MASSACHUSETTS DEPARTMENT OF
ELEMENTARY AND SECONDARY
EDUCATION

2019 Next-Generation MCAS and MCAS-Alt Technical Report

Prepared by Cognia and the
Massachusetts Department of Elementary and Secondary Education

cognia™

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

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

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

Table of Contents

List of Tables

Chapter 1. OVERVIEW

1.1 Purposes of the MCAS and This Report

The Massachusetts Comprehensive Assessment System (MCAS) was originally 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 defines the purposes of the MCAS in Chapter 69 of the Massachusetts General Laws as follows:

- Establish "whether students are meeting the academic standards described," in the state curriculum frameworks, ensuring that "such instruments shall be criterion referenced." (Ch 69, Sec 1I).
- Provide "a comprehensive diagnostic assessment of individual students" in the required grades (Ch. 69, Sec 1I);
- Support the annual publication of assessment results in all public schools, districts, and the state (Ch. 69, Sec 1I);
- Provide a "competency determination," defined as the requirement that all high school graduates have fulfilled a measure of the "mastery of a common core of skills and knowledge" in mathematics, science and technology, English, and history and social sciences. (Ch. 69, Sec. 1D);
- Set and activate goals for high standards of innovation, quality, and accountability in schools (Ch 69, Sec. 1B).

Additional tests and requirements have been added to the MCAS program to meet the requirements of the No Child Left Behind Act of 2001 and the Every Student Succeeds Act (ESSA) of 2015.

The purpose of this *2019 Next-Generation MCAS and MCAS-Alt Technical Report* is to document the technical quality and characteristics of the 2019 next-generation MCAS ELA, mathematics, and grades 5 and 8 science and technology/engineering (STE) tests and of the 2019 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 2019. A companion document, the *2019 Legacy MCAS Technical Report*, provides information regarding the technical quality of the legacy tests administered in 2019: the STE tests in high school.

Technical reports for previous testing years are available on the DESE 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, 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.

In addition, this report provides technical evidence for how the MCAS is designed to fulfill the requirements of the Act described above, as well as federal requirements under ESSA for assessments in English language arts, mathematics, and science and technology/engineering. The MCAS is designed to:

- Assess all students who are educated with Massachusetts public funds in designated grades, including students with disabilities and English learner (EL) students. Massachusetts has an annual state participation rate over 98% across all grades, subjects, and assessments (see section 3.3.3).
- Measure student, school, and district performance in meeting the state's learning standards as detailed in the Massachusetts curriculum frameworks. As described throughout this document, the MCAS tests are designed to measure the standards in the curriculum frameworks. The process for ensuring alignment to the standards begins with the test and item specifications and test blueprints, continues

through the development process with rigorous review by educators and other experts, and culminates with the release of test information (including standards alignment) to students, schools, and districts.

- Provide measures of student achievement that will enable improvements in student outcomes. The scales and achievement levels for the next-generation tests are designed to indicate students' readiness to engage in academic work at the next grade level, and to provide information to parents and students if they are not on track.
- Massachusetts releases significant numbers of test items each year—and provides item descriptions, standards, and other related information for all test questions, whether released or unreleased—to help families and educators better understand how students are being assessed on the content standards and how instruction can be targeted to achieve better outcomes at the individual or aggregate levels.
- Report on the performance of individual students, schools, districts, and the state. Massachusetts provides comprehensive reporting on the results of individual students, schools, districts, and the state through reporting on achievement and growth to parents and families (*Parent/Guardian Reports*), and through dissemination of full results to schools, districts, and the public (see section 3.8 and section 3.9.5).
- Help determine ELA, mathematics, and STE competency for the awarding of high school diplomas. Students must achieve a passing score on the ELA, mathematics, and STE tests (or successfully file an MCAS appeal) as one condition for high school graduation (see the *2010 MCAS and MCAS-Alt Technical Report* as well).

1.2 Organization of This Report

This report provides detailed information regarding test design and development, scoring, and analysis and reporting of 2019 next-generation MCAS and MCAS-Alt 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 appendices contain detailed item-level and summary statistics related to each 2019 next-generation 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 2019. Chapter 2 explains the guiding philosophy, purposes, uses, components, and validity evidence of MCAS. The next two chapters cover 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 validity evidence of results. Numerous appendices are referenced throughout the report.

1.3 Current Year Updates

In 2017, Massachusetts began a transition from the legacy paper-based MCAS tests (administered since 1998) to next-generation MCAS tests that are administered primarily via computer and aligned with the most recent Massachusetts curriculum frameworks. The 2019 MCAS administration marked a continuation of that transition.

Table 1-1 shows which MCAS tests were administered at each grade level in spring 2019 and whether the tests were next-generation (NG) or legacy (L) assessments.

Table 1-1. Spring 2019 MCAS Tests Administered, by Grade Level

Content Area	*Grade Level*							
	3	*4*	*5*	*6*	*7*	*8*	*9*	*10*
English Language Arts	NG	NG	NG	NG	NG	NG		NG
Mathematics	NG	NG	NG	NG	NG	NG		NG
Science and Technology/Engineering			NG			NG	L*	L*

** Students may take one of four high school STE tests offered in biology, chemistry, introductory physics, and technology/engineering in grade 9 or grade 10. Additional information about these tests is available in a separate document.*

1.3.1 About the MCAS Assessments

In 2019, computer-based administration was required for all content areas at grades 3–8 and for grade 10 ELA and mathematics, but paper-based tests were available as a test accommodation at all grades. Because of the transition from legacy MCAS to next-generation MCAS tests, the presentation of psychometric results for the 2019 next-generation MCAS tests in grade 10 ELA and mathematics and grades 5 and 8 STE does not include comparisons to prior administrations.

