math Call Center	8	Day Night	8:00 a.m.–3:00 p.m. 3:00 p.m.–11:00 p.m.
Virginia Beach, VA			
ELA	4	Day	8:00 a.m.–4:30 p.m.

The following staff members were involved with scoring the 2016–17 MCAS responses:

- **Measured Progress Staff**
  - The *Scoring Project Manager* was located in Dover, New Hampshire, and oversaw communication and coordination of MCAS scoring between Measured Progress and Pearson.
  - A *Scoring Content Specialist* in mathematics and ELA ensured consistency of content area benchmarking and scoring across all grade levels at each scoring location. Scoring Content Specialists prepared all training material, handed off the relevant training materials to Pearson, and fielded any questions between Pearson and Measured Progress to ensure a consistent scoring approach among the scoring groups and across years.
- **Pearson Staff**
  - The *Scoring Portfolio Manager* was located in Iowa City, Iowa, and was responsible for the coordination, management, and oversight of MCAS scoring for Pearson.
  - The *Scoring Project Manager* was located in Iowa City, Iowa, and oversaw communication and coordination of MCAS scoring between Pearson and Measured Progress.
  - A *Scoring Content Specialist* in mathematics and ELA ensured consistency of content area scoring across all grade levels at each scoring location. Scoring Content Specialists monitored the quality of scoring and worked closely with a group of Scoring Directors to

ensure the accurate and timely completion of scoring. Scoring Content Specialists coordinated communication with their counterparts at Measured Progress regarding the training material.

- *Scoring Directors* were responsible for the training and qualification of scorers and scoring supervisors, and ensuring quality targets for their assigned items.
- *Scoring Supervisors* provided support and direction to scorers on quality, accuracy, and timely scoring completion.

3.4.3.2 Scorer Recruitment and Qualifications

MCAS scorers, a diverse group of individuals with a wide range of backgrounds, ages, and experiences, were recruited to meet contract requirements. These requirements included that all MCAS scorers had successfully completed at least two years of college, although hiring preference was given to individuals with a four-year college degree.

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 of qualifications, such as résumés and transcripts, which were carefully reviewed. Regardless of their qualifications, 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. A summary of scorers' backgrounds across the scoring sites and shifts are summarized in Table 3-15 below.

Table 3-15. 2017 Next-Generation MCAS: Summary of Scorers' Backgrounds across Scoring Shifts and Scoring Locations (Operational Scoring)

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	0	0.00	0	0.00
Bachelor's degree	860	54.53	78	63.93
Master's degree/doctorate	717	45.47	44	36.07
Teaching Experience				
No teaching certificate or experience	517	32.78	63	51.64
Teaching certificate or experience	998	63.29	52	42.62
College instructor	62	3.98	7	5.74
Scoring Experience				
No previous experience as scorer	632	40.08	35	28.69
1–3 years of experience	586	37.16	40	32.79
3+ years of experience	359	22.76	47	35.52

3.4.3.3 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.

Scoring 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
- recorded modules (used for individuals, small groups, or large groups)

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 represents a typical response, rather than an unusual or uncommon one; solid, rather than controversial, content; and a true score, meaning that this response has a precise score that will not be changed. Anchor sets are used to exemplify each score point.
- **Practice sets** include unusual, discussion-provoking responses, illustrating the range of responses encountered in operational scoring (including exceptionally creative approaches; extremely short or disorganized responses; responses that demonstrate attributes of both higher-score anchor papers and lower-score anchor papers; and responses that show traits of multiple score points). Practice sets are used to refine the scorers' understanding of how to apply the scoring rules across a wide range of responses.
- **Qualifying sets** consist of 10 responses that are clear, typical examples of each of the possible 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.

PARCC Item Scoring Training

Some items on the spring 2017 test were PARCC-developed items that were included on the MCAS test. Training materials for these items were provided by PARCC. Scorers who were trained and qualified to score these items from the PARCC assessment also scored the PARCC items embedded in the MCAS test in accordance with the same scoring parameters and scoring procedures as for the MCAS items.

3.4.3.4 Leadership Training

Scoring content specialists also had overall responsibility for ensuring that scoring leadership (scoring supervisors and scoring directors) 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 was required to meet or surpass a qualification standard of at least 80% exact and 90% exact-plus-adjacent scoring accuracy.

3.4.3.5 Methodology for Scoring Constructed-Response Item Responses and Essays

Score Options

The MCAS tests included constructed-response items requiring students to generate a brief response. Constructed-response items included short-answer items (mathematics only) with assigned scores of 0–1; short-response items (grades 3 and 4 ELA only) with assigned scores of 0–3; constructed-response items requiring longer or more complex responses with assigned scores of 0–4, 0–3, or 0–2 (mathematics only); and ELA essays with assigned scores of 0–8, depending on grade level and type of essay.

The sample 4-point mathematics constructed-response item scoring guide below (Table 3-16) illustrates the item-specific MCAS scoring guides used in 2017.

Table 3-16. 2017 Next-Generation MCAS: Four-Point Constructed-Response Item Scoring Guide – Grade 8 Mathematics

Score	Description
4	The student response demonstrates an exemplary understanding of the Statistics and Probability concepts involved in constructing and assessing a line of best fit informally. For scatter plots that suggest a linear association, the student informally fits a straight line and informally assesses the model fit by judging the closeness of the data points to the line.
3	The student response demonstrates a good understanding of the Statistics and Probability concepts involved in constructing and assessing a line of best fit informally. 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 constructing and assessing a line of best fit informally. 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 constructing and assessing a line of best fit informally.
0	The student response contains insufficient evidence of an understanding of the Statistics and Probability concepts involved in constructing and assessing a line of best fit informally 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.
- **Send for Review**

Scorers were instructed to mark any potential nonscorable essay as "Send for Review." At the review of a "Send for Review" response, scoring leadership would review the response and either assign a number score or apply one of the following condition codes:

- **Unreadable:** The response cannot be read because of poor penmanship, spelling cannot be deciphered, writing is too small, too faint to see, or only partially visible.
- **Non-English:** Response was written entirely in a language other than English, or without enough English or numbers to provide a score.
- **Off Topic:** Response does not address the topic or task for the item. The response is irrelevant to the item prompt, or response states that the student is refusing to participate in testing.
- **Direct Copy:** Direct copy of text from the passage or item prompt.

Scorers could also flag a response as a "Crisis" response, which resulted in sending the response to the scoring leadership for immediate attention.

A response could be 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).

Single-Scoring, Double-Blind Scoring, and Read-Behind Scoring

Student responses were either single-scored (response was scored only once by a single scorer) or double-blind scored (response was independently read and scored by two different scorers).

Double-Blind Scoring

In double-blind scoring, neither scorer knew whether the response had been scored before or were aware of any prior score it had been given. For a double-blind response with discrepant scores between the two scorers that were within one point of each other, the higher score was used. Any double-blind response with discrepant scores greater than one point (for items with three or more score points) was sent to the arbitration queue and read by a scoring supervisor or a scoring director.

Double-blind scoring was conducted on 10% of constructed-response item responses on both the ELA and mathematics tests at grades 3–8.

Read-Behind Scoring

In addition to the 10% double-blind scoring, scoring leadership, at random points throughout the scoring shift, engaged in read-behind (back-reading) scoring for each of the scorers at his or her table. This process involved scoring leadership viewing responses recently scored by a particular scorer and assigning his or her own score to that same response. Scoring leadership would then compare scores and advise or counsel the scorer as necessary.

Table 3-17 illustrates the rules for instances when two read-behinds or two double-blind scores were not identical (i.e., adjacent or discrepant).

Table 3-17. 2017 Next-Generation MCAS: Read-Behind and Double-Blind Resolution Charts

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

* If double-blind scores are adjacent, the higher score is used as the final score. If double-blind scores are neither identical nor adjacent, the resolution score is used as the final score.

3.4.3.6 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 indicators, 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:

- embedded validity responses
- read-behind scoring (RBs)/back-reading
- double-blind scoring (DBs)
- compilation reports (summary of scoring agreement statistics)

Embedded validity responses were used to monitor the scorer's accuracy of scoring. These responses were approved by the scoring content specialist or scoring director and distributed to scorers based on a percentage of their total number of responses scored. For the first two days, validity responses were routed to scorers to comprise 6% of their responses for ELA and 3% for mathematics. Starting with the third day of live scoring, these rates were reduced to 4% for ELA and 2% for mathematics. At the third-day rate, a full shift of scoring was expected to result in 6–19 validity responses per day in ELA and around 8 validity responses per day in mathematics, based on expected read rates.

Read-behinds involved responses that were first read and scored by a scorer, then read and scored by a member of scoring leadership. Scoring leadership would, at various points during the scoring shift, conduct a review of submitted scorer work. After the scorer scored the response, scoring leadership 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. Read-behinds 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 agreement standard were counseled, given extra monitoring assignments such as additional read-behinds, and allowed to resume scoring if they demonstrated the ability to meet the scoring standards after the intervention.

Double-blinds involved responses scored independently by two different scorers. Scorers knew in advance that some of the responses they scored were going to be scored by others, but they had no way of knowing what responses would be scored by another scorer, or whether they were the first, second, or only scorer. Responses given discrepant scores by two independent scorers were read and scored by scoring leadership. Scorers who fell below the 70% exact and 90% exact-plus-adjacent score agreement standard during the scoring shift were counseled, given extra monitoring assignments such as additional read-behinds, and were allowed to resume scoring if they demonstrated the ability to meet the scoring standards after the intervention.

Compilation reports combined all scorer statistics, including the percentage of exact, adjacent, and discrepant scores on the validity responses, and the scorer's percentage of exact, adjacent, and discrepant scores on the read-behinds. As scoring leadership conducted read-behinds, 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.

Warnings were issued to scorers that did not meet minimum validity metrics after a minimum of 10 validity responses. If after an additional five validity responses, the scorer had not improved, ePEN automatically locked out that scorer and a 10-response targeted calibration set was administered. The scorer was required to attain at least 70% exact agreement and 90% exact-plus-adjacent agreement on this calibration set to continue scoring the project. If the scorer passed the targeted calibration, ePEN was unlocked and the scorer regained admission to continue scoring. The scorer was required to continue maintaining scoring standards for validity, as validity statistics continued to be checked every 10 validity responses. If validity fell below scoring standards at any of these subsequent intervals, the scorer was released from the project and all scores assigned immediately reset.

3.4.3.7 Interrater Consistency

Interrater consistency statistics are the result of the processes implemented to ensure valid and reliable hand-scoring of items and, as such, provide evidence of scoring stability. 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. For student constructed-response questions in grades 3–8, 10% 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 scoring consistency on the MCAS tests.

A summary of the interrater consistency results is presented in Table 3-18. Results in the table for hand-scored items are organized by content area and grade. The table shows the number of score categories, the number of included scores, the percent exact agreement, the 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 F. The percentage of third reads is affected by new scoring procedures and by the scoring resolution process instituted, and is higher this year than in prior years due to additional oversight applied in the first year of the next-generation MCAS scoring with the two scoring vendors.

Table 3-18. 2017 Next-Generation MCAS: Summary of Interrater Consistency Statistics Organized across Items by Content Area and Grade

Content Area	Grade	Number of		Percent*		Correlation	Percent of Third Scores
		Score Categories	*Included Scores*	*Exact*	*Adjacent*		
ELA	3	4	27,706	70.02	28.69	0.71	23.83
		5	6,868	67.17	31.16	0.71	15.39
	4	4	27,826	68.58	30.47	0.73	40.10
		5	6,941	68.91	30.23	0.76	29.32
	5	4	27,353	68.46	30.41	0.76	30.63
		5	13,659	70.85	28.46	0.73	35.81
	6	4	20,016	69.14	29.72	0.78	24.41
		5	6,740	65.21	30.61	0.82	10.28
		6	13,276	66.45	32.73	0.78	31.58
	7	4	20,475	67.03	30.89	0.76	26.61
		5	6,847	65.15	32.61	0.83	28.09
		6	13,628	65.03	32.12	0.79	25.87
	8	4	20,563	72.70	26.15	0.80	30.51
		5	6,890	65.14	31.23	0.82	29.59
		6	13,673	69.06	30.82	0.86	30.97
Mathematics	3	3	14,019	93.28	6.68	0.94	26.18
		4	6,975	88.03	11.37	0.93	39.61
	4	4	14,142	87.43	12.20	0.93	19.78
		5	14,152	85.37	12.82	0.95	27.91
	5	4	13,794	85.08	14.21	0.94	18.87
		5	13,897	86.72	12.01	0.96	24.03
	6	4	13,481	89.10	9.75	0.93	13.12
		5	13,774	90.12	8.70	0.98	19.44
	7	4	13,491	84.92	13.96	0.91	69.63
		5	13,934	91.42	7.61	0.97	32.84
	8	4	6,876	81.17	17.84	0.91	41.48
		5	13,904	88.33	10.82	0.96	21.32

*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 predominantly assess the knowledge and 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 next-generation MCAS items meet these standards. Qualitative analyses such as those conducted by the ADC committees 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 next-generation MCAS assessments in spring 2017. Note that the information presented in this section is based only on the operational items, since those are the items on which student scores are calculated. (Item analyses, not included in this report, are also performed for field-test items, and the statistics are then used during the item review process and during form assembly for future administrations.)

As there were two test administration modes—online and paper—in spring 2017, there was a concern that the testing mode might introduce a construct-irrelevant variance to test scores. A mode comparability study was conducted on tests for the online-optional grades (grades 3, 5, 6, and 7), which suggested there is a small but significant mode effect for some grades and subjects (see Appendix J: IRT & Mode Linking Report).

One set of psychometric procedures were adjusted to address these mode differences: dimensionality analysis was conducted separately for online and paper test forms, since dimensionality analysis is based on between-item covariance after conditioning on total test scores, and will be affected if the testing mode introduces an irrelevant variance to the total test scores.

Other psychometric procedures were not affected. Evaluations of the difficulty indices, item-test correlations, and DIF statistics are less affected by testing mode, so these evaluations were conducted based on the whole population (analyses were conducted on the entire population that took each item; in instances where the item was unique to one testing mode, the statistics presented here are for the sample of the population taking that unique item).

It should be noted that 2017 is the first administration of the next-generation MCAS assessment, so no comparisons to previous years' results are provided.

3.5.1 Classical Difficulty and Discrimination Indices

All selected-response and constructed-response items are evaluated in terms of item difficulty according to standard classical test theory practices. Difficulty is defined as the average proportion of points achieved on an item and is measured by obtaining the average score on an item and dividing it by the maximum possible score for the item. Selected-response items are scored dichotomously (correct vs. incorrect), so, for these items, the difficulty index is simply the

proportion of students who correctly answered the item. Constructed-response items and essay 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 constructed-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 earned 0% of the item points, and an index of 1.0 indicates that all students received all of the item points or 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 selected-response items or essentially zero for constructed-response items) to 0.90, with the majority of items generally falling between 0.4 and 0.7. However, on a standards-referenced assessment such as the MCAS, it may be appropriate to include some items with very low or very high item difficulty values to ensure sufficient content coverage.