1.3.2 About the 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 include the following elements:

- high-quality test items aligned to the Massachusetts learning standards;
- item types that assess both skills and knowledge, such as writing to text in ELA and solving complex problems in mathematics;
- achievement levels that send clear signals to students, parents, and educators about readiness for work at the next level (including results at grade 10 that signal readiness for college and career);
- a full range of student accessibility features and accommodations; and
- both computer-based and paper-based test administrations, with computer-based testing as the primary method.

In 2019, all students in grades 3–8 and 10 took the next-generation assessments in ELA and mathematics and students in grades 5 and 8 took the next-generation assessments in STE. In addition, a standalone next-generation field test was administered in biology and introductory physics to grade 9 students enrolled in the corresponding courses. These field-test items will be used to populate the June 2020 next-generation High School Biology and Introductory Physics tests.

1.3.3 Background on the Transition to Next-Generation Assessments

The following are some key milestones for developing and implementing the next-generation MCAS tests:

- **2010**: 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. Approximately one-half of the students at those grade levels take the MCAS assessments, and about one-half take the PARCC assessments.
- **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 in ELA and mathematics grades 3–8 for the first time. The tests include a mixture of MCAS and PARCC items.
- **Spring 2018**: The second administration of next-generation MCAS tests in ELA and mathematics grades 3–8. PARCC items are used only for a small number of items on the mathematics tests.
- **Spring 2019**: The third administration of next-generation MCAS tests in ELA and mathematics grades 3–8. The first administration in ELA and mathematics grade 10 and STE grades 5 and 8. The tests include only MCAS items and PARCC items are no longer included.

1.4 Special Issues

1.4.1 Grade 10 ELA Essay Prompt

DESE chose not to score students' responses to one of the grade 10 ELA essay prompts, a question that asked students to respond to a passage from Colson Whitehead's novel *The Underground Railroad*. Students were asked to write from the perspective of a character who, while sheltering a girl who has run away from slavery, uses derogatory language toward her. Researchers from the Stanford Graduate School of Education were invited by Commissioner Jeffrey Riley to examine whether there was any statistical evidence that the performance of African American students was affected on the parts of the test following the essay prompt. The essay prompt was administered on the second day of testing. The questions that followed the essay prompt included a small number of common operational items, as well as other questions that were for field testing only and did not count toward student scores. The researchers estimated students' scores for the portion of the test after the essay prompt based on students' performance on the first part of the test and compared this to the actual performance. They found a statistically significant but small negative difference in performance on questions after the essay prompt for black students as compared to the performance of white students (a –.06 σ points difference on questions after the essay prompt, or an approximate difference of –.0061 σ points on the entire test). Researchers noted that this small difference was within the typical range of differences for these two groups on the later portions of the tests from other grades and years. Researchers emphasized the small impact by indicating that about three students could have been misclassified as "Failing" the test. To ensure fairness to all students, the state allowed students who believed they were negatively impacted by the question to take the November ELA Retest and use the higher of the two scores for the graduation requirement and/or John and Abigail Adams Scholarship eligibility. Additional information on the results of the Stanford Study is available at www.doe.mass.edu/news/news.aspx?id=25710; the study is available at cepa.stanford.edu/sites/default/files/mcas_study_20190831.pdf.

1.4.2 Paper-Based Testing in One School District

Due to a computer virus and an extended lack of internet operability in 2019, the Lynn School District tested using paper-based testing materials only. An unusually large performance spike was seen for Lynn, using non-adjusted paper scores. As in past years when larger numbers of districts used paper-based forms, a mode adjustment study was conducted to correct for mode effects using matched samples and propensity scores (matching online to paper results for two comparable groups of students on each test). This process was repeated ten times (i.e., ten matched samples were generated for each test) to obtain comparable sets of matched samples (in which the standardized differences between groups for each variable was controlled to be <0.1). Lookup tables identifying

equivalent online test scores for paper tests were generated using an equipercentile linking approach. Finally, a kernel smoothing was applied to get smoothed adjusted score linkages. A full description of this analysis is provided in section 3.6.4 of this report.

1.4.3 Computer-Based Testing Drag-and-Drop Item Issue

A two-point drag-and-drop item in Grade 10 ELA exhibited a minor error when taken on smaller screen devices (such as iPads). The error made it difficult for a small percentage of students to drag a portion of the answer into the correct answer box; however, students could successfully answer the item using a workaround. Schools that reported the error were provided with the workaround. To evaluate the impact on schools using the smaller screen devices, several analyses were conducted including DIF, multi-group IRT analyses, and logistic regression analyses. Results showed a differential negative impact on student results in one district that did not receive the work-around method for this item. Consequently, this item was not counted towards students' scores in that district (Methuen), which required the use of a separate lookup table for generating student scores. On parent/guardian reports, and in some Edwin Analytics reports, aggregate results were not reported for that district.[1]

[1] See the Parent/Guardian report example in Appendix Q. The aggregated district results in ELA are labeled "Average Points in District."

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, DESE 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 1.3, Massachusetts is in the process of transitioning from the legacy MCAS tests to next-generation MCAS assessments that

- align MCAS items with the 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 a student's 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 2019 next-generation MCAS tests were aligned to the 2017 Massachusetts curriculum frameworks for English Language Arts and Mathematics and the 2016 Massachusetts curriculum frameworks for Science and Technology/Engineering.