A desirable characteristic of an item is for higher-ability students to perform better on the item than lower-ability students. 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 2017 next-generation MCAS constructed-response items, the item discrimination index used was the Pearson product-moment correlation; for selected-response 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 for selected-response items 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 students with lower total scores. Given this definition, an item can discriminate between low-performing examinees and high-performing examinees. Discrimination indices were very useful to consider when selecting items for the new next-generation MCAS tests and were provided to the ADC committees along with other item-level statistics, such as difficulty. Very low or negative point-biserial coefficients on field-tested new items can indicate that the items are flawed and should not be considered for the operational tests.

A summary of the item difficulty and item discrimination statistics for each grade and content area combination is presented in Table 3-19. Note that the statistics are presented for all items as well as separately by item type: selected response (SR), constructed response (CR), and essay (ES). 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 3-19. 2017 Next-Generation MCAS: Summary of Item Difficulty and Discrimination Statistics by Content Area and Grade

Content Area	Grade	Item Type	Number of Items	*Difficulty*		Discrimination	
				Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
ELA	3	ALL	27	0.63	0.19	0.39	0.12
		SR	18	0.70	0.17	0.35	0.09
		CR	5	0.53	0.14	0.41	0.06
		ES	4	0.42	0.06	0.59	0.03
	4	ALL	28	0.68	0.14	0.45	0.11
		SR	18	0.75	0.10	0.42	0.07
		CR	6	0.60	0.13	0.38	0.13
		ES	4	0.52	0.07	0.64	0.02
	5	ALL	28	0.71	0.14	0.43	0.14
		SR	18	0.78	0.09	0.38	0.08
		CR	4	0.68	0.13	0.34	0.11
		ES	6	0.52	0.09	0.65	0.04
	6	ALL	28	0.63	0.15	0.48	0.13
		SR	18	0.69	0.13	0.40	0.07
		CR	4	0.60	0.10	0.52	0.04
		ES	6	0.47	0.10	0.69	0.02
	7	ALL	28	0.66	0.12	0.44	0.15
		SR	18	0.69	0.11	0.36	0.07
		CR	4	0.67	0.13	0.42	0.10
		ES	6	0.56	0.12	0.70	0.02
	8	ALL	28	0.69	0.16	0.46	0.17
		SR	18	0.78	0.10	0.40	0.07
		CR	4	0.51	0.09	0.32	0.16
		ES	6	0.54	0.12	0.73	0.02
Mathematics	3	ALL	45	0.63	0.16	0.47	0.10
		SR	24	0.69	0.14	0.43	0.08
		CR	21	0.57	0.17	0.51	0.09
	4	ALL	43	0.62	0.19	0.45	0.12
		SR	21	0.68	0.15	0.41	0.09
		CR	22	0.56	0.20	0.50	0.12
	5	ALL	45	0.59	0.16	0.45	0.10
		SR	23	0.63	0.16	0.40	0.08
		CR	22	0.55	0.15	0.50	0.09
	6	ALL	39	0.55	0.20	0.49	0.14
		SR	17	0.62	0.20	0.40	0.12
		CR	22	0.49	0.17	0.56	0.12
	7	ALL	38	0.54	0.19	0.50	0.12
		SR	18	0.64	0.14	0.44	0.09
		CR	20	0.45	0.18	0.56	0.12
	8	ALL	36	0.56	0.18	0.49	0.12
		SR	18	0.65	0.16	0.40	0.07
		CR	18	0.47	0.16	0.58	0.10

Caution should be exercised when comparing indices across grade levels. Differences may be due not only to differences in the item statistics on the test, but may also be affected by differences in student abilities and/or differences in the standards and/or curricula taught in each grade.

Difficulty indices for selected-response items tend to be higher (indicating that students performed better on these items) than the difficulty indices for constructed-response items because selected-response items can be answered correctly by simply identifying rather than providing the correct answer, and also by guessing. Similarly, discrimination indices for those constructed-response items with more than two points tend to be larger than those for dichotomous items because of the greater variability of the former (i.e., the partial credit these items allow). The restriction of range (i.e., only two score categories) in dichotomous items tends to make the discrimination indices lower. Note that these patterns are more consistent within item type, so when interpreting classical item statistics, comparisons should be emphasized among items of the same type.

In addition to the item difficulty and discrimination summaries presented above, item-level statistics along with item-level score point distributions, based on the combined sample of online and paper tests, are provided in Appendix G. On next-generation MCAS items, the item difficulty and discrimination indices are within generally acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that students who performed well on individual items tended to perform well overall. There are a small number of items with discrimination indices below 0.20, but none was negative. While it is acceptable to include items with low discrimination values or with very high or very low item difficulty values when their content is needed to ensure that the content specifications are appropriately covered, there were very few such cases on the next-generation MCAS. Item-level score point distributions are provided for constructed-response items in Appendix H; for each item, the percentage of students (online and paper) 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 be examined when sample sizes permit and that actions 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 next-generation MCAS items in terms of DIF statistics. One application of the DIF statistics is to use them to evaluate item quality in the ADC and bias committee item review process.

For the next-generation 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. The minimum group N to calculate DIF is 75.

DIF for items is evaluated initially at the time of field-testing. 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 actual item bias. Consequently, all items with DIF are examined by content experts and educators to try to identify the cause. 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. If content experts do not identify a source of bias on the item, the item may be eligible for operational form construction.

Computed DIF indices have a theoretical range from -1.0 to 1.0 for selected-response items, and an adjusted index with the same scale (-1.0 to 1.0) for constructed-response items. Dorans and Holland (1993) suggest that index values between -0.05 and 0.05 denote either a negligible amount of DIF or the absence of DIF. The majority of next-generation MCAS items fell within this range. Dorans and Holland further state 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 operationally.

For the 2017 next-generation MCAS administration, DIF analyses were conducted for all subgroups (as defined in the No Child Left Behind Act) for which the sample size was adequate. Six subgroup comparisons were evaluated for DIF:

- male/female
- not LEP-FLEP/LEP-FLEP
- not economically disadvantaged/economically disadvantaged
- white/African American
- white/Hispanic
- special education/no special education

The tables in Appendix I present the number of items classified as either "low" or "high" DIF, in total and by group favored. The moderate number of items that exhibited low DIF and several that exhibited high DIF were reviewed by content and educational experts to rule out a source of bias prior to being included on the operational 2017 next-generation MCAS tests.

3.5.3 Dimensionality Analysis

Because tests are constructed with multiple content area subcategories and their associated knowledge and skills, the potential exists for the invocation of multiple dimensions beyond the common primary dimension. Generally, the subcategories are highly correlated with each other; therefore, a primary dimension typically explains the majority of variance in test scores. The presence of one dominant primary dimension is the primary psychometric assumption to support the use of the unidimensional item response theory (IRT) models that are used for calibrating and scaling the next-generation MCAS assessments.

The purpose of dimensionality analysis is to investigate whether violation of the assumption of test unidimensionality is statistically detectable and, if so, (a) the degree to which unidimensionality is violated and (b) the nature of the multidimensionality. Dimensionality analyses were performed on common items for all next-generation MCAS test forms administered during the spring 2017 administrations. Test forms in the two administration modes were analyzed separately. A total of 24 test forms were analyzed; the results for these analyses are reported below.

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 across all possible conditioning scores. When a test is strictly unidimensional, all conditional covariances are expected to take on values within random noise of zero, indicating statistically independent item responses for examinees with equal expected scores. Nonzero conditional covariances are essentially violations of the principle of local independence, and such local dependence implies multidimensionality. Thus, nonrandom patterns of positive and negative conditional covariances are indicative of multidimensionality.

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

DETECT is an effect-size measure of multidimensionality. As with DIMTEST, the data are first randomly divided into a training sample and a cross-validation sample (these samples are drawn independently of those used with DIMTEST). The training sample is used to find a set of mutually exclusive and collectively exhaustive clusters of items that best fit a systematic pattern of positive conditional covariances for pairs of items from the same cluster and negative conditional covariances for pairs comprised 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 operational items of the next-generation MCAS tests administered during spring 2017. For all grades except grades 4 and 8, there were over 26,000 student examinees per test form in both mathematics and ELA. For grades 4 and 8, a majority of the student population took the online test form, resulting in over 60,000 students per online form, and less than 6,000 students per paper form. The data for each grade were split into a training sample and a cross-validation sample. Because DIMTEST had an upper limit of 24,000 students, the training and cross-validation samples for the tests that had over 24,000 students were limited to 12,000 each, randomly sampled from the total sample. DETECT, on the other hand, had an upper limit of 500,000 students, so every training sample and cross-validation sample used all the available data. After randomly splitting the data into training and cross-validation samples, DIMTEST was applied to each data set to see if the null hypothesis of unidimensionality would be rejected. DETECT was then applied to each data set for which the DIMTEST null hypothesis was rejected in order to estimate the effect size of the multidimensionality.

3.5.3.1 DIMTEST Analyses

The results of the DIMTEST analyses indicated that the null hypothesis was rejected at a significance level of 0.01 for every data set. Because strict unidimensionality is an idealization that almost never holds exactly for a given data set, the statistical rejections in the DIMTEST results were not surprising. Indeed, because of the very large sample sizes involved in most of the data sets (over 26,000 in 20 of the 24 test forms), 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 3-20 below displays the multidimensionality effect-size estimates from DETECT.

Table 3-20. 2017 Next-Generation MCAS: Multidimensionality Effect Sizes by Grade, Content Area, and Test Mode

Content Area	Grade	Multidimensionality Effect Size	
		Online	*Paper*
ELA	3	0.25	0.25
	4	0.30	0.36
	5	0.35	0.33
	6	0.38	0.42
	7	0.34	0.36
	8	0.38	0.48
	Average	0.33	0.37
Mathematics	3	0.20	0.21
	4	0.19	0.22
	5	0.19	0.18
	6	0.21	0.22
	7	0.13	0.13
	8	0.11	0.09
	Average	0.17	0.17

The DETECT values indicate very weak to weak multidimensionality for all the 2017 next-generation mathematics test forms. The 2017 next-generation ELA test forms in both modes show moderate multidimensionality.

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 selected-response and constructed-response item types. Inspection of the DETECT clusters indicated that selected-response/ constructed-response separation generally occurred much more strongly with ELA than with mathematics, a pattern that has been consistent across all previous years of dimensionality analyses for the MCAS legacy tests. Specifically, for the next-generation ELA test forms, every grade had one set of clusters dominated by selected-response items and another set of clusters dominated by constructed-response items. On the next-generation mathematics test forms, there was less clear evidence of consistent separation of selected-response and constructed-response items.

In summary, for the 2017 dimensionality analyses, the violations of local independence, as evidenced by the DETECT effect sizes, were either weak or very weak in mathematics test forms, and were weak to moderate in ELA test forms. The patterns with respect to the selected-response

and constructed-response items were consistent with those in the legacy MCAS tests, with ELA tending to display more separation than mathematics.

3.6 MCAS IRT Linking and Scaling

This section describes the procedures used to calibrate, link, and scale the next-generation MCAS tests. As this is the first administration of the next-generation MCAS tests, there is no equating between this year and last year. However, given the existence of two testing modes (online and paper), extra effort was taken to ensure that scores from the two modes were placed on the same scale. As the first step, linking was conducted between the two modes in an attempt to place them on the same scale. Then a mode comparability study was conducted to evaluate whether scale differences still remained after linking. Section 3.6.3 describes the linking study, and section 3.6.4 describes the mode comparability study.

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 linking items between testing modes (e.g., delta analyses, a-a plots, b-b plots);
- evaluation of the scaling results (e.g., parallel processing by the Psychometrics and Research Department and the Data and Reporting Services [DRS] Department); and
- replication and confirmation of calibration, linking, and scaling results from Pearson.

The testing vendor, Measured Progress, developed procedures to (1) evaluate the impact of mode in the online-optional grades (i.e., grades 3, 5, 6, and 7) and (2) link the paper forms to the online forms using items that performed similarly across modes in all grades. Methods for this analysis were evaluated by the MCAS Technical Advisory Committee and by the Massachusetts ESE.

A mode linking report, which provides 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. This report, *2016–2017 MCAS IRT and Mode Linking Report*, prepared by Measured Progress's Psychometrics and Research Department, is provided in Appendix J. Methodologies for conducting the mode analysis and linking study are provided in chapter 3 of the *Linking Report*. A key feature of the study, linking the paper scales to the online scales, is described in section 3.6.3 below.

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 next-generation MCAS grades 3–8 mathematics and ELA tests, the three-parameter logistic (3PL) model was used for traditional four-option selected-response items, and the two-parameter logistic (2PL) model was used for binary-scored selected-response and technology-enhanced items (Hambleton & van der Linden, 1997; Hambleton, Swaminathan, & Rogers, 1991). The graded-response model (GRM) was used for polytomous items (Nering & Ostini, 2010), including polytomously scored multi-part items, constructed-response items, and essays.

The 3PL model for selected-response 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,
a 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 the 2PL model, this reduces to the following:

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

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 2-PL form:

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

where
U indexes the scored response on an item,
i indexes the items,
j indexes students,
k indexes threshold,
θ is the student ability,

α represents item discrimination,
b represents item difficulty,
d represents threshold, and
D is a normalizing constant equal to 1.701.

After computing k ICTCs in the GRM, $k+1$ item category characteristic curves (ICCCs), which indicate the probability of responding to a particular category given θ, are derived by subtracting adjacent ICTCs:

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

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

The GRM is also commonly expressed as:

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

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

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

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

Note that for a dichotomously scored item, $E(U_i|\theta_j) = P_i(\theta_j)$. For more information about item calibration and determination, see Lord and Novick (1968), Hambleton and Swaminathan (1985), or Baker and Kim (2004).

3.6.2 IRT Results

The tables in Appendix J give the IRT item parameters and associated standard errors of all operational scoring items on the 2017 next-generation MCAS grades 3–8 ELA and mathematics tests. 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 3-21 lists items that required intervention during item calibration. For each flagged item, the table shows the reason it was flagged (e.g., the *c*-parameter could not be estimated or poor model fit) and what action was taken. In most cases, items flagged during this step were identified because of the guessing parameter (*c*-parameter) being poorly estimated. Difficulty in estimating the *c*-parameter is not at all unusual and is well documented in the psychometric literature (see, 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. In our analyses, the *c*-parameters for items with convergence problems were fixed to either 0 or 0.05. These choices were made for better comparisons with *c*-parameter estimates from IRTPRO (Cai, Thissen, & du Toit, 2011), which is the software used by Pearson. For one item (IA00349D) on the grade 6 ELA paper-based test, the *a*-parameter was fixed to the initial value supplied by PARSCALE,[2] as we found this intervention helped improve the model fit to a large extent. Furthermore, for one item (IA00731) on the grade 4 mathematics online test, the calibration was changed from 2PLM to 3PLM. The 2PLM was chosen initially because this item was developed as a technology-enhanced item that involved minimal guessing. However, by taking a closer look at the item content shown in Figure 3-1 below, it was determined that it was essentially a selected-response item with three options, and one option could be easily ruled out. We found changing to the 3PLM greatly improved the item fit for this item.