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 from the next-generation ELA and mathematics tests in grades 3–8 and 10 and the next-generation STE tests in grades 5 and 8 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, and
- 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 include 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. 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 Appendices E and F	Scoring consistency, interrater agreement, and scoring accuracy
	3.5 Appendices G and H	Classical item analyses
	3.7 Appendix O	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 Appendices A, M, S, and W	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 Appendices I, J, K, L, and T	Item response theory modeling; dimensionality; scaling; linking online to paper results; differential item functioning
Consequential validity	3.8 Appendices N, P, and Q	MCAS reporting
	3.9.5	Supporting the valid use of MCAS data

MCAS-Alt assessment results are sometimes aggregated with other MCAS results. Therefore, validity information with respect to reliability and content-related validity provided for MCAS also pertains, to some extent, to the MCAS-Alt. 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. Summary of Validity Evidence for MCAS-Alt

Type of Validity Evidence	*Section*	*Description of Information Provided*
Content-related validity evidence	4.2.1 Appendix C	Assessment design (test blueprints aligned to MCAS blueprints but with modifications made for the range and complexity of standards); descriptions of primary evidence and supporting documentation
Reliability and subgroup statistics and scoring consistency	4.4, 4.7.3, and 4.8 Appendices F, P, V, and W	Procedures to ensure consistent scoring; interrater scoring statistics
	4.5 Appendix G	Classical item statistics
	4.7.1 and 4.7.2 Appendix O	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. 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.

2.5.1 General Achievement-Level Descriptors

The general ALDs for the next-generation MCAS tests at grades 3–8 and 10 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.

2.5.2 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, and students are assessed in grades 5, 8, and 9/10 in STE. In 2017, MCAS began the transition to a "next-generation" test that is administered primarily through a computer-based platform.

The MCAS program is managed by DESE staff with assistance and support from the assessment contractor, Cognia, and its subcontractor, Pearson. The next-generation computer-based tests were administered through Pearson's TestNav application. Massachusetts educators play a key role in MCAS through service on a variety of committees related to the development of MCAS test items, the development of MCAS achievement-level descriptors, and the setting of performance standards. The program is supported by a five-member national Technical Advisory Committee (TAC).

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

3.2 Next-Generation Test Design and Development

In 2019, the MCAS next-generation operational tests were administered at grades 3–8 and 10 in both ELA and mathematics and grades 5 and 8 in STE. In 2019, the next-generation tests in ELA, mathematics, and STE were administered primarily on a computer with paper accommodations available. (Legacy tests were administered on paper. Additional information about Legacy tests can be found in the *2019 Legacy MCAS Technical Report*.)

3.2.1 Test Specifications

3.2.1.1 CRITERION-REFERENCED TEST

In 2019, the items used on the next-generation MCAS tests were developed specifically for Massachusetts. All items were aligned to content standards in the Massachusetts Curriculum Frameworks. These content standards are the basis for the reporting categories in each content area and are used to guide the development of test items. Items on the 2019 next-generation MCAS tests were coded to the 2017 Massachusetts Curriculum Frameworks in ELA and Mathematics and the 2016 Massachusetts Curriculum Framework for Science and Technology/Engineering. All items were coded to at least one content standard and some were coded to more than one standard. In the next-generation STE tests, items were also coded to a science practice, if applicable. See section 3.2.4.1 for more information about science practices.

3.2.1.2 ITEM TYPES

The types of items and their functions, by content area, are described below.

English Language Arts (ELA)

- **Selected-response items (SR)** are worth one or two points and consist of the following:
  - **Multiple-choice items** (computer and 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, multiple-choice item requires students to select the single best answer from four response options. Items are machine-scored: students earn 1 point for a correct response and receive 0 points for an incorrect or blank response.
  - **Two-part, multiple-choice items** (computer and 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. (In some cases, item directions instruct students to select two correct answers in the second part.) The items are machine-scored: correct responses are worth 2 points, partially correct answers are worth 1 point, and incorrect and blank responses receive 0 points. Students who answer the first part incorrectly receive a score of 0; students must answer the first part correctly in order to receive 1 or 2 points.
  - **Two-point, technology-enhanced items** (computer only) use computer-based interactions such as such as inline choice, hot spots, and drag and drop that require the student to choose from a range of options presented. The items are machine-scored: correct responses are worth 2 points, partially correct answers are worth 1 point, and incorrect and blank responses receive 0 points.
- **Constructed-response (CR) items** (computer and 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 and receive a score of 3, 2, 1, or 0 points.
- **Essays (ES)** (computer and paper) are on all tests in grades 3–8 and 10 and are text-based. Students are required to type or write an essay in response to a prompt which is based on the passage or passage set they have read. Essays are hand-scored and receive a score of 0–7 possible score points for grades 3–5 and 0–8 possible score points for grades 6–8 and 10.

See section 3.4 for more details on the scoring of CR and ES items.

Mathematics

- **Selected-response (SR) items** (computer and paper) are worth one or two points and consist of the following:
  - **Multiple-choice items** make efficient use of limited testing time and allow for coverage of a wide range of knowledge and skills within a content area. The items require students to select the single best answer from four response options. Items are machine-scored: students earn 1 point for a correct response and receive 0 points for an incorrect or blank response.
  - **Multiple-select items** require students to select two or more correct answers from a set of answer options. Students are instructed to select a certain number of options. There are typically five to seven options to choose from. Items are machine-scored: students earn 1 point for a correct response and receive 0 points for an incorrect or blank response.
  - **Technology-enhanced (TE) items** (computer only) use interactions such as inline choice, hot spots, and drag and drop that require the student to choose from a range of options presented. These TE items are machine-scored. For one-point TE items, correct responses are worth 1 point,

and incorrect and blank responses receive 0 points. Two-point TE items are assessed in grades 4–8 and 10. For two-point TE items, there are two parts, and each part is worth 1 point. The two parts are scored independently from each other. Students earn 2 points for 2 correct parts, 1 point for only 1 correct part, and receive 0 points for no correct parts.