[2] The initial value for the *a*-parameter was set to be $\frac{\rho'}{\sqrt{1-\rho'^2}}$, where ρ' is the polyserial correlation of an item. The logic for using this initial value can be found in Lord (1980, pp. 31–33).

Figure 3-1. 2017 Next-Generation MCAS: Item IA00731

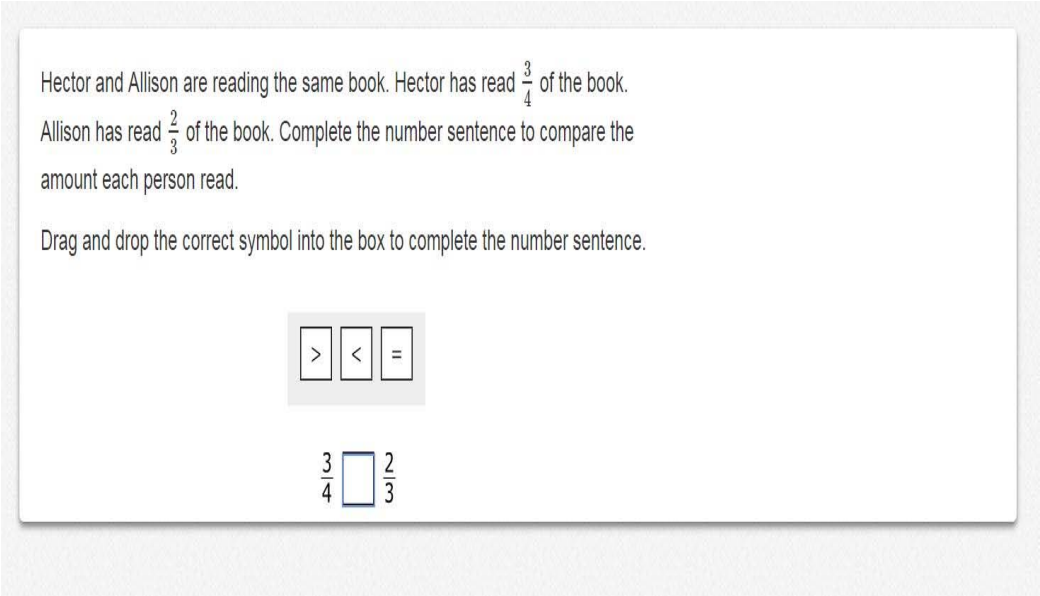


Table 3-21. 2017 Next-Generation MCAS: Items That Required Intervention During IRT Calibration

Content Area	Grade	Mode	Item ID	Reason	Action
ELA	3	Online	IA00268	*c*-parameter	set c = 0.05
		Online	IA00275	*c*-parameter	set c = 0.05
		Paper	IA00275	*c*-parameter	set c = 0.00
		Paper	IA00268	*c*-parameter	set c = 0.00
	5	Online	IA00146	*c*-parameter	set c = 0.00
		Paper	IA00146	*c*-parameter	set c = 0.00
	6	Online	IA00347	*c*-parameter	set c = 0.05
		Paper	IA00347	*c*-parameter	set c = 0.05
		Paper	IA00349D	*a*-parameter	a set to initial
	7	Online	IA00072	*c*-parameter	set c = 0.05
		Online	IA00074	*c*-parameter	set c = 0.05
		Online	IA00077	*c*-parameter	set c = 0.05
		Paper	IA00072	*c*-parameter	set c = 0.00
		Paper	IA00077	*c*-parameter	set c = 0.00
	8	Online	IA00207	*c*-parameter	set c = 0.00
		Online	IA00200	*c*-parameter	set c = 0.00
Mathematics	4	Online	IA00731	Model fit	2PL => 3PL
	5	Online	IA00809	*c*-parameter	set c = 0.00
		Online	IA00802	*c*-parameter	set c = 0.05
		Paper	IA00809	*c*-parameter	set c = 0.00
		Paper	IA00802	*c*-parameter	set c = 0.05
	6	Online	IA01191	*c*-parameter	set c = 0.00
		Paper	IA01210	*c*-parameter	set c = 0.00
		Paper	IA01191	*c*-parameter	set c = 0.00
	7	Paper	IA00883	*c*-parameter	set c = 0.05
		Paper	IA00768	*c*-parameter	set c = 0.05
	8	Online	IA00877	*c*-parameter	set c = 0.00
		Online	IA00760	*c*-parameter	set c = 0.05

The number of Newton cycles required for convergence for each grade and content area during the IRT analysis can be found in Table 3-22. The number of cycles required fell within acceptable ranges (less than 150) for all tests.

Table 3-22. 2017 Next-Generation MCAS: Number of Newton Cycles Required for Convergence

Content Area	Grade	Online Initial Cycles	Paper Initial Cycles
ELA	Grade 3	35	33
	Grade 4	31	59
	Grade 5	45	79
	Grade 6	38	30
	Grade 7	31	56
	Grade 8	40	32
Mathematics	Grade 3	30	27
	Grade 4	45	45
	Grade 5	26	60
	Grade 6	41	42
	Grade 7	68	56
	Grade 8	39	58

3.6.3 Linking

The purpose of linking is to "put scores from two or more tests on the same scale" (Kolen & Brennan, 2010, p. 423). After successful linking, scores from different tests are comparable to one another, and thus 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. Linking may be used if multiple test forms are administered in the same year, or one year's forms may be linked to those used in the previous year. Since 2017 was the first administration of the next-generation MCAS assessment, there is no year-to-year linking. However, given the existence of two testing modes (online vs. paper), linking was conducted to put scores from test forms in the two modes on the same scale.

A common concern is that different testing modes might introduce a construct-irrelevant variance to students' scores (i.e., mode effect). However, if there is no mode effect, scores from the two testing modes can be treated as equivalent after linking. Otherwise, the linking relationship established through the anchor items may not be accurate; as a result, the difficulty difference between test forms in different modes is not accurately adjusted by linking.

Additionally, the groups of students who take the online forms are not equivalent to the groups who take the paper form. IRT is particularly useful for linking that involves nonequivalent groups (Allen & Yen, 1979). Mode linking used the anchor test–nonequivalent groups design described by Petersen, Kolen, and Hoover (1989). In this linking design, the examinee groups taking different test forms do not need to be equivalent (i.e., naturally occurring groups are assumed), and group difference will be adjusted through linking, as long as the difference is not too large. Comparability is instead evaluated by using a set of anchor items and assuming they perform in the same way on both forms; they can, thus, accurately measure the differences in the two groups.

Specifically, the online form in each test was used as the reference form, given that all tests will gradually shift to online administration in the future. Scores from the paper form were linked to scores from the online form for each ELA and mathematics test in grades 3–8. Note that for ELA nonaccommodated tests, all items used for scoring were common between the two testing modes; for mathematics nonaccommodated tests, a majority of items used for scoring were common between the two testing modes, but there were two to five unique items for each mode per test.

Item parameter estimates of anchor items from the paper-based administration were placed on the online scale by using the Stocking-Lord method (SL; Stocking & Lord, 1983), which is based on the IRT principle of item parameter invariance. According to this principle, the anchor items from both modes should have the same item parameters. Thus, prior to implementing this method, two evaluations were conducted to check whether the item parameter invariance holds between the testing modes. In other words, anchor items between the two testing modes were evaluated for DIF between modes. These evaluations included delta analysis and IRT-based analysis. In delta analysis, delta values of anchor items from one mode were plotted against those from the other mode. In IRT-based analysis, the item parameters for each test were first freely estimated using PARSCALE (Muraki & Bock, 2003). The resulting estimated *b*-parameters of anchor items in one mode were plotted against those in the other mode. In both delta and *b*-*b* plots, any items that appeared as outliers were flagged.

For most of the tests, the delta and *b*-*b* analyses did not detect mode DIF in any items. In only four tests (ELA grades 3 and 8, mathematics grades 5 and 6) mode DIF items were identified, and for only one or two items per test. Appendix J presents results from the delta analysis and the *b*-*b* analysis. The discard status presented in the appendix indicates whether the item was flagged as potentially inappropriate for use in linking.

Note that for ELA grade 8, the one item that was identified by delta analysis was retained as a linking item, because the flagging index barely crossed the flagging criterion, and that item was not identified by *b*-*b* analysis. There was also a strong motivation for using the online item parameters for paper-based items for ELA grade 8, because the paper and online forms shared exactly the same set of items in ELA grade 8, and more than 95% of students took the online administration. Therefore, the online item parameters were directly used for items on the paper forms to build the look-up table for paper-based tests for ELA grade 8. Although a majority of students took online test forms in ELA at grade 4 and in mathematics at grades 4 and 8, the items in these grades and content areas were not entirely the same between online and paper test forms. Therefore, online parameters were not directly used for paper-based items for these three tests.

For the other tests, the anchor items that successfully survived these evaluation procedures were then employed in the Stocking-Lord (SL) method, and the linking relationship obtained from the SL method was used to transform the item parameters for all items in the paper-based administration onto the online scale. The transformed item parameters were then used to build the raw score to theta look-up tables for the paper-based tests.

3.6.4 Mode Comparability and Adjustment

As mentioned earlier, there is a common concern of construct-irrelevant variance due to different testing modes. If the delta and *b*-*b* methods effectively detect all the mode-DIF items, and if linking is conducted with non-DIF items only, the mode effect can be minimized. However, there is no guarantee that the delta and *b*-*b* methods have perfect power to detect all mode-DIF items, especially if a mode effect exists in all the anchor items. Therefore, a mode comparability study was conducted after linking to evaluate whether equivalent groups have the same performance on the two testing modes. As a majority of students took online test forms in grades 4 and 8 for both ELA and mathematics, the mode comparability study and adjustment were not conducted for those tests.

The rationale behind mode comparability evaluation is to compare two equivalent groups' performance on the two testing modes. Given the preexisting ability difference between the online- and paper-tested student groups, the propensity score matching technique (Rosenbaum & Rubin,

1983; Stuart, 2010) was used to adjust group ability difference and create matched groups between modes. Specifically, the propensity scores were estimated by fitting a logistic regression to the 2016–17 testing mode (online vs. paper) on a number of covariates, including prior year score and demographic variables (race, gender, LEP status, economically disadvantaged status, special education status, years in Massachusetts). Note that in 2015–16, a subpopulation took the MCAS legacy test; another subpopulation took the PARCC assessment, which was also administered in two modes (online and paper). So for the 2017 next-generation MCAS tests, there were three types of applicable prior-year scores: MCAS, PARCC online, and PARCC paper. The propensity score estimation and matching were conducted for each type of prior-year score, respectively. In other words, the total population was divided into three subpopulations based on students' prior-year scores, and the propensity score estimation and matching were conducted for each subpopulation, respectively.

After propensity scores were estimated, nearest neighbor matching with a caliper size of 0.02 was conducted on propensity scores. A bi-directional matching approach was further implemented to make the resulting matched sample representative of the state population. The bi-directional matching was conducted in three steps:

(1) For each student in the online group, a student from the paper group with the closest propensity score was selected. The resulting matched sample was denoted as the online-equivalent group.
(2) For each student in the paper group, a student from the online group with the closest propensity score was selected. The resulting matched sample was denoted as the paper-equivalent group.
(3) The original paper group was combined with the online-equivalent group to form the matched paper group; similarly, the original online group was combined with the paper-equivalent group to form the matched online group.

Because the matched sample is a combination of the original group in one mode and an equivalent group in the other mode, the combination results in a population-representative sample. This process is equivalent to adjusting the weight of the observations in the original sample in one mode so as to make it similar to the population.

As a prior score was not available for 2016–17 grade 3 students, a pseudo prior-score approach was used to create "prior" scores for those students. An implicit assumption in the pseudo prior-score approach is that grade 3 students across two years in the same school can be considered as equivalent groups in terms of their ability. To implement the pseudo prior-score approach, the following three steps were conducted for each school:

(1) Find the grade 3 students' score for that school in year 1.
(2) Find the grade 3 students' score at that school in year 2; if there were fewer than 10 students in the school in either year, the school was deleted from the analysis.
(3) Conduct equipercentile linking between scores in steps 1 and 2 to find the year 1 equivalent score (i.e., pseudo prior-score) for each year 2 student.

After matched samples were generated, two analyses were conducted to evaluate the matching effectiveness. The first analysis was to calculate the balance, which was the standardized difference between two matched groups on each matching variable. The standardized difference before matching was also calculated as a reference for comparison. The second analysis was to calculate the standardized difference between each matched group and the total population on each matching

variable. This analysis was to evaluate the population representativeness of the matched sample. The results are presented in Appendix J. The results for the next-generation 2017 tests suggested that both the balance and the population representativeness index were smaller than 0.1 in their absolute values after matching, thus suggesting matching was effective (Austin & Mamdani, 2006).

After matched samples were created, mode effect was calculated as the standardized difference between matched samples on students' scale scores. A nonparametric permutation test was further conducted to evaluate the statistical significance of the differences. Results are presented in Appendix J. The results suggested there was a small but significant mode effect for each test, and ELA tended to have a larger effect size than mathematics.

With the presence of a significant mode effect, scores on the paper forms were adjusted to minimize the mode effect. Specifically, equipercentile linking was conducted between θ estimates (after linking) from the two matched groups. The paper group was treated as the reference group, so that the online equivalent score was calculated for each θ on the paper scale. Online equivalent scores were used as the adjusted paper scores. The adjusted look-up tables are presented in Appendix J, and plots showing the adjustment results are presented in Appendix J. Mode effect analysis was conducted again—this time between the online scores and the adjusted paper scores. Results are presented in Appendix J. The nonsignificant mode difference in each test suggested the mode adjustment was effective.

3.6.5 Achievement Standards

Cutpoints for next-generation MCAS grades 3–8 ELA and mathematics tests were set via standard setting in 2017, establishing the theta cuts used for reporting each year. These theta cuts are presented in Table 3-23. 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. The 2017 *Standard Setting Report* in Appendix L provides a full description of how these cutpoints were established.