- **Short-answer (SA) items** (computer and paper) are worth one or two points and consist of the following:
  - **Short-answer items** 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: students earn 1 point for a correct response and receive 0 points for an incorrect or blank response. For the paper versions of these items, students write their numbers in boxes and then complete a number grid, which is machine-scored.
  - **Technology-enhanced (TE) items** (computer only) use interactions such as fraction model or line plot that require the students to demonstrate knowledge and skills by generating an answer or selecting an answer from a wide range of options. These TE items are machine-scored. For one-point TE items, students earn 1 point for a correct response and receive 0 points for an incorrect or blank response. Two-point TE items are assessed in grades 4–8 and 10. For two-point TE items, there are two parts, and each part is worth 1 point. The two parts are scored independently from each other. Students earn points for 2 correct parts, 1 point for only 1 correct part, and receive 0 points for no correct parts.
- **Constructed-response (CR) items** (computer and paper) require students to solve problems and generate responses to prompts. Students are required to use higher-order thinking skills, such as analyzing and explaining, to construct responses. Some CR items include a technology-enhanced part, such as creating a graph or completing a model using drag and drop technology. Student responses are hand-scored. CR items are worth either three or four points.
  - ***Three-point constructed-response items*** are used only on the grade 3 test. Students are expected to solve problems and generate one to two sentences in response to a prompt. Students earn 3, 2, 1, or 0 score points for these items.
  - ***Four-point constructed-response items*** are used on the grades 4–8 and 10 tests. Student responses are hand scored and assigned score points ranging from zero to four, depending on the item type. Students earn 4, 3, 2, 1, or 0 score points for these items.

Science and Technology/Engineering (STE)—Grades 5 and 8

- **Selected-response (SR) items** (computer and paper) are worth one or two points and consist of the following:
  - **Multiple-choice items** make efficient use of limited testing time and allow for coverage of a wide range of knowledge and skills within a content area. The items require students to select the single best answer from four response options. Items are machine-scored: students earn 1 point for a correct response and receive 0 points for an incorrect or blank response.
  - **Multiple-select items** require students to select two or more correct answers from a set of answer options. Students are instructed to select a certain number of options. There are typically four to seven options to choose from. Items are machine-scored: students earn 1 point for a correct response and receive 0 points for an incorrect or blank response.

- **Technology-enhanced (TE) items** (computer only) use interactions such as inline choice, hot spots, and drag and drop that require the student to choose from a range of options presented. These TE items are machine-scored. For one-point TE items, students earn 1 point for a correct response and receive 0 points for an incorrect or blank response. For two-point TE items, there are two parts, and each part is worth one point. The two parts are scored independently from each other. Students earn 2 points for 2 correct parts, 1 point for only 1 correct part, and receive 0 points for no correct parts.

- **Constructed-response (CR) items** (computer and paper) require students to process information about a scenario and to use higher-order thinking skills, such as analyzing and explaining, to construct responses to prompts (e.g., identify, describe, explain) about the scenario. The scenario information may include narrative descriptions, models, and data tables or graphs. Some CR items include a technology-enhanced part, such as completing a model using drag and drop technology. Student responses are hand-scored and each item is worth either 2 or 3 score points. For two-point CR items, students may earn 2, 1, or 0 score points. For three-point CR items, students may earn 3, 2, 1, or 0 score points.

3.2.1.3 DESCRIPTION OF TEST DESIGNS

The MCAS assessments contain both common and matrix items. The common items are administered to all students and count toward a student's overall score. Matrix items are either field-test items or equating items. Field-test items are tried out to see how they perform and do not count toward a student's score. Equating items are used to link one year's results to those of previous years and do not count toward a student's score. Equating and field-test items are distributed among multiple forms of the test for each grade and content area.

The number of test forms varies by grade and content area and typically ranges between 10 and 30 forms. Each student takes one form of the test and therefore answers a subset of matrix items. Common and matrix items are not distinguishable to test takers. Because all students are given matrix items, an adequate sample size (typically a minimum of 1,500 responses per item) is obtained to produce data that can be used to inform equating decisions and common item selection for future tests.

In 2019, two common forms were developed for the grades 3–8 and 10 ELA and mathematics assessments and the grades 5 and 8 STE assessments: one form designated as the computer-based (CBT) common form and one form designated as the paper-based (PBT) common form. To create the PBT common form, technology-enhanced items that appeared on the CBT form were revised and made into "paper cousins." Paper cousins are items that test the same content as the technology-enhanced items on the CBT, but they have been changed into multiple-choice or multiple-select items.

3.2.2 ELA Test Specifications

3.2.2.1 STANDARDS

The 2019 MCAS grades 3–8 and 10 ELA tests, including all matrix items, were aligned to and measured the following learning standards from the *2017 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)
  - Production and Distribution of Writing (Standards 4–6)

The *2017 Massachusetts Curriculum Framework for English Language Arts and Literacy* can be found at www.doe.mass.edu/frameworks/ela/2017-06.pdf.