Table 3-23. 2017 Next-Generation MCAS: Cut Scores on the Theta Metric and Reporting Scale by Content Area and Grade

Content Area	Grade	Theta			Scaled Score				
		Cut 1	*Cut 2*	*Cut 3*	*Min*	*Cut 1*	*Cut 2*	*Cut 3*	*Max*
ELA	3	-1.581	0.011	1.604	440	470	500	530	560
	4	-1.561	0.031	1.623	440	470	500	530	560
	5	-1.659	0.038	1.734	440	470	500	530	560
	6	-1.591	-0.011	1.570	440	470	500	530	560
	7	-1.560	0.011	1.582	440	470	500	530	560
	8	-1.456	0.051	1.559	440	470	500	530	560
Mathematics	3	-1.377	0.027	1.432	440	470	500	530	560
	4	-1.379	0.054	1.487	440	470	500	530	560
	5	-1.551	0.025	1.601	440	470	500	530	560
	6	-1.518	-0.008	1.502	440	470	500	530	560
	7	-1.414	0.031	1.476	440	470	500	530	560
	8	-1.496	-0.008	1.479	440	470	500	530	560

3.6.6 Reported Scaled Scores

Because the θ scale used in IRT calibrations is not understood by most stakeholders, reporting scales were developed for the next-generation MCAS ELA and mathematics tests in grades 3–8. The reporting scales are linear transformations of the underlying θ scale. As the three θ cutpoints from the standard setting have equal intervals, one single linear transformation was sufficient to transform the θ scale from each performance level category on one reporting scale. Student scores on the next-generation MCAS tests are reported in integer values from 440 to 560. Because the same transformation is applied to all achievement-level categories and the reported scale scores preserve the interval scale properties (except for the truncated scale scores at the lower and upper end of the score scale), it is appropriate to calculate means and standard deviations with scaled scores.

By providing information that is more specific about the position of a student's results, scaled scores supplement achievement-level scores. Students' raw scores (i.e., total number of points) on the 2017 next-generation MCAS tests were translated to scaled scores using a data analysis process called *scaling*, which simply converts from one scale to another. In the same way that a given temperature can be expressed on either the Fahrenheit or the Celsius scale, or the same distance can be expressed in either miles or kilometers, student scores on the 2017 next-generation MCAS tests can be expressed in raw or scaled scores.

It is important to note that converting from raw scores to scaled scores does not change students' achievement-level classifications. Given the relative simplicity of raw scores, it is fair to question why scaled scores for the MCAS are reported instead of raw scores. The answer is that scaled scores make the reporting of results consistent. To illustrate, standard setting typically results in different raw cut scores across content areas. The raw cut score between *Partially Meeting Expectations* and *Meeting Expectations* could be, for example, 35 in grade 3 mathematics but 33 in grade 4 mathematics, yet both of these raw scores would be transformed to scaled scores of 500. It is this uniformity across scaled scores that facilitates the understanding of student performance. The psychometric advantage of scaled scores over raw scores comes from their being linear transformations of θ. Since the θ scale is used for equating, scaled scores are comparable from one year to the next. Raw scores are not.

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

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

where
m is the slope and
b is the intercept.

A separate linear transformation is used for each grade and content area combination. Table 3-24 shows the slope and intercept terms used to calculate the scaled scores for each grade and content area. Note that the values in Table 3-24 will not change unless the standards are reset.

Appendix J contains raw-score-to-scaled-score look-up tables. The tables show the scaled score equivalent of each raw score for the 2017 next-generation MCAS tests. Additionally, Appendix J

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

Table 3-24. 2017 Next-Generation MCAS: Scaled Score Slopes and Intercepts by Content Area and Grade

Content Area	Grade	Slope	Intercept
ELA	3	18.839	499.785
	4	18.846	499.421
	5	17.686	499.335
	6	18.984	500.202
	7	19.098	499.791
	8	19.900	498.981
Mathematics	3	21.357	499.413
	4	20.938	498.869
	5	19.039	499.525
	6	19.870	500.165
	7	20.758	499.353
	8	20.172	500.170

3.7 MCAS Reliability

Although an individual item's performance is an important factor in evaluating an assessment, an evaluation of the test as a whole 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 next-generation 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 impact 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 3-25 presents descriptive statistics, Cronbach's α coefficient, and raw score SEMs for each content area and grade. Statistics are based on operational items only. For next-generation MCAS ELA tests, the items are the same between online and paper forms,[3] while there are some item differences for mathematics tests. Cronbach's α coefficient was calculated separately for mathematics online and paper forms,[4] but not for ELA. The reliability estimates range from 0.85 to 0.93, which generally are in acceptable ranges.

Table 3-25. 2017 Next-Generation MCAS: Raw Score Descriptive Statistics, Cronbach's Alpha, and SEMs by Content Area and Grade

Content Area	Grade	Mode	Number of Students	Raw Score			Alpha	SEM
				Maximum	*Mean*	*Standard Deviation*		
ELA	3	--	69805	42	23.64	7.23	0.85	2.81
	4	--	63939	42	27.28	7.67	0.88	2.69
	5	--	69098	46	29.56	7.98	0.88	2.76
	6	--	68908	49	27.56	9.83	0.91	3.02
	7	--	70167	49	30.14	9.42	0.89	3.11
	8	--	70135	49	30.05	9.23	0.89	3.09
Mathematics	3	Online	24029	48	29.03	10.42	0.92	3.02
		Paper	45877	48	28.93	10.90	0.92	3.00
	4	Online	59748	54	32.71	11.17	0.91	3.33
		Paper	10473	54	23.91	11.87	0.92	3.37
	5	Online	26635	54	31.35	11.28	0.90	3.51
		Paper	42491	54	29.88	12.10	0.91	3.54
	6	Online	27468	54	28.24	12.17	0.91	3.66
		Paper	41410	54	26.02	13.15	0.92	3.72
	7	Online	28206	54	25.57	12.00	0.92	3.39
		Paper	41934	54	24.26	12.58	0.93	3.42
	8	Online	63102	54	30.10	11.73	0.91	3.48
		Paper	6960	54	19.27	11.13	0.91	3.27

[3] ELA grade 4 had a separate online accommodated form, but it was administered to less than 10% of the state population, so it was not reported in the reliability analyses.
[4] As online accommodated forms share the same items with paper forms, the calculation for paper forms included those students taking the online accommodated forms.

Because of the dependency of the alpha coefficients on the test-taking population and the test characteristics, cautions need be taken when making 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 test-taking population characteristics such as the range of individual differences in the group (i.e., variability within the population), average ability level of the population 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 next-generation MCAS tests. Appendix M 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 0.81 to 0.93 across the tests, with a median of 0.90 and a standard deviation of 0.026, indicating that reliabilities are generally within a reasonable range.

For several reasons, the subgroup reliability results should be interpreted with caution. 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 M shows that subgroup 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 when the population of interest is a single subgroup.

3.7.3 Reporting Subcategory Reliability

Reliabilities were calculated for the reporting subcategories within next-generation 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 M. The reliability coefficients for the reporting subcategories range from 0.25 to 0.85, with a median of 0.70 and a standard deviation of 0.10. Lower reliabilities on subcategory scores are associated with very low numbers of items. 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 the classical test theory (Haertel, 2006), and interpretations should take this into account. Qualitative differences among 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 next-generation MCAS tests, students are classified into one of four achievement levels: *Not Meeting Expectations*, *Partially Meeting Expectations*, *Meeting Expectations*, or *Exceeding Expectations*. Appendix K shows achievement-level distributions by content area and grade for the 2017 next-generation MCAS tests.

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 the consistency of classifications based on a single administration of a test. The Livingston and Lewis (1995) technique was used for the 2017 next-generation MCAS tests because it is easily adaptable to all types of testing formats, including mixed formats.

The DAC estimates reported in Tables 3-27 to 3-30 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 next-generation MCAS tests, after various technical adjustments (described in Livingston & Lewis, 1995), a four-by-four contingency table of accuracy was created for each content area and grade, where cell [*i,j*] represented the estimated proportion of students whose true score fell into classification *i* (where *i* = 1 to 4) and observed score fell into classification *j* (where *j* = 1 to 4). The sum of the diagonal entries (i.e., the proportion of students whose true and observed classifications matched) signified overall accuracy.

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

Measured Progress also measured consistency on the 2017 next-generation 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

DAC analyses were conducted both for the overall population and for subpopulations at each performance achievement level. Due to the adjustment to students' scores on paper forms, achievement-level classifications were based on the adjusted cut scores on paper. Due to mode effect and mode adjustment, DAC estimates were calculated separately for online and paper forms.

Results of the DAC analyses are provided in Table 3-26 and Table 3-27 for the 2017 next-generation MCAS online and paper tests, respectively. The tables include overall accuracy indices with consistency indices displayed in parentheses next to the accuracy values, as well as overall kappa values. Overall ranges for accuracy (0.77–0.84), consistency (0.68–0.77), and kappa (0.51–0.65) 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 0.76 for *Not Meeting Expectations* for the grade 3 ELA online form. This figure indicates that among the students whose true scores placed them in this classification, 76% would be expected to be in this classification when categorized according to their observed scores. Similarly, a consistency value of 0.54 indicates that 54% of students with observed scores in the *Not Meeting Expectations* 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 the Every Student Succeeds Act (ESSA), the primary concern is distinguishing between students who are proficient and those who are not yet proficient. In this case, accuracy at the *Partially Meeting Expectations/Meeting Expectations* threshold is critically important, which summarizes the percentage of students who are correctly classified either above or below the particular cutpoint. Tables 3-28 and 3-29 provide, for the 2017 next-generation MCAS online and paper tests, respectively, the accuracy and consistency estimates and false positive and false negative decision rates at each cutpoint. 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 *Partially Meeting Expectations/Meeting Expectations* threshold shown in Tables 3-28 and 3-29 range from 0.87–0.94 and 0.82–0.91, respectively. The false positive and false negative decision rates at the *Partially Meeting Expectations/Meeting Expectations* threshold range from 4%–7% in Table 3-28 and 4%–6% in Table 3-29. These results indicate that nearly all students were correctly classified with respect to being above or below the *Partially Meeting Expectations/Meeting Expectations* cutpoint.

Table 3-26. 2017 Next-Generation MCAS: Summary of Decision Accuracy and Consistency Results by Content Area and Grade—Overall and Conditional on Achievement Level (CBT Forms)

Content Area	Grade	Overall	Kappa	Conditional on Achievement Level			
				Not Meeting Expectations	*Partially Meeting Expectations*	*Meeting Expectations*	*Exceeding Expectations*
ELA	3	0.78 (0.69)	0.51	0.76 (0.54)	0.81 (0.75)	0.75 (0.67)	0.78 (0.59)
	4	0.80 (0.72)	0.56	0.79 (0.62)	0.83 (0.78)	0.75 (0.67)	0.80 (0.64)
	5	0.82 (0.74)	0.58	0.78 (0.60)	0.84 (0.79)	0.80 (0.74)	0.79 (0.60)
	6	0.82 (0.75)	0.59	0.82 (0.69)	0.84 (0.78)	0.82 (0.77)	0.64 (0.43)
	7	0.82 (0.74)	0.59	0.80 (0.63)	0.83 (0.78)	0.81 (0.74)	0.79 (0.62)
	8	0.80 (0.71)	0.56	0.80 (0.64)	0.82 (0.77)	0.76 (0.69)	0.80 (0.64)
Mathematics	3	0.82 (0.75)	0.62	0.82 (0.71)	0.83 (0.78)	0.83 (0.79)	0.70 (0.52)
	4	0.83 (0.76)	0.62	0.83 (0.71)	0.82 (0.76)	0.84 (0.79)	0.74 (0.54)
	5	0.82 (0.75)	0.60	0.76 (0.59)	0.82 (0.76)	0.83 (0.78)	0.81 (0.64)
	6	0.83 (0.76)	0.62	0.79 (0.64)	0.82 (0.76)	0.85 (0.81)	0.78 (0.61)
	7	0.83 (0.76)	0.62	0.72 (0.55)	0.83 (0.78)	0.85 (0.79)	0.84 (0.72)
	8	0.82 (0.75)	0.61	0.74 (0.58)	0.82 (0.77)	0.83 (0.78)	0.81 (0.68)

Table 3-27. 2017 Next-Generation MCAS: Summary of Decision Accuracy and Consistency Results by Content Area and Grade—Overall and Conditional on Achievement Level (PBT Forms)

Content Area	Grade	Overall	Kappa	Conditional on Achievement Level			
				Not Meeting Expectations	*Partially Meeting Expectations*	*Meeting Expectations*	*Exceeding Expectations*
ELA	3	0.77 (0.68)	0.51	0.78 (0.60)	0.81 (0.75)	0.73 (0.64)	0.78 (0.59)
	4	0.80 (0.72)	0.53	0.81 (0.68)	0.83 (0.79)	0.70 (0.60)	0.77 (0.57)
	5	0.80 (0.72)	0.55	0.80 (0.64)	0.83 (0.78)	0.75 (0.66)	0.80 (0.63)
	6	0.81 (0.74)	0.59	0.84 (0.72)	0.85 (0.80)	0.79 (0.74)	0.61 (0.42)
	7	0.81 (0.73)	0.57	0.81 (0.65)	0.83 (0.78)	0.77 (0.69)	0.80 (0.64)
	8	0.79 (0.71)	0.57	0.86 (0.78)	0.81 (0.74)	0.76 (0.70)	0.56 (0.34)
Mathematics	3	0.82 (0.75)	0.61	0.84 (0.73)	0.84 (0.78)	0.81 (0.76)	0.64 (0.44)
	4	0.84 (0.77)	0.65	0.86 (0.78)	0.83 (0.77)	0.84 (0.78)	0.74 (0.52)
	5	0.83 (0.76)	0.62	0.77 (0.62)	0.84 (0.79)	0.83 (0.78)	0.80 (0.64)
	6	0.83 (0.76)	0.63	0.77 (0.63)	0.82 (0.76)	0.85 (0.80)	0.82 (0.68)
	7	0.81 (0.74)	0.60	0.70 (0.55)	0.80 (0.75)	0.84 (0.78)	0.85 (0.74)
	8	0.81 (0.74)	0.60	0.77 (0.68)	0.81 (0.74)	0.85 (0.79)	0.80 (0.65)

Table 3-28. 2017 Next-Generation MCAS: Summary of Decision Accuracy and Consistency Results by Content Area and Grade—Conditional on Cutpoint (CBT Forms)

Content Area	Grade	Not Meeting Expectations / Partially Meeting Expectations			Partially Meeting Expectations / Meeting Expectations			Meeting Expectations / Exceeding Expectations		
		Accuracy (consistency)	*False*		*Accuracy (consistency)*	*False*		*Accuracy (consistency)*	*False*	
			Positive	*Negative*		*Positive*	*Negative*		*Positive*	*Negative*
ELA	3	0.96 (0.95)	0.01	0.03	0.87 (0.82)	0.07	0.06	0.94 (0.92)	0.04	0.02
	4	0.97 (0.95)	0.01	0.02	0.89 (0.85)	0.06	0.05	0.94 (0.92)	0.04	0.02
	5	0.97 (0.96)	0.01	0.02	0.89 (0.85)	0.06	0.05	0.96 (0.94)	0.03	0.01
	6	0.97 (0.95)	0.01	0.02	0.91 (0.87)	0.05	0.05	0.95 (0.92)	0.04	0.01
	7	0.97 (0.95)	0.01	0.02	0.90 (0.85)	0.05	0.05	0.95 (0.94)	0.03	0.01
	8	0.96 (0.94)	0.01	0.03	0.89 (0.85)	0.06	0.05	0.94 (0.92)	0.04	0.02
Mathematics	3	0.96 (0.94)	0.02	0.03	0.92 (0.89)	0.04	0.04	0.95 (0.93)	0.04	0.02
	4	0.96 (0.94)	0.02	0.03	0.91 (0.87)	0.05	0.05	0.96 (0.94)	0.03	0.01
	5	0.95 (0.93)	0.02	0.03	0.91 (0.87)	0.05	0.05	0.96 (0.95)	0.03	0.01
	6	0.96 (0.94)	0.02	0.03	0.91 (0.88)	0.04	0.04	0.96 (0.95)	0.02	0.01
	7	0.95 (0.92)	0.02	0.04	0.92 (0.88)	0.04	0.04	0.97 (0.95)	0.02	0.01
	8	0.95 (0.93)	0.02	0.03	0.91 (0.88)	0.04	0.04	0.96 (0.94)	0.03	0.01

Table 3-29. 2017 Next-Generation MCAS: Summary of Decision Accuracy and Consistency Results by Content Area and Grade—Conditional on Cutpoint (PBT Forms)

Content Area	Grade	Not Meeting Expectations / Partially Meeting Expectations			Partially Meeting Expectations / Meeting Expectations			Meeting Expectations / Exceeding Expectations		
		Accuracy (consistency)	*False*		*Accuracy (consistency)*	*False*		*Accuracy (consistency)*	*False*	
			Positive	*Negative*		*Positive*	*Negative*		*Positive*	*Negative*
ELA	3	0.95 (0.93)	0.01	0.03	0.88 (0.83)	0.06	0.06	0.94 (0.92)	0.04	0.02
	4	0.94 (0.91)	0.02	0.04	0.90 (0.86)	0.06	0.04	0.96 (0.95)	0.03	0.01
	5	0.96 (0.95)	0.01	0.03	0.89 (0.85)	0.06	0.05	0.94 (0.92)	0.04	0.02
	6	0.97 (0.95)	0.01	0.02	0.91 (0.88)	0.04	0.05	0.93 (0.91)	0.05	0.02
	7	0.96 (0.94)	0.01	0.03	0.90 (0.85)	0.05	0.05	0.95 (0.93)	0.03	0.02
	8	0.94 (0.92)	0.02	0.03	0.90 (0.87)	0.05	0.05	0.95 (0.93)	0.05	0.01
Mathematics	3	0.96 (0.94)	0.02	0.03	0.92 (0.88)	0.04	0.04	0.94 (0.92)	0.04	0.02
	4	0.93 (0.91)	0.03	0.04	0.92 (0.89)	0.04	0.04	0.98 (0.97)	0.01	0.01
	5	0.95 (0.93)	0.02	0.03	0.91 (0.88)	0.05	0.04	0.97 (0.95)	0.02	0.01
	6	0.95 (0.92)	0.02	0.03	0.92 (0.88)	0.04	0.04	0.97 (0.95)	0.02	0.01
	7	0.93 (0.90)	0.03	0.04	0.92 (0.89)	0.04	0.04	0.96 (0.95)	0.02	0.01
	8	0.89 (0.85)	0.05	0.05	0.94 (0.91)	0.04	0.03	0.98 (0.97)	0.01	0.01

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

As with other methods of evaluating reliability, DAC statistics that are calculated based on small groups can be expected to be lower than those calculated based on larger groups. For this reason, the values presented in Tables 3-26 through 3-29 should be interpreted with caution. In addition, it is important to remember that it might be inappropriate to compare DAC statistics across grades and content areas.

3.8 Reporting of Results

The next-generation MCAS tests are designed to measure student achievement on 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 for ELA and mathematics for students in grades 3–8: *Not Meeting Expectations, Partially Meeting Expectations, Meeting Expectations,* and *Exceeding Expectations.* Students receive a separate achievement-level classification in each content area. Reports are generated at the student level, school level, and district level.

Parent/Guardian Reports and student results labels are the only printed reports, and are mailed to districts for distribution to parents/guardians and schools. See section 3.8.1 below for additional details of the *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* was completely redesigned in 2017 to support the next-generation MCAS assessments. The *Parent/Guardian Report* is a stand-alone single page (11" x 17") color report that is folded, and is generated for each student eligible to take the MCAS tests. Two full color copies of each student's report are printed: one for the parent/guardian and one for the school's records. Two sample reports are provided in Appendix N. 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 information. The 2017 revisions were undertaken with input from the MCAS Technical Advisory Committee, and also from parent focus groups held in several towns across the state, with participants from various backgrounds.

The front cover of the *Parent/Guardian Report* provides student identification information, including student name, grade, date of birth, ID (SASID), school name, and district name. The cover also presents general information about the test, website information for parent/guardian resources, and, new for 2017, a summary of the student's results for each content area. This summary provides important information for each content area at a glance, including the student's achievement level, scaled score, and range of scores.

The inside portion of the report contains the achievement level, scaled score, and standard error of the scaled score for each content area tested. If the student does not receive a scaled score, the reason is displayed under the heading "Your Child's Achievement Level." Each achievement level has its own distinct color, and that color is used throughout the report, to highlight important report elements based on the student's achievement level and score. These report elements include the student's earned achievement level, scaled score, the visual scale's achievement-level title and achievement-level cut scores, and the comparison of the student's scaled score to the average scaled score at the student's school, district, and the state levels.

For ELA and mathematics, the student's scaled score is compared to the average scaled score earned by all students at the school, district, and state levels. These scaled score values are color-coded based on the corresponding achievement levels. The student's performance in each content area's reporting categories is also displayed using pictographs and text that indicates the points earned by the student versus the total points possible in that reporting category. For each reporting category, the average number of points earned by students scoring close to 500 is also displayed for comparison purposes. The student's performance on individual test questions is reported at the bottom of the results page in a simplified item response grid. The grid indicates the points earned and points possible for each test question. A link to an external resource is also provided for parents/guardians who wish to review test question descriptions on the department's website. Students who tested only in ELA and mathematics received a report with a back page that provides important information about the Leading the Nation program for 2017–18.

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

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

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

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

One copy of each student label was 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 which students will be excluded from school-, district-, and state-level summary computations. Copies of the decision rules for the 2017 next-generation MCAS administration are included in Appendix O.

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 the data team, the sending unit verifies that the data are accurate before handoff. Additionally, when a unit receives a data set, the first step is to verify the accuracy of the data. Once new report designs were approved by the ESE, reports were run using demonstration data to test the application of the decision rules. The populated reports were 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 next-generation 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 education stakeholders 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 ensures that evidence of response process validity is collected and reported for all new MCAS item types used in the next-generation assessments.

In this testing cycle, ESE conducted a learning lab to study the readiness of students and educators in Massachusetts schools to respond to the new ELA essay items that require students to write an essay in response to reading one or two (related) passages or other literary genres. This learning lab was conducted, prior to forms construction for the 2017 next-generation MCAS assessments, in 10 districts in the state (37 classrooms) in grades 4–8. Findings from this study spurred the development of additional online practice materials, and some students and educators indicated they were not completely prepared to respond to this new item type. Details on the study and results are provided in Appendix P.

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 linking 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. 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.

3.9.4 Validity Evidence in Relationship to Other Variables

The ESE has begun collecting evidence to evaluate the extent to which the next-generation MCAS assessments measure "student readiness for the next level" of schooling, such as readiness for the next grade level, or readiness for postsecondary education. One early piece of predictive validity suggests that the test is identifying students who are not ready for the next grade. First-year analysis of the relationship between student scores and retention in grade indicates that, in all grades except grade 8, students who did not *Meet Expectations* on both assessments showed higher rates for retention, as shown in Appendix P.

Evidence of convergent validity, also provided in Appendix P, indicated that the portion of the test with selected-response items and the portion of the test with constructed-response items were more highly correlated with the scaled score within the same subject than with the scaled score of a different subject area. Additional analyses will be conducted in this area, including examining the relationship of results on the next-generation MCAS tests with student grades and course-taking patterns.

3.9.5 Efforts to Support the Valid Use of Next-Generation MCAS Data

The ESE takes many steps to support the intended uses of MCAS data. (The intended uses are listed in section 2.3 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 achievement results and the longitudinal student growth percentiles derived from them are used as key indicators in the state's accountability formulas for schools and districts.[5] The accountability formulas for schools and districts also factor in assessment participation rates. Information on the state's accountability system is available on the ESE website at www.mass.gov/edu/government/departments-and-boards/ese/programs/accountability/reports/.

As documented on the accountability web page listed above, the ESE carefully weighs all available evidence prior to rendering accountability decisions for schools and districts. No school, for

[5] Accountability for educators is addressed in the ESE's *Educator Evaluation Framework* document, which is available at www.doe.mass.edu/edeval/.

instance, is placed in Levels 4 or 5 without an agency-wide review of data, which factors in trends and subjective indicators alongside several years of assessment data. Assignment to a lower accountability level comes with increased assistance and involvement of the ESE with local education agencies (LEAs).

In 2017, schools that administered the next-generation MCAS tests were not assigned an accountability level unless participation rates fell below 90%. A new school and district accountability and assistance framework is expected to be adopted and put into place in spring 2018.

Finally, students with significant disabilities who are unable to take the MCAS exams even when accommodations are provided can participate in the MCAS-Alt program, which allows students to submit a portfolio of work that demonstrates their proficiency on the state standards. Technical information on the MCAS-Alt program is presented in Chapter 4 of this report.

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

3. 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 samples provided in Appendix N) were designed with input from groups of parents. These reports contain scaled scores and achievement levels from the current year and prior years, as well as norm-referenced student growth percentiles, which calculate how a student's current score compares to that of students who scored similarly on the prior one or two tests in that subject. 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, and to use state-level data for programmatic and diagnostic purposes. These reports can help educators review patterns in the schools and classrooms that students attended in the past, or make plans for the schools and classrooms the students are assigned to in the coming year. The ESE monitors trends in report usage in Edwin Analytics. Between June and November (the peak reporting season for MCAS), over one million reports are run in Edwin Analytics, with approximately 400,000 reports generated in August when schools review their preliminary assessment results in preparation for the return to school.

Examples of two of the most popular reports are provided on the following pages. The *MCAS School Results by Standards* report, shown in Figure 3-2, 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.

The *MCAS Growth Distribution* report, shown in Figure 3-3, presents the distribution of students by student growth percentile band across years. For each year, the report also shows the median student growth percentile and the percentage of students scoring *Proficient or Higher* (or, for 2017, *Meeting or Exceeding Expectations*). 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.

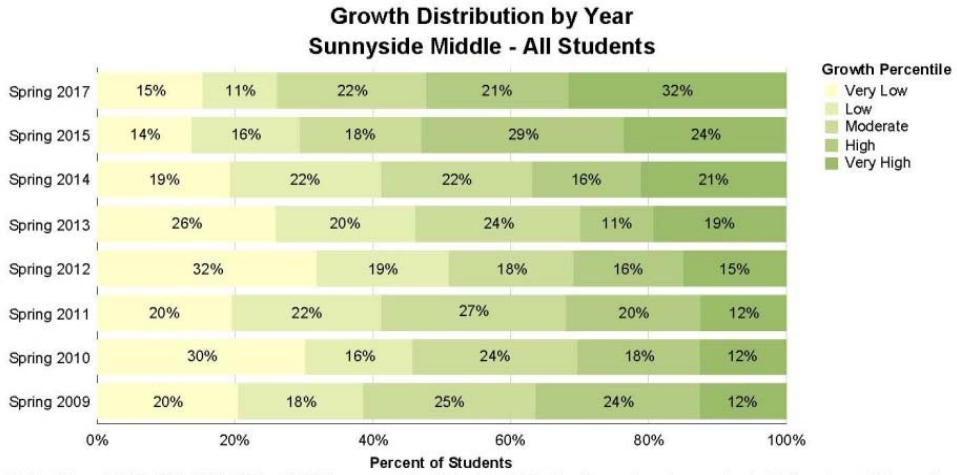
Figure 3-2. 2017 Next-Generation MCAS: School Results by Standards Report

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

Figure 3-3. 2017 Next-Generation MCAS: Growth Distribution Report



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

Test	Very Low	Low	Moderate	High	Very High	Median SGP	N Students (SGP)	% Proficient or Higher	% Meeting or Exceeding Expectations	N Students (Ach. Level)
Spring 2017	17	12	24	23	35	63.0	111		42	116
Spring 2015	14	16	18	30	24	63.0	102	50		110
Spring 2014	22	25	25	18	24	47.0	114	57		119
Spring 2013	27	21	25	11	20	42.5	104	45		111
Spring 2012	30	18	17	15	14	40.0	94	53		99
Spring 2011	19	21	26	19	12	51.0	97	58		105

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

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

Chapter 4 MCAS Alternate Assessment (MCAS-Alt)

4.1 Overview

4.1.1 Background

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

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

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

4.1.2 Purposes of the Assessment System

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

The purposes of the MCAS-Alt are to

- include difficult-to-assess students in statewide assessment and accountability systems;
- determine whether students with significant disabilities are receiving a program of instruction based on the state's academic learning standards;
- determine how much the student has learned in the specific areas of the academic curriculum being assessed;
- assist teachers in providing challenging academic instruction; and
- provide an opportunity for some students with significant disabilities to earn a CD and become eligible to receive a high school diploma.

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

4.1.3 Format

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

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

4.2 Test Design and Development

4.2.1 Test Content and Design

MCAS-Alt assessments are required for all grades and content areas in which standard MCAS tests are administered. However, in the MCAS-Alt, the range and level of complexity of the standards being assessed have been modified without altering the essential components or meanings of the standards. Specific MCAS-Alt content areas and strands/domains required for students in each grade level are listed in Table 4-1.