3.2.2.2 ELA ITEM TYPES

The grades 3–8 and 10 ELA tests used several item types, as shown in Table 3-1.

Table 3-1. ELA Item Types and Score Points

Item Type	*Possible Raw Score Points*	*Grade Levels*
Multiple-choice (SR)	0 or 1	3–8, 10
Two-part, multiple-choice (SR)	0, 1, or 2	3–8, 10
Technology-enhanced (SR)	0, 1, or 2	3–8, 10
Constructed-response (CR)	0, 1, 2, or 3	3–4
Essay (ES)	0 to 7	3–5
	0 to 8	6-8, 10

SR = selected-response, CR = constructed-response, ES = essay

3.2.2.3 PASSAGE TYPES

Passages used in the ELA tests are authentic published passages selected for the MCAS assessment. Test developers, including DESE test developers, 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 a sufficient number of unique assessment items; and
- be free of bias and sensitivity issues.

Passages ranged in length from approximately 600 to 2500 words per passage set. Word counts were slightly reduced at lower grades. Passage sets consisted of either a single passage or paired/tripled 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. Informational passages are drawn from a variety of sources, including magazines, newspapers, and books.

In grades 3–8, each common form included three passage sets, with some forms containing two literary passage sets and one informational passage set, while other forms contained one literary passage set and two informational passage sets. In grade 10, each common form included four passage sets with three literary and one informational set. Across the forms, sets may be single, paired, or tripled selections.

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 and 10 is their degree of complexity, which results from increasing levels of sophistication in language and concepts, as well as passage length. Test developers use 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).

3.2.2.4 ELA TEST DESIGN

All items are coded to ELA framework standards. There are no stand-alone items on the tests; all vocabulary, grammar, and mechanics questions are associated with a passage set.

Students read a passage set and answer questions that follow. Question types include selected-response items, constructed-response items (grades 3 and 4 only), and essay items. Please see section 3.2.1.2 above for additional details on item types. Approximately 20% of the items were technology-enhanced items.

Test Design by Grade

Grades 3–4

The common portion of each test at grades 3 and 4 included three passage sets, and the matrix portion included one passage set. Two of the common passage sets included seven or eight 1- or 2-point selected-response items plus one 7-point text-based essay item. The other common passage set included seven 1- or 2-point selected-response items and one 3-point constructed-response item. Each test contained a total of 44 common points distributed across two testing sessions.

Grade 5

The common portion of each test at grade 5 included three passage sets, and the matrix portion included one passage set. Each passage set included seven or eight 1- or 2-point selected-response items, and one 7-point text-based essay item. The test contained a total of 48 common points distributed across two testing sessions.

Grades 6–8

The common portion of each test at grades 6–8 included three passage sets, and the matrix portion included one passage set. Each passage set included seven or eight 1- or 2-point selected-response items, and one 8-point text-based essay item. The test contained a total of 51 common points distributed across two testing sessions.

Grade 10

The common portion of each test at grade 10 included four passage sets. Two passage sets in the common portion included seven or eight 1- or 2-point selected-response items and one 8-point text-based essay item. Two common passage sets included four 1- or 2-point selected-response items. The test contained a total of 51 common points distributed across two testing sessions.

Matrix

For grades 3–8, the matrix portion contained included one passage set. In grades 3–4, the matrix passage set included eight or nine 1- or 2-point selected-response items, and either two constructed-response items or one

essay. In grades 5–8, the matrix passage set included eight or nine 1- or 2-point selected-response items, and one essay item.

The grade 10 matrix portion included two passage sets. One matrix passage set included eight 1- or 2-point selected-response items, and one 8-point text-based essay item. The other matrix passage set included four 1- or 2-point selected-response items.

Table 3-2 shows the recommended testing times. MCAS tests are untimed; therefore, times shown in the table are approximate.

Table 3-2. ELA Recommended Testing Times, Grades 3–8 and 10

Grade	*Session 1 recommended testing time (min)*	*Session 2 recommended testing time (min)*	*Total recommended testing time (min)*
3	120–150	120–150	240–300
4	120–150	120–150	240–300
5	120–150	120–150	240–300
6	120–150	120–150	240–300
7	120–150	120–150	240–300
8	120–150	120–150	240–300
10	150	150	300

Common and Matrix Item Distribution

The grades 3–8 and 10 ELA tests were administered to most students on the computer and to some students with accommodations on a paper form. Tables 3-3 (for the computer-based forms) and 3-4 (for the paper-based forms) list the distribution of common and matrix items in each 2019 ELA test, by grade.

Table 3-3. Distribution of ELA Common and Matrix Items by Grade and Item Type—Computer-based (CBT)

Grade and Test			*Items per Form*							
			Common				*Equating/Matrix*			
Grade	*Test*	*# of Forms*	*SR (1 pt.)*	*SR (2 pt.)*	*CR*	*ES*	*SR (1 pt.)*	*SR (2 pt.)*	*CR*[1]	*ES*
3	ELA	12	15	6	1	2	6-8	0-2	2	1
4	ELA	12	15	6	1	2	6-8	0-2	2	1
5	ELA	12	17	5	0	3	6-8	0-2	0	1
6	ELA	12	15	6	0	3	6-8	0-2	0	1
7	ELA	12	15	6	0	3	6-8	0-2	0	1
8	ELA	12	15	6	0	3	6-8	0-2	0	1
10[2]	ELA	20	21	7	0	2	9-10	2-3	0	1

[1] Each grade 3 and grade 4 matrix form contained either two constructed-response items or one essay item.
[2] The common item numbers for grade 10 reflect the removal of one 8-point essay from the common form.

Table 3-4. Distribution of ELA Common and Matrix Items by Grade and Item Type—Paper (PBT)[1]

Grade and Test			*Items per Form*							
Grade	*Test*	*# of Forms*	*Common*				*Equating*			
			SR (1 pt.)	*SR (2 pt.)*	*CR*	*ES*	*SR (1 pt.)*	*SR (2 pt.)*	*CR*	*ES*
3	ELA	1	15	6	1	2	8	0	2	0
4	ELA	1	15	6	1	2	8	0	2	0
5	ELA	1	17	5	0	3	8	0	0	1
6	ELA	1	15	6	0	3	8	0	0	1
7	ELA	1	15	6	0	3	8	0	0	1
8	ELA	1	15	6	0	3	8	0	0	1
10[2]	ELA	1	21	7	0	2	9	3	0	1

[1] *The paper form is derived from Form 1 of the CBT.*
[2] *The common item numbers for grade 10 reflect the removal of one 8-point essay from the common form.*

3.2.2.5 ELA BLUEPRINTS

Table 3-5 shows the target and actual percentages of common item points by reporting category. Reporting categories are based on the anchor standards in the 2017 Massachusetts curriculum framework for ELA.

Table 3-5. 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*	*10[1]*
Language	25 (21)	25 (25)	30(27)	25 (24)	25 (24)	25 (24)	25 (18)
Reading	55 (61)	55 (57)	45 (48)	45 (47)	45 (47)	45 (47)	55 (63)
Writing	20 (18)	20 (18)	25(25)	30 (29)	30 (29)	30 (29)	20 (20)
Total	100	100	100	100	100	100	100

[1] *The reporting category percentage in grade 10 reflects the removal of one 8-point essay from the common form.*

3.2.2.6 ELA COGNITIVE LEVELS

Each item on the ELA tests 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 tests are described below.

- **Level I (Identify/Recall)**—Level I items require that the student recognize basic information presented in the text. Examples of skills at this level include identifying main ideas/facts/details; recalling and locating details; identifying genre or setting; and identifying definitions, parts of speech, or functions of punctuation. Key words include identify, list, match, recognize, describe, and distinguish.
- **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. Examples of skills at this level include understanding the whole text (Big Picture)/generalizing; interpreting, making connections, visualizing, and forming questions; explaining a character's role/motives; determining whether an idea is fact or opinion; filtering important information and key concepts; and determining the meaning of a word in context. Key words include infer, analyze, describe, interpret, determine, conclude, explain, summarize, and classify.
- **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. Examples of skills at this level include understanding another point of view; analyzing/evaluating an author's purpose, style, and message; arguing/defending a point of view with evidence from the text; using reasoning to determine

an outcome; applying information from the text; and synthesizing elements of text(s) in order to create a whole. Key words include critique, evaluate, analyze, predict, agree/disagree, argue/defend, apply, synthesize, judge, compare, and contrast.

Each cognitive level is represented in the ELA tests.

3.2.2.7 ELA 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 MATHEMATICS STANDARDS

The 2019 MCAS grades 3–8 and 10 mathematics tests, including all field-test items, were aligned to, and measured the learning standards from the *2017 Massachusetts Curriculum Framework for Mathematics*. The 2017 standards are grouped by domains at grades 3–8 and 10, as shown below.

- ***Domains for grades 3–5***
  - Operations and Algebraic Thinking
  - Number and Operations in Base Ten
  - Number and Operations—Fractions
  - Geometry
  - Measurement and Data
- ***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
- ***Domains for grade 10***
  - Number and Quantity
  - Algebra
  - Functions

- Geometry
- Statistics and Probability

The *2017 Massachusetts Curriculum Framework for Mathematics* can be found at www.doe.mass.edu/frameworks/math/2017-06.pdf.

3.2.3.2 MATHEMATICS ITEM TYPES

The 2019 mathematics tests included several item types, as shown in Table 3-6. Approximately 25–30% of the items were technology-enhanced items.

Table 3-6. Mathematics Item Types and Score Points

Item Type	*Possible Raw Score Points*	*Grade Levels*
Multiple-choice (SR)	0 or 1	3–8, 10
Multiple-select (SR)	0 or 1	3–8, 10
Technology-enhanced (SA)/(SR)/(CR)	0 or 1 0, 1, or 2	3 4–8, 10
Short-answer (SA)	0 or 1	3–8, 10
Constructed-response (CR)	0, 1, 2, or 3 0, 1, 2, 3, or 4	3 4–8, 10

SA = short-answer, SR = selected-response, CR = constructed-response

3.2.3.3 MATHEMATICS TEST DESIGN

Test Design by Grade

Grade 3

The common portion of the grade 3 test included thirty-six 1-point selected-response or short- answer items and four 3-point constructed-response items. The matrix portion included three 1-point selected-response or short-answer items and one 3-point constructed-response item. The test contained a total of 48 common points distributed across two testing sessions.

Grades 4–6

The common portion of the grades 4–6 tests included thirty-four 1-point selected-response or short-answer items, two 2-point selected-response items, and four 4-point constructed-response items. The matrix portion included two 1-point selected-response or short-answer items, one 2-point selected-response or short-answer item, and one 4-point constructed-response item. Each test contained a total of 54 common points distributed across two testing sessions.