Table 4-1. 2017 MCAS-Alt: Requirements

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

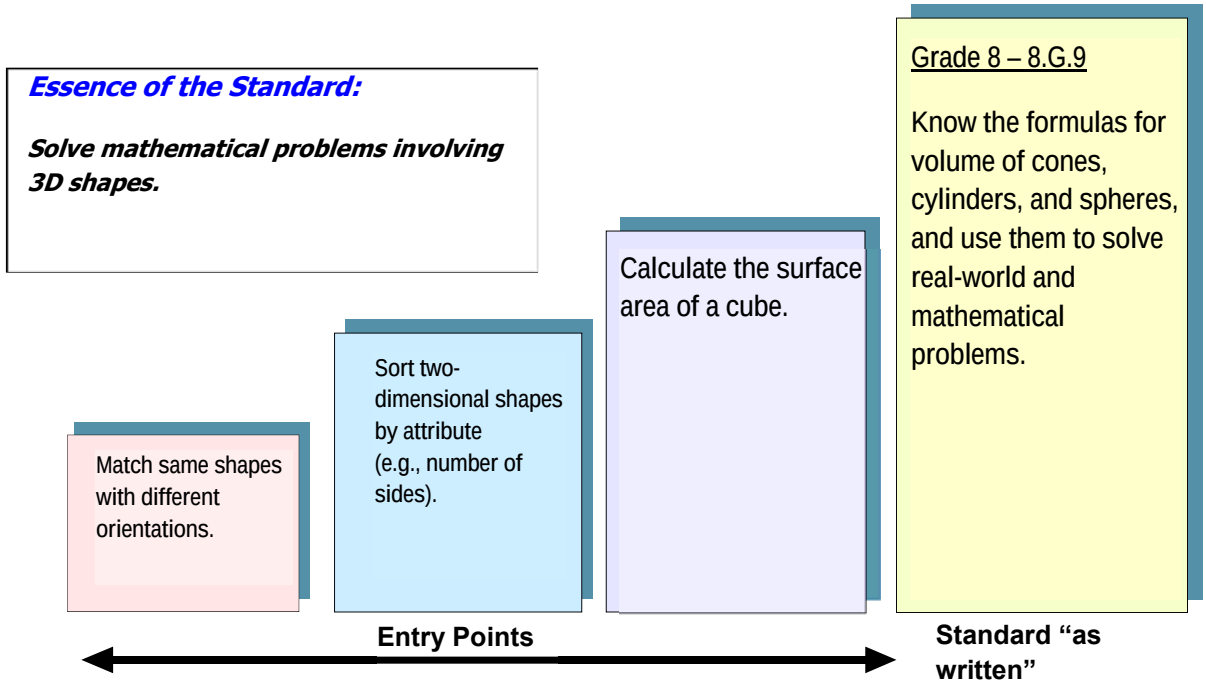
* Earth and Space Science, Life Science, Physical Sciences, Technology/Engineering

4.2.1.1 Access to the Grade-Level Curriculum

The Fall 2016 *Resource Guide* is the basis for determining appropriate curriculum goals that engage and challenge each student based on the curriculum framework learning standards at each grade level.

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

Figure 4-1. 2017 MCAS-Alt: Access to the Grade-Level Curriculum (Mathematics Example) Through Entry Points That Address the Essence of the Standard



4.2.1.2 Assessment Design

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

Primary Evidence

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

One of the three pieces must be a data chart (e.g., field data chart, line graph, or bar graph) that indicates

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

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

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

Each piece of primary evidence must clearly show the final product of an instructional activity and be labeled with

- the student's name,
- the date of the activity,
- a brief description of how the task or activity was conducted and what the student was asked to do,
- the percentage of accuracy for the performance, and
- the percentage of independence for the performance (i.e., the degree to which the student demonstrated knowledge and skills without the use of prompts or cues from the teacher).

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

Supporting Documentation

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

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

4.2.1.3 Assessment Dimensions (Scoring Rubric Areas)

Trained and qualified scorers examine each piece of evidence in the strand and apply the criteria described in the Guidelines for Scoring MCAS-Alt Portfolios (available at www.doe.mass.edu/mcas/alt/results.html), using the Rubric for Scoring Portfolio Strands, to produce a subscore for the strand based on the following:

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

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

Each portfolio strand is scored in each of five rubric dimensions, further described in section 4.4.3.1:

- Level of Complexity (score range of 1–5)
- Demonstration of Skills and Concepts (M, 1–4)
- Independence (M, 1–4)
- Self-Evaluation (M, 1, 2)
- Generalized Performance (1, 2)

(Note: a score of "M" means there was insufficient evidence or information to generate a numerical score in a dimension.)

Scores in Level of Complexity, Demonstration of Skills and Concepts, and Independence are combined, as shown on pp. 104–105, to yield a strand subscore; those subscores are combined, as shown in Appendix T, to yield an overall score in the content area.

4.2.1.4 MCAS-Alt Competency and Grade-Level Portfolios

A relatively small number of MCAS-Alt competency portfolios and grade-level portfolios are submitted each year for students who address learning standards at or near grade-level expectations but who are unable to participate in standard MCAS testing, even when accommodations are provided, due to a significant disability. Participation rates for 2017 are provided in section 4.3.3.3.

The participation guidelines section of the *Educator's Manual for MCAS-Alt* (available at www.doe.mass.edu/mcas/alt/resources.html) describes the characteristics of those students for whom it may be appropriate to submit grade-level and/or competency portfolios. For additional information on how the 2017 MCAS-Alt grade-level and competency portfolios were evaluated, see section 4.4 of this report.

Competency Portfolios

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

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

Grade-Level Portfolios

For students in grades 3–8, a grade-level portfolio may be submitted that assesses a broader range of standards than those assessed in the basic MCAS-Alt portfolio, if the student is working at or close to grade-level expectations and wishes to earn a score of *Partially Meeting Expectations* (ELA, Mathematics) or *Needs Improvement* (Science and Tech/Engineering) or higher on the assessment. Otherwise, a student's score on the MCAS-Alt will be rated as either *Progressing*, *Emerging*, or *Awareness*.

4.2.2 Test Development

4.2.2.1 Rationale

Alternate assessment is the component of the state's assessment system that measures the academic performance of students with the most significant disabilities. Students with disabilities are required by federal and state laws to participate in the MCAS so that their performance of skills and knowledge of content described in the state's curriculum frameworks can be assessed, and also so they can be visible and accountable in reports of results for each school and district.

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

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

4.2.2.2 Role of the Advisory Committee

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

4.3 Test Administration

4.3.1 Evidence Collection

Strands Other Than ELA–Writing

Each portfolio strand (with the exception of ELA–Writing) must include a data chart documenting the student's performance (i.e., the percentage of accuracy and independence of the performance) and progress (whether the rates of accuracy and/or independence increase over time) in learning a new academic skill related to the standard(s) required for assessment. Data are collected on at least

eight different dates to determine whether progress has been made and the degree to which the skill has been mastered. On each date, the data point must indicate the percentage of correct versus inaccurate responses given on that date and whether the student required cues or prompts to respond (i.e., the overall percentage of independent responses given by the student). Data charts include a brief description of the activity (or activities) conducted on each date, and describe how the task relates to the measurable outcome being assessed. Data may be collected either during routine classroom instruction or during tasks and activities set up specifically for the purpose of assessing the student and may include performance data either from a collection of work samples or from a series of responses to specific tasks summarized for each date.

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

ELA–Writing Strand

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

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

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

4.3.2 Construction of Portfolios

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

- **Artistic cover** designed and produced by the student and inserted in the front window of the three-ring portfolio binder
- **Portfolio cover sheet** containing important information about the student
- **Student's introduction** to the portfolio produced as independently as possible by the student using his or her primary mode of communication (e.g., written, dictated, or recorded on video or audio) describing "What I want others to know about me as a learner and about my portfolio"
- **Verification form** signed by a parent, guardian, or primary care provider signifying that he or she has reviewed the student's portfolio or, at minimum, was invited to do so (in the event no signature was obtained, the school must include a record of attempts to invite a parent, guardian, or primary care provider to view the portfolio)

- **Signed consent form to photograph or audio/videotape a student** (kept on file at the school), if images or recordings of the student are included in the portfolio
- **Weekly schedule** documenting the student's program of instruction, including participation in the general academic curriculum
- **School calendar** indicating dates in the current academic year on which the school was in session
- **Strand cover sheet** describing the accompanying set of evidence addressing a particular outcome
- **Work sample description** attached to each piece of primary evidence, providing required labeling information. (if work sample descriptions are not used, this information must be written directly on each piece).
- **Scoring Rubric (ELA–Writing only)** completed by the teacher submitting the portfolio (as detailed in section 4.4.3.2)

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

4.3.3 Participation Requirements

4.3.3.1 Identification of Students

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

4.3.3.2 Participation Guidelines

A student's IEP team (or 504 plan coordinator, in consultation with other staff) determines how the student will participate in the MCAS for each content area scheduled for assessment, either by taking the test routinely or with accommodations, or by taking the alternate assessment if the student is unable to take the standard test, even when accommodations are provided, because of the severity of his or her disabilities. The participation guidelines section of the *Educator's Manual for MCAS-Alt* (available at www.doe.mass.edu/mcas/alt/resources.html) describes the characteristics of those students who should be considered for the MCAS-Alt. This information is documented in the student's IEP or 504 plan and must be revisited on an annual basis. A student may take the general assessment, with or without accommodations, in one subject and the alternate assessment in another subject.

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

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

The student's team must review the options provided in Figure 4-2. Additional guidance on MCAS-Alt participation is provided in the Commissioner's memo and attachments available at www.doe.mass.edu/mcas/alt/essa/.

Figure 4-2. 2017 MCAS-Alt: Participation Guidelines

OPTION 1

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

OPTION 2

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

OPTION 3

Characteristics of Student's Instructional Program and Local Assessment	Recommended Participation in MCAS
If the student is	***Then***
a) working on learning standards at or near grade-level expectations; ***and*** is b) **sometimes able** to take a paper-and-pencil test, either without accommodations or with one or more accommodation(s); ***but*** c) has a complex and significant disability that does not allow the student to fully demonstrate knowledge and skills on a test of this format and duration, (Examples of complex and significant disabilities for which the student may require an alternate assessment are provided on the following page.)	the student should take the **standard MCAS test**, if possible, with necessary accommodations that are consistent with the instructional accommodation(s) used in the student's instructional program (according to the ESE's accommodations policy) and that are documented in an approved IEP or 504 plan prior to testing. ***However,*** the team may recommend the MCAS-Alt when the nature and complexity of the disability prevent the student from fully demonstrating knowledge and skills on the standard test, even with the use of accommodations; in this case, the **MCAS-Alt grade-level portfolio** (in grades 3–8) or **competency portfolio** (in high school) should be compiled and submitted.

While the majority of students who take alternate assessments have significant *cognitive* disabilities, participation in the MCAS-Alt is not limited to these students. When the nature and complexity of a student's disability present significant barriers or challenges to standardized testing, even with the use of accommodations, the student's IEP team or 504 plan may determine that the student should take the MCAS-Alt through either the grade-level (grades 3–8) or competency portfolio (high school) option, even though the student may be working at or near grade-level expectations.

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

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

4.3.3.3 MCAS-Alt Participation Rates

Across all content areas, a total of 8,242 students, or 1.6% of the assessed population, participated in the 2017 MCAS-Alt in grades 3–10. A slightly higher relative proportion of students in grades 3–8 took the MCAS-Alt compared with students in grade 10, and slightly more students were alternately assessed in ELA than in mathematics. Additional information about MCAS-Alt participation rates by content area is provided in Appendix C, including the comparative rate of participation in each MCAS assessment format (i.e., routinely tested, tested with accommodations, or alternately

assessed). The 2017 MCAS-Alt State Summary is available at www.doe.mass.edu/mcas/alt/results.html.

4.3.4 Educator Training

During October 2016, a total of 3,056 educators and administrators received training on conducting the 2017 MCAS-Alt. Attendees had the option of participating in one of three sessions: an introduction to MCAS-Alt for educators new to the assessment, an update for those with previous MCAS-Alt experience, and an administrator's overview. Topics for the introduction session included the following:

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

Topics for the update session included the following:

- a summary of the statewide 2016 MCAS-Alt results
- changes to the MCAS-Alt requirements for 2017
- how best to address the ELA–Writing strand requirements
- avoiding mistakes that lead to an achievement level *Incomplete*
- reporting results
- using data charts to improve teaching and learning
- competency and grade-level portfolio requirements
- accessing the general curriculum and preparing alternate assessment portfolios for students with the most severe cognitive disabilities

Topics for the administrator's session included the following:

- purposes of MCAS-Alt
- who should take MCAS-Alt
- what MCAS-Alt assesses
- MCAS-Alt results
  - *participation*
  - *performance*
  - *trends over time*
  - *supporting teachers who conduct MCAS-Alt*
- principal's role in MCAS-Alt

During January 2017, a total of 1,287 educators attended training sessions in which they were able to review and discuss their students' portfolios and have their questions answered by MCAS-Alt training specialists (i.e., expert teachers).

These training sessions were repeated in February and March 2017, with an additional 911 educators in attendance.

4.3.5 Support for Educators

A total of 86 MCAS-Alt Training Specialists were trained by the ESE to provide assistance and support for teachers conducting the MCAS-Alt in their districts, as well as to assist the Department at eight Department-sponsored portfolio review training sessions in January, February, and March 2017. In addition, ESE staff provided ongoing technical assistance throughout the year via e-mail and telephone to educators with specific questions about their portfolios.

The MCAS Service Center provided toll-free telephone support to district and school staff regarding test administration, reporting, training, materials, and other relevant operations and logistics. The Measured Progress project management team provided extensive training to the MCAS Service Center staff on the logistical, programmatic, and content-specific aspects of the MCAS-Alt, including web-based applications used by the districts and schools to order materials and schedule shipment pickups. Informative scripts were used by the Service Center coordinator and approved by the ESE to train Service Center staff in relevant areas such as web support, enrollment inquiries, and discrepancy follow-up and resolution procedures.

4.4 Scoring

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

Scorers were rigorously trained and qualified based on the *2017 Guidelines for Scoring MCAS-Alt Portfolios*. The *MCAS-Alt Rubric for Scoring Portfolio Strands* has been used as the basis for scoring portfolios since 2001 when it was first developed with assistance from teachers and the statewide advisory committee. The criteria for scoring portfolios are listed and described in detail on the following pages.

4.4.1 Scoring Logistics

MCAS-Alt portfolios were scored in Dover, New Hampshire, during April and May 2017. The ESE and Measured Progress trained and closely monitored scorers to ensure that portfolio scores were accurate.

Portfolios were reviewed and scored by trained scorers according to the procedures described throughout section 4.4. Scores were entered into a computer-based scoring system designed by Measured Progress and the ESE, and scores were frequently monitored for accuracy and completeness.

Security was maintained at the scoring site by restricting access to unscored portfolios to ESE and Measured Progress staff, and by locking portfolios in a secure location before and after each scoring day.

MCAS-Alt scoring leadership staff included several floor managers (FMs) who monitored the scoring room. Each FM managed a group of tables at the elementary, middle, or secondary level. A Table Leader (TL) was responsible for managing a single table with four to five scorers. Communication and coordination among scorers was maintained through daily meetings between

FMs, TLs, and scoring leadership to ensure that critical information and scoring rules were implemented across all grade clusters.