Grades 7–8

The common portion of the grades 7–8 tests included thirty-four 1-point selected-response or short-answer items, two 2-point selected-response items, and four 4-point constructed-response items. The matrix portion included two 1-point selected-response or short-answer items, two 2-point selected-response or short-answer items, and two 4-point constructed-response items. Each test contained a total of 54 common points distributed across two testing sessions. Items in session 2 were developed to assess content where the students may need a calculator. These items were 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).

Grade 10

The common portion of the grade 10 test included thirty-two 1-point selected-response or short-answer items, six 2-point selected-response items, and four 4-point constructed-response items. The matrix portion included four 1-

point selected-response or short-answer items, two 2-point selected-response or short-answer items, and four 4-point constructed-response items. Each test contained a total of 60 common points distributed across two testing sessions. Items in session 2 were developed to assess content where the students may need a calculator. These items were 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).

Table 3-7 shows the distribution of common and matrix points on the 2019 mathematics tests, as well as recommended testing times. Since MCAS tests are untimed, the times shown are approximate.

Table 3-7. Mathematics Recommended Testing Times and Common/Matrix Points per Test, Grades 3–8 and 10

Grade	*# of Sessions*	*Session 1 Recommended Testing Time (in minutes)*	*Session 2 Recommended Testing Time (in minutes)*	*Total Recommended Testing Time (in minutes)*	*Common Points*	*Matrix Points*
3	2	90	90	180	48	6
4–6	2	90	90	180	54	8–9
7–8	2	90	90	180	54	12–14
10	2	90 – 120	90 – 120	180 – 240	60	24

The grades 3–8 and 10 mathematics tests were administered to most students on the computer and to some students with accommodations on a paper form. Tables 3-8 (for the computer-based forms) and 3-9 (for the paper form) show the distribution of common and matrix item types.

Table 3-8. Distribution of Mathematics Common and Matrix Items by Grade and Item Type—Computer-based (CBT)

Grade	*# of Forms*	Common				Matrix	
		SR/MS SA TE		*CR*		*SR/MS SA/ TE*	*CR*
		(1 pt.)	*(2 pt.)*	*(3 pt.)*	*(4 pt.)*	*(1 or 2 pt.)*	*(3 or 4 pt.)*
3	28	36	0	4	0	3	1
4	28	34	2	0	4	3	1
5	28	34	2	0	4	3	1
6	28	34	2	0	4	3	1
7	21	34	2	0	4	4	2
8	22	34	2	0	4	4	2
10	28	32	6	0	4	6	4

Table 3-9. Distribution of Mathematics Common and Matrix Items by Grade and Item Type—Paper (PBT)

Grade	*# of Forms*	Common				Matrix	
		SR/MS SA		*CR*		*SR/MS/SA*	*CR*
		(1 pt.)	*(2 pt.)*	*(3 pt.)*	*(4 pt.)*	*(1 or 2 pt.)*	*(3 or 4 pt.)*
3	1	36	0	4	0	3	1
4	1	34	2	0	4	3	1
5	1	34	2	0	4	3	1
6	1	34	2	0	4	3	1
7	1	34	2	0	4	4	2
8	1	34	2	0	4	4	2
10	1	32	6	0	4	6	4

3.2.3.4 MATHEMATICS BLUEPRINTS

Tables 3-10 through 3-13 show the target and actual percentages of common item points by reporting category. Reporting categories are based on the Massachusetts curriculum framework domains.

Table 3-10. 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	30 (31)	20 (20)	15 (15)
Number and Operations in Base Ten	15 (17)	20 (20)	30 (30)
Number and Operations – Fractions	20 (19)	30 (30)	25 (24)
Geometry	10 (8)	10 (11)	10 (11)
Measurement and Data	25 (25)	20 (19)	20 (20)
Total	100	100	100

Table 3-11. 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	20 (20)	20 (20)
The Number System	20 (20)	20 (19)
Expressions and Equations	30 (30)	25 (26)
Geometry	15 (15)	15 (15)
Statistics and Probability	15 (15)	20 (20)
Total	100	100

Table 3-12. Target (and Actual) Distribution of Math Common Item Points by Reporting Category, Grade 8

Domain	*% of Points at Each Grade (+/-5%)*
The Number System and Expressions and Equations	40 (39)
Functions	20 (20)
Geometry	30 (30)
Statistics and Probability	10 (11)
Total	100

Table 3-13. Target (and Actual) Distribution of Math Common Item Points by Reporting Category, Grade 10

Domain	*% of Points at Each Grade (+/-5%)*
Number and Quantity	15 (15)
Algebra & Functions	35 (35)
Geometry	35 (35)
Statistics and Probability	15 (15)
Total	100

3.2.3.5 MATHEMATICS 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 described below.

- **Level I (Recall and Recognition)**—Level I items require that the student recall mathematical definitions, notations, simple concepts, and procedures, and applies 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 that the student engages 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 that the student performs more abstract reasoning, planning, and evidence-gathering. In order to answer questions of this cognitive level, a student must engage in reasoning about an open-ended situation with multiple decision points, 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 and across item types. Cognitive Level III is best represented by constructed-response items; an attempt was made to include cognitive Level III items at each grade.

3.2.3.6 MATHEMATICS REFERENCE MATERIALS

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

Reference sheets were provided to students at grades 5–8 and 10. These sheets contain information, such as formulas, that students may need to answer certain items.