4.4.2 Recruitment, Training, and Qualification of Scorers, Table Leaders, and Floor Managers

4.4.2.1 Scorer Training Materials

The MCAS-Alt Project Leadership Team (PLT), including ESE and Measured Progress staff plus four teacher consultants, met daily over the course of scoring in 2017, and periodically throughout the 2016–2017 school year to accomplish the following:

- nominate prospective scorers and scoring leaders for the 2017 scoring center
- select sample portfolio strands to use during training the following fall, and to train, calibrate, and qualify scorers in 2017
- discuss issues and themes to be addressed during the following fall educator training sessions

All sample strands were scored using the 2017 scoring guidelines, noting any scoring problems that arose during the review. Concerns were resolved by using the *Educator's Manual for MCAS-Alt* and by following additional scoring rules agreed upon by the PLT and subsequently addressed in the final 2017 scoring guidelines.

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

Each of these samples was triple-scored. Of the triple scores, only scores in exact agreement in all five scoring dimensions—Level of Complexity, Demonstration of Skills and Concepts, Independence, Self-Evaluation, and Generalized Performance—were considered as possible exemplars.

4.4.2.2 Recruitment

Through Kelly Services, Measured Progress recruited prospective scorers and TLs for the MCAS-Alt Scoring Center. All TLs and many scorers had previously worked on scoring projects for other states' test or alternate assessment administrations, and all had four-year college degrees.

Additionally, the PLT recruited MCAS-Alt Training Specialists, many of whom had previously served as TLs or scorers, to assist the ESE and Measured Progress. Sixteen MCAS-Alt Training Specialists were selected to participate in portfolio scoring and were designated as expert scorers who assisted in verifying that scores of "M" (indicating that evidence was missing or insufficient to determine a score) were accurate, and in the training/retraining of TLs.

4.4.2.3 Training

Scorers

Scorers were rigorously trained in all rubric dimensions. Scorers reviewed scoring rules and participated in the "mock scoring" of numerous sample portfolio strands selected to illustrate examples of each rubric score point. Scorers were given detailed instructions on how to review data charts and other primary evidence to tally the rubric area scores using a strand organizer. Trainers

facilitated discussions and review among scorers to clarify the rationale for each score point and describe special scoring scenarios and exceptions to the general scoring rules.

Table Leaders and Floor Managers

In addition to the training received by scorers, TLs and FMs received training in logistical, managerial, and security procedures.

4.4.2.4 Qualification

Scorers

Before scoring actual student portfolios, each scorer was required to take a qualifying assessment consisting of 23 questions and to score four sample portfolio strands (i.e., 20 scoring dimensions). To qualify as a scorer, the threshold score on the 23 questions was 85% (20 correct out of 23 total questions); and the threshold score on the portfolio strands was 85% exact agreement overall for the five scoring dimensions (i.e., exact agreement on 17 out of 20 scorable dimensions for the four strands).

Scorers who did not achieve the required percentages were retrained using another qualifying assessment. Those who achieved the required percentages were authorized to begin scoring student portfolios. If a scorer did not meet the required accuracy rate on the second qualifying assessment, he or she was released from scoring.

Table Leaders and Floor Managers

TLs and FMs were qualified by the ESE using the same methods and criteria used to qualify scorers, except they were required to achieve a score of 90% correct or higher on both portions of the qualifying test.

4.4.3 Scoring Methodology

4.4.3.1 All Subjects Except ELA - Writing

Guided by a TL, four or five scorers at each table reviewed and scored portfolios from the same grade. Scorers were permitted to ask TLs questions as they reviewed portfolios. In the event a TL could not answer a question, the FM provided assistance. In the event the FM was unable to answer a question, ESE staff members were available to provide clarification.

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

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

A. Level of Complexity
B. Demonstration of Skills and Concepts
C. Independence

D. Self-Evaluation
E. Generalized Performance

The 2017 MCAS-Alt score distributions for all scoring dimensions are provided in Appendix H.

During spring 2017, scorers used an automated, customized scoring program called *AltScore* to score MCAS-Alt portfolios. Scorers were guided through the scoring process by answering a series of yes/no and fill-in-the-blank questions onscreen which were used by the program to calculate the correct score. Use of the computer-based scoring application allowed scorers to focus exclusively and sequentially on each portfolio product and record the necessary information, rather than keeping track of products they had previously reviewed and calculating the score.

A. Level of Complexity

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

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

Table 4-2. 2017 MCAS-Alt: Scoring Rubric for Level of Complexity

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

B. Completeness

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

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

- the student's name

- the date of performance
- a brief description of the activity
- the percentage of accuracy
- the percentage of independence

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

Brief descriptions of each activity on the data chart were also considered in determining the completeness of a data chart. Educators had been instructed during educator training workshops and in the *2017 Educator's Manual for MCAS-Alt* that "each data chart must include a brief description beneath each data point that clearly illustrates how the task or activity relates to the measurable outcome being assessed." One- or two-word descriptions were likely to be considered insufficient to document the relationship between the activity and the measurable outcome and therefore would result in the exclusion of those data points from being scored.

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

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

- The data chart listed the percentages of both accuracy and independence at or above 80% at the beginning of the data collection period, indicating that the student did not learn a challenging new skill in the strand and was instead addressing a skill he or she already had learned.
- The data chart did not document the measurable outcome on at least 8 distinct dates; the measurable outcome was not based on a required learning standard or strand; and/or the evidence did not indicate the student's accuracy and independence on each task or trial.
- Two additional pieces of primary evidence did not address the same measurable outcome as the data chart or were not labeled with all required information.

C. Demonstration of Skills and Concepts

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

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

Based on the average percentage of accuracy in the data points and evidence in the final 1/3 time frame, the overall score in the strand was determined using the rubric shown in Table 4-3.

Table 4-3. 2017 MCAS-Alt: Scoring Rubric for Demonstration of Skills and Concepts

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

D. Independence

The score for Independence shows the degree to which the student responded without cues or prompts during tasks or activities based on the measurable outcome being assessed. For strands that included a "core set of evidence," Independence was scored first by identifying the final 1/3 time frame on the data chart (or the final three data points, if fewer than 12 points were listed). Then an average percentage was calculated based on the percent of independence for

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

Based on the average percent of independence of the data points and evidence in the final 1/3 time frame, the overall score in the strand was determined using the rubric shown in Table 4-4 below.

A score of M was given both in Demonstration of Skills and Concepts and in Independence if any of the following was true:

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

Table 4-4. 2017 MCAS-Alt: Scoring Rubric for Independence

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

E. Self-Evaluation

The score for Self-Evaluation indicates the frequency of activities in the portfolio strand that involve self-correction, task-monitoring, goal-setting, reflection, and overall awareness by the student of his or her own learning. Each strand was given a score of M, 1, or 2 based on the scoring rubric shown in Table 4-5.

Table 4-5. 2017 MCAS-Alt: Scoring Rubric for Self-Evaluation, Individual Strand Score

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

F. Generalized Performance

The score for Generalized Performance reflects the number of contexts and instructional approaches used by the student to demonstrate knowledge and skills in the portfolio strand. Each strand was given a score of either 1 or 2 based on the rubric shown in Table 4-6.

Table 4-6. 2017 MCAS-Alt: Scoring Rubric for Generalized Performance

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

4.4.3.2 ELA–Writing

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

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

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

Additional Information about ELA–Writing:

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

4.4.4 Monitoring Scoring Quality

The FM oversaw the general flow of work in the scoring room and monitored overall scoring consistency and accuracy, particularly among TLs. The TLs ensured that scorers at their table were consistent and accurate in their scoring. Scoring consistency and accuracy were maintained using two methods: double-scoring and resolution (i.e., read-behind) scoring.

4.4.4.1 Double-Scoring

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

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

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

Table 4-10 in section 4.7.3 shows the percentages of interrater agreement for the 2017 MCAS-Alt.

4.4.4.2 Resolution Scoring

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

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

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

4.4.4.3 Tracking Scorer Performance

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

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

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

Specific requirements for submission of grade-level and competency portfolios are described in the *Educator's Manual for MCAS-Alt*. Section 4.2.1.4 of this report also discusses grade-level and competency portfolios.

4.4.5.1 Grade-Level Portfolios in Grades 3–8

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

Each grade-level portfolio was evaluated by a panel of content area experts to determine whether it achieved a score of *Partially Meeting Expectations* (ELA, mathematics) or *Needs Improvement* (STE) or higher. To receive an achievement level at or above *Partially Meeting Expectations* or *Needs Improvement*, the portfolio must have demonstrated

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

4.4.5.2 Competency Portfolios in High School

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

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

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

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

4.5 MCAS-Alt Classical Item Analyses

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

Quantitative analyses presented here are based on the statewide administration of the 2017 MCAS-Alt and include three of the five dimension scores on each task (Level of Complexity, Demonstration of Skills and Concepts, and Independence). Although the other two dimension scores (Self-Evaluation and Generalized Performance) are reported, they do not contribute to a student's overall achievement level; therefore, they are not included in quantitative analyses.

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

Statistical evaluations of MCAS-Alt include difficulty and discrimination indices, structural relationships (correlations among the dimensions), and bias and fairness. Item-level classical statistics—item difficulty and discrimination values—are provided in Tables G-14 and G-15 of Appendix G. Item-level score distributions for each item (i.e., the percentage of students who

received each score point) are provided in Tables H-3 and H-4 of Appendix H. Note that the Self-Evaluation and Generalized Performance dimension scores are also included in Appendix H.

4.5.1 Difficulty

Following from the definition of dimensions and dimension scores as similar to traditional test items and scores, all items are evaluated in terms of difficulty according to standard classical test theory practices. Difficulty is traditionally described according to an item's *p*-value, which is calculated as the average proportion of points achieved on the item. Dimension scores achieved by each student are divided by the maximum possible score to return the proportion of points achieved on each item; *p*-values are then calculated as the average of these proportions. Computing the difficulty index in this manner places items on a scale that ranges from 0.0 to 1.0. This statistic is properly interpreted as an "easiness index," because larger values indicate easier items. An index of 0.0 indicates that all students received no credit for the item, and an index of 1.0 indicates that all students received full credit for the item.

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

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

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

4.5.2 Discrimination

Discrimination indices can be thought of as measures of how closely an item assesses the same knowledge and skills assessed by other items contributing to the criterion total score. That is, the discrimination index can be thought of as a measure of construct consistency. The correlation between student performance on a single item and total test score is a commonly used measure of this characteristic of an item. Within classical test theory, this item-test correlation is referred to as the item's discrimination, because it indicates the extent to which successful performance on an item discriminates between high and low scores on the test. A desirable feature of an item is that the higher-ability students perform better on the item than lower-ability students or that the item demonstrates strong, positive item-test correlation.

In light of this interpretation, the selection of an appropriate criterion total score is crucial to the interpretation of the discrimination index. For the MCAS-Alt, the sum of the three dimension scores, excluding the item being evaluated, was used as the criterion score. For example, in grade 3 ELA, total test score corresponds to the sum of scores received on the three dimensions included in quantitative analyses (i.e., Level of Complexity, Demonstration of Skills and Concepts, and Independence) across both Language and Reading strands.

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

Table 4-7. 2017 MCAS-Alt: Summary of Item Difficulty and Discrimination Statistics by Content Area and Grade

Content Area	Grade	Number of Items	*p*-Value		Discrimination	
			Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
ELA	3	9	0.79	0.20	0.39	0.05
	4	9	0.79	0.19	0.40	0.08
	5	9	0.80	0.19	0.36	0.09
	6	9	0.80	0.19	0.39	0.09
	7	9	0.79	0.19	0.40	0.08
	8	9	0.80	0.19	0.42	0.09
	HS	9	0.79	0.19	0.37	0.07
Mathematics	3	9	0.84	0.20	0.60	0.09
	4	12	0.85	0.20	0.62	0.08
	5	9	0.85	0.20	0.61	0.06
	6	9	0.85	0.20	0.60	0.14
	7	9	0.85	0.20	0.58	0.07
	8	9	0.84	0.19	0.59	0.10
	HS	15	0.84	0.18	0.38	0.07
STE	5	12	0.85	0.19	0.37	0.05
	8	12	0.85	0.19	0.40	0.09
Biology	HS	12	0.85	0.19	0.37	0.07
Chemistry	HS	12	0.85	0.19	0.51	0.19
Introductory Physics	HS	12	0.85	0.15	0.60	0.17
Technology/ Engineering	HS	9	0.83	0.19	0.44	0.17

4.5.3 Structural Relationships Among Dimensions

By design, the achievement-level classification of the MCAS-Alt is based on three of the five scoring dimensions (Level of Complexity, Demonstration of Skills and Concepts, and Independence). As with any assessment, it is important that these dimensions be carefully examined. This was achieved by exploring the relationships among student dimension scores with Pearson

correlation coefficients. A very low correlation (near zero) would indicate that the dimensions are not related, a low negative correlation (approaching -1.00) indicates that they are inversely related (i.e., that a student with a high score on one dimension had a low score on the other), and a high positive correlation (approaching 1.00) indicates that the information provided by one dimension is similar to that provided by the other dimension. The average correlations among the three dimensions by content area and grade level are shown in Table 4-8.

Table 4-8. 2017 MCAS-Alt: Average Correlations Among the Three Dimensions by Content Area and Grade

Content Area	Grade	Number of Items Per Dimension	Average Correlation Between:*			Correlation Standard Deviation*		
			Comp/ Ind	*Comp/ Sk*	*Ind/ Sk*	*Comp/ Ind*	*Comp/ Sk*	*Ind/ Sk*
ELA	3	3	0.16	0.24	0.19	0.09	0.11	0.09
	4	3	0.10	0.20	0.19	0.05	0.13	0.03
	5	3	0.12	0.13	0.19	0.11	0.17	0.05
	6	3	0.17	0.18	0.15	0.01	0.14	0.08
	7	3	0.13	0.20	0.14	0.06	0.14	0.09
	8	3	0.18	0.25	0.16	0.03	0.06	0.07
	HS	3	0.16	0.21	0.18	0.07	0.13	0.02
Mathematics	3	2	0.13	0.19	0.16	0.04	0.06	0.07
	4	2	0.15	0.21	0.24	0.02	0.02	0.01
	5	2	0.18	0.18	0.21	0.02	0.05	0.03
	6	2	0.21	0.11	0.11	0.04	0.04	0.03
	7	2	0.11	0.18	0.16	0.08	0.01	0.01
	8	2	0.19	0.18	0.13	0.05	0.03	0.01
	HS	5	0.18	0.15	0.17	0.06	0.06	0.06
STE	5	4	0.13	0.13	0.21	0.04	0.02	0.04
	8	4	0.23	0.24	0.13	0.03	0.07	0.07
Biology	HS	4	0.15	0.12	0.22	0.03	0.05	0.03
Chemistry	HS	4	-0.05	0.27	0.21			0.12
Introductory Physics	HS	4	0.26	0.16	0.36	0.08	0.08	0.24
Technology/ Engineering	HS	3	0.27	0.12	0.06	0.07	0.19	0.18

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

The average correlations between every two dimensions range from very weak (0.00 to 0.20) to weak (0.20 to 0.40). It is important to remember in interpreting the information in Table 4-8 that the correlations are based on small numbers of item scores and small numbers of students and should therefore be interpreted with caution.

4.5.4 Differential Item Functioning

The *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004) explicitly states that subgroup differences in performance should be examined when sample sizes permit and that actions should be taken to ensure that differences in performance are because of construct-relevant, rather than irrelevant, factors. *Standards for Educational and Psychological Testing* (AERA et al., 2014) includes similar guidelines.

When appropriate, the standardization differential item functioning (DIF) procedure (Dorans & Kulick, 1986) is employed to evaluate subgroup differences. The standardization DIF procedure is

designed to identify items for which subgroups of interest perform differently, beyond the impact of differences in overall achievement. However, because of the small number of students who take the MCAS-Alt, and because those students take different combinations of tasks, it was not possible to conduct DIF analyses. Conducting DIF analyses using groups of fewer than 200 students would result in inflated type I error rates.

4.6 Bias/Fairness

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

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

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

4.7 Characterizing Errors Associated with Test Scores

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

4.7.1 MCAS-Alt Overall Reliability

In the previous section, individual item characteristics of the 2017 MCAS-Alt were presented. Although individual item performance is an important focus for evaluation, a complete evaluation of an assessment must also address the way in which items function together and complement one

another. Any assessment includes some amount of measurement error; that is, no measurement is perfect. This is true of all academic assessments—some students will receive scores that underestimate their true ability, and others will receive scores that overestimate their true ability. When tests have a high amount of measurement error, student scores are very unstable. Students with high ability may get low scores and vice versa. Consequently, one cannot reliably measure a student's true level of ability with such a test. Assessments that have less measurement error (i.e., errors are small on average, and therefore students' scores on such tests will consistently represent their ability) are described as reliable.

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

The split-half method requires psychometricians to select items that contribute to each half-test score. This decision may have an impact on the resulting correlation, since each different possible split of the test into halves will result in a different correlation. Another problem with the split-half method of calculating reliability is that it underestimates reliability, because test length is cut in half. All else being equal, a shorter test is less reliable than a longer test. Cronbach (1951) provided a statistic, alpha (α), that eliminates the problem of the split-half method by comparing individual item variances to total test variance. Cronbach's α was used to assess the reliability of the 2017 MCAS-Alt. The formula is as follows:

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

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

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

Table 4-9. 2017 MCAS-Alt: Cronbach's Alpha and SEMs by Content Area and Grade

Content Area	Grade	Number of Students	Raw Score			Alpha	SEM
			Maximum Score	*Mean*	*Standard Deviation*		
ELA	3	1,158	39	29.13	3.39	0.63	2.07
	4	1,161	39	29.57	3.36	0.62	2.06
	5	1,226	39	29.5	3.35	0.61	2.10
	6	1,092	39	29.6	3.29	0.63	1.99
	7	1,022	39	29.55	3.37	0.63	2.05
	8	1,041	39	29.54	3.46	0.67	1.98
	HS	821	39	29.02	3.47	0.64	2.08
Mathematics	3	1,039	26	21.43	1.20	0.63	0.73
	4	1,098	26	21.52	1.15	0.67	0.66

continued

Content Area	Grade	Number of Students	Raw Score			Alpha	SEM
			Maximum Score	*Mean*	*Standard Deviation*		
Mathematics	5	1,118	26	21.48	1.20	0.65	0.71
	6	1,017	26	21.45	1.21	0.65	0.72
	7	963	26	21.42	1.21	0.61	0.76
	8	951	26	21.41	1.27	0.62	0.78
	HS	786	39	30.83	3.44	0.86	1.31
STE	5	1,098	39	31.61	2.75	0.79	1.26
	8	969	39	31.46	2.93	0.81	1.28
Biology	HS	609	39	31.27	3.13	0.76	1.54
Chemistry	HS	51	39	32.14	1.71	0.73	0.89
Introductory Physics	HS	44	39	32.11	4.25	0.83	1.75
Technology/ Engineering	HS	74	39	30.5	3.49	0.77	1.66

An alpha coefficient toward the high end (greater than 0.50) is taken to mean that the items are likely measuring very similar knowledge or skills; that is, they complement one another and suggest that the 2017 MCAS-Alt is a reliable assessment.

4.7.2 Subgroup Reliability

The reliability coefficients discussed in the previous section were based on the overall population of students who participated in the 2017 MCAS-Alt. Tables M-7 and M-8 in Appendix M present reliabilities for various subgroups of interest taking MCAS-Alt. Subgroup Cronbach's α coefficients were calculated using the formula defined on the previous page, based only on the members of the subgroup in question in the computations; values are calculated only for subgroups with 10 or more students.

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

4.7.3 Interrater Consistency

Section 4.4 of this chapter describes the processes that were implemented to monitor the quality of the hand-scoring of student responses. One of these processes was double-blind scoring of at least 20% of student responses in all portfolio strands. Results of the double-blind scoring, used during the scoring process to identify scorers who required retraining or other intervention, are presented here as evidence of the reliability of the MCAS-Alt. A third score was required for any score category in which there was not an exact agreement between scorer one and scorer two. A third score was also required as a confirmation score when either scorer one and/or scorer two provided a score of M for Demonstration of Skills and Concepts and Independence or a score of 1 for Level of Complexity.

A summary of the interrater consistency results is presented in Table 4-10. Results in the table are aggregated across the tasks by content area, grade, and number of score categories (five for Level of Complexity and four for Demonstration of Skills and Concepts and Independence). The table shows the number of items, number of included scores, percent exact agreement, percent adjacent agreement, correlation between the first two sets of scores, and the percent of responses that required a third score. This information is also provided at the item level in Tables F-3 and F-4 of Appendix F.

Table 4-10. 2017 MCAS-Alt: Summary of Interrater Consistency Statistics Aggregated across Items by Content Area and Grade

Content Area	Grade	Number of			Percent		Correlation	Percent of Third Scores
		Items	*Score Categories*	*Included Scores*	*Exact*	*Adjacent*		
ELA	3	6	4	1,150	97.91	1.39	0.97	3.65
		3	5	656	98.17	0.76	0.73	5.79
	4	6	4	1,504	98.54	1.20	0.98	2.93
		3	5	860	98.26	0.23	0.67	5.70
	5	6	4	1,828	98.25	1.37	0.98	4.10
		3	5	1,034	98.94	0.29	0.67	5.80
	6	6	4	1,138	98.77	1.05	0.99	2.99
		3	5	634	98.74	0.79	0.84	4.73
	7	6	4	1,648	98.48	1.33	0.98	3.28
		3	5	927	99.03	0.11	0.75	4.75
	8	6	4	2,574	98.37	1.28	0.97	4.39
		3	5	1,451	97.79	0.76	0.65	8.13
	HS	6	4	2,566	97.54	1.83	0.97	7.60
		3	5	1,581	97.22	1.14	0.67	12.97
Mathematics	3	4	4	754	99.60	0.40	0.99	0.80
		2	5	436	98.39	0.92	0.75	1.61
	4	4	4	1,038	98.94	0.96	0.95	1.54
		2	5	585	98.8	0.51	0.76	1.71
	5	4	4	1,184	99.24	0.68	0.96	1.27
		2	5	695	98.85	0.72	0.78	1.44
	6	4	4	760	99.47	0.53	0.98	1.18
		2	5	430	98.14	1.40	0.82	1.86
	7	4	4	1,130	99.12	0.88	0.97	1.24
		2	5	634	98.42	0.63	0.72	2.05
	8	2	5	992	98.69	0.30	0.76	4.33
		10	4	2,618	99.16	0.80	0.98	2.29
	HS	5	5	1,604	95.57	0.94	0.56	8.54
		4	4	754	99.60	0.40	0.99	0.80
STE	5	8	4	1,686	99.70	0.30	0.99	0.71
		4	5	942	99.36	0.11	0.75	0.74
	8	8	4	2,486	99.72	0.28	0.99	1.01
		4	5	1,423	98.81	0.21	0.67	3.65
Biology	HS	6	4	1,726	99.07	0.93	0.97	2.43
		3	5	1,045	96.94	0.48	0.38	6.22
Chemistry	HS	6	4	138	100.00	0.00	1.00	1.45
		3	5	88	98.86	0.00	0.57	5.68
Introductory Physics	HS	6	4	98	100.00	0.00	1.00	4.08
		3	5	59	100.00	0.00		0.00
Technology/ Engineering	HS	6	4	218	99.54	0.46	0.99	0.92
		3	5	126	98.41	1.59	0.94	3.97

4.8 MCAS-Alt Comparability Across Years

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

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

Finally, comparability across years is encouraged through the classification of students into achievement-level categories, using a look-up table that remains consistent each year. The description of each achievement level, shown in Table 4-11, while transitioning in grades 3-8, remains relatively consistent, which ensures that the meaning of students' scores is comparable from one year to the next. Table 4-12 shows the achievement-level look-up table (i.e., the achievement level corresponding to each possible combination of dimension scores), which is used each year to combine and tally the overall content area achievement level from the individual portfolio strand scores. In addition, achievement-level distributions for each of the last three years are provided in Appendix K.

Table 4-11. 2017 MCAS-Alt Achievement-Level Descriptions

Achievement Level	Description
Incomplete (1)	**Insufficient evidence** and information were included in the portfolio to allow a performance level to be determined in the content area.
Awareness (2)	Students at this level demonstrate **very little understanding** of learning standards and core knowledge topics contained in the Massachusetts curriculum framework for the content area. Students require extensive prompting and assistance, and their performance is mostly inaccurate.
Emerging (3)	Students at this level demonstrate a **simple understanding below grade-level expectations** of a limited number of learning standards and core knowledge topics contained in the Massachusetts curriculum framework for the content area. Students require frequent prompting and assistance, and their performance is limited and inconsistent.
Progressing (4)	Students at this level demonstrate a **partial understanding below grade-level expectations** of selected learning standards and core knowledge topics contained in the Massachusetts curriculum framework for the content area. Students are steadily learning new knowledge, skills, and concepts. Students require minimal prompting and assistance, and their performance is basically accurate.
Partially Meeting Expectations (Grades 3-8)/ Needs Improvement (High School) (5)	PME: A student who performed at this level partially met grade-level expectations in this subject. NI: Students at this level demonstrate a **partial understanding of grade-level subject matter** and solve some simple problems.
Meeting Expectations (Grades 3-8)/ Proficient (High School) (6)	ME: A student who performed at this level met grade-level expectations and is academically on track to succeed in the current grade in this subject. P: Students at this level demonstrate a **solid understanding of challenging grade-level subject matter** and solve a wide variety of problems
Exceeding Expectations (Grades 3-8)/ Advanced (High School (7)	EE: A student who performed at this level exceeded grade-level expectations by demonstrating mastery of the subject matter. A: Students at this level demonstrate a **comprehensive understanding of challenging grade-level subject matter** and provide sophisticated solutions to complex problems.

Table 4-12. 2017 MCAS-Alt: Strand Achievement-Level Look-Up

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

continued

Level of Complexity	Demonstration of Skills	Independence	Achievement Level
2	4	4	2
3	1	1	1
3	1	2	1
3	1	3	1
3	1	4	1
3	2	1	1
3	2	2	1
3	2	3	2
3	2	4	2
3	3	1	1
3	3	2	2
3	3	3	3
3	3	4	3
3	4	1	1
3	4	2	2
3	4	3	3
3	4	4	3
4	1	1	1
4	1	2	1
4	1	3	1
4	1	4	1
4	2	1	1
4	2	2	1
4	2	3	2
4	2	4	2
4	3	1	1
4	3	2	2
4	3	3	3
4	3	4	3
4	4	1	1
4	4	2	2
4	4	3	3
4	4	4	3
5	1	1	1
5	1	2	1
5	1	3	2
5	1	4	2
5	2	1	1
5	2	2	2
5	2	3	3
5	2	4	3
5	3	1	1
5	3	2	2
5	3	3	3
5	3	4	4
5	4	1	1
5	4	2	2
5	4	3	3
5	4	4	4

4.9 Reporting of Results

4.9.1 Primary Reports

Measured Progress created two primary reports for the MCAS-Alt: the *Portfolio Feedback Form* and the *Parent/Guardian Report.*

4.9.1.1 Portfolio Feedback Form

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

4.9.1.2 Parent/Guardian Report

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

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

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

4.9.2 Decision Rules

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

4.9.3 Quality Assurance

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

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

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

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

4.10 MCAS-Alt Validity

One purpose of the *2017 Next-Generation MCAS and MCAS-Alt Technical Report* is to describe the technical aspects of the MCAS-Alt that contribute validity evidence in support of MCAS-Alt score interpretations. According to the *Standards for Educational and Psychological Testing* (AERA et al., 2014), considerations regarding establishing intended uses and interpretations of test results and conforming to these uses are of paramount importance in regard to valid score interpretations. These considerations are addressed in this section.

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

4.10.1 Test Content Validity Evidence

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

4.10.2 Internal Structure Validity Evidence

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

4.10.3 Response Process Validity Evidence

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

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

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

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

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

Another important reason to include students with significant disabilities in standards-based instruction is to explore their capacity to learn standards-based knowledge and skills. While "daily living skills" are critical for these students to function as independently as possible, academic skills are extremely important. Standards in the Massachusetts curriculum frameworks are defined as "valued outcomes for all students." Evidence indicates that students with significant disabilities learn more than anticipated when given opportunities to engage in challenging instruction with the necessary support.

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

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

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

4.10.5 Summary

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

REFERENCES

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

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

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

Austin, P. C., & Mamdani, M. M. (2006). A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine 25*, 2084–2106.

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

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

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

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

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

Cai, L., Thissen, D., du Toit, S. H. C. (2011). IRTPRO for Windows [Computer software]. Lincolnwood, IL: Scientific Software International.

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.

Haertel, E. H. (2006). Reliability. In R.L. Brennan (Ed). *Educational measurement (pp. 65-110)*. Westport, CT: Praeger Publishers.

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: Joint Committee on Testing Practices. Available from 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. (2010). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer-Verlag.

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. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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*. Unpublished manuscript.

Measured Progress Psychometrics and Research Department. (2017). *MCAS 2016–2017 IRT & Mode Linking 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.

Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of American Statistical Association 79*, 516–524.

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.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied psychological measurement 7*, 201–210.

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.

Stuart, A. (2010) Matching methods for causal inference: a review and a look forward. *Statistical Science*. *25*(1), 1–21.

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

APPENDICES