The second session of the grades 7, 8, and 10 mathematics tests was a calculator session. All items included in this session were either calculator-neutral (calculators are permitted but not required to answer the question) or calculator-active (students are expected to use a calculator to answer the question). Each student taking the computer-based grade 7 mathematics test had access to a five-function calculator during session 2 of the mathematics test. Each student taking the computer-based grade 8 and grade 10 mathematics tests had access to a scientific calculator during session 2 of the mathematics test. Students taking the paper-based mathematics tests in grades 7, 8, and 10 had access to comparable handheld calculators.

3.2.4 Science and Technology/Engineering Test (STE) Specifications

3.2.4.1 STE STANDARDS AND PRACTICES

The next-generation STE MCAS tests for grades 5 and 8 were first administered in 2019. The tests were aligned to the standards in the 2016 Massachusetts STE Curriculum Framework. The grade 5 test was based on the grades 3–5 standards and the grade 8 test was based on the grades 6–8 standards. The 2016 PreK–8 standards are grouped into the following four domains:

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

In addition, the grades 5 and 8 next-generation STE MCAS tests assessed the science and technology/engineering practices incorporated into the standards. There are eight practices included in the standards:

1. Asking questions (for science) and defining problems (for engineering)
2. Developing and using models
3. Planning and carrying out investigations
4. Analyzing and interpreting data
5. Using mathematics and computational thinking
6. Constructing explanations (for science) and designing solutions (for engineering)
7. Engaging in argument from evidence
8. Obtaining, evaluating, and communicating information

The *2016 Massachusetts Science and Technology/Engineering Curriculum Framework* can be found at www.doe.mass.edu/frameworks/scitech/2016-04.pdf. In addition, Instructional Guidelines were developed to help clarify some standards and can be found at www.doe.mass.edu/stem/ste/g3-g5.docx and www.doe.mass.edu/stem/ste/g6-g8.docx.

3.2.4.2 STE ITEM TYPES

The 2019 next-generation MCAS STE tests included several item types, as shown in Table 3-14.

Table 3-14. STE Item Types and Score Points

Item Type	*Possible Raw Score Points*	*Grade Level*
Multiple-choice (SR)	0 or 1	5 and 8
Multiple-select (SR)	0 or 1	5 and 8
Technology-enhanced (SR)	0 or 1 0, 1, or 2	5 and 8
Constructed-response (CR)	0, 1, or 2 0, 1, 2, or 3	5 and 8

SR = selected-response, CR = constructed-response

3.2.4.3 STE TEST DESIGN

Test Design

The common portion of the grades 5 and 8 tests included thirty-two 1-point selected-response items, three 2-point selected-response items, two 2-point constructed-response items, and four 3-point constructed-response items. The next-generation STE tests included two common modules which are scenario-based groups of items. Each module contained three 1-point selected-response items and one 3-point constructed response item. Module items made up 12 points of the test, while discrete items made up 42 points of the test. The matrix portion included five 1-point selected-response items, one 2-point selected-response or constructed-response item, and one 3-point constructed-response item, for a total of 10 points. Some forms contained matrix modules (equating or field test) while other forms only included discrete items. The test contained a total of 54 common points distributed across two testing sessions. Approximately 25–30% of the items were technology-enhanced items.

Table 3-15 shows the distribution of common and matrix points on the 2019 STE tests, as well as recommended testing times. Since MCAS tests are untimed, the times shown are approximate.

Table 3-15. STE Recommended Testing Times and Common/Matrix Points per Test, Grades 5 & 8

Grade	*# of Sessions*	*Session 1 Recommended Testing Time (in minutes)*	*Session 2 Recommended Testing Time (in minutes)*	*Total Recommended Testing Time (in minutes)*	*Common Points*	*Matrix Points*
5	2	75–90	75–90	150–180	54	10
8	2	60–75	60–75	120–150	54	10

The grades 5 and 8 STE tests were administered to most students on the computer and to some students with accommodations on a paper form. Tables 3-16 (for the computer-based forms) and 3-17 (for the paper form) show the distribution of common and matrix item types.

Table 3-16. Distribution of STE Common and Matrix Items by Grade and Item Type—Computer-based (CBT)

Grade	*# of Forms*	*Common*				*Matrix*		
		SR1 (1 pt.)	*SR2 (2 pt.)*	*CR2 (2 pt.)*	*CR3 (3 pt.)*	*SR1 (1 pt.)*	*SR2/CR2 (2 pt.)*	*CR3 (3 pt.)*
5	18	32	3	2	4	5	1	1
8	18	32	3	2	4	5	1	1

Table 3-17. Distribution of STE Common and Matrix Items by Grade and Item Type—Paper (PBT)

Grade	*# of Forms*	*Common*				*Matrix*	
		SR1	*SR2*	*CR2*	*CR3*	*SR1/SR2*	*CR2/CR3*
		(1 pt.)	*(2 pt.)*	*(2 pt.)*	*(3 pt.)*	*(1 or 2 pt.)*	*(2 or 3 pt.)*
5	1	32	3	2	4	5	2
8	1	32	3	2	4	6	1

3.2.4.4 STE BLUEPRINTS

Table 3-18 shows the target and actual percentages of common item points by content reporting category. Content reporting categories are based on the Massachusetts curriculum framework domains.

Table 3-18. Target (and Actual) Distribution of STE Common Item Points by Reporting Category, Grades 5 & 8

Domain	*% of Points at Each Grade (+/-5%)*	
	5	*8*
Earth and Space Sciences	25 (26)	25 (26)
Life Science	25 (26)	25 (26)
Physical Science	25 (26)	25 (26)
Technology/Engineering	25 (22)	25 (22)
Total	100	100

In addition to the content reporting categories, over 50% of the items were coded to an MCAS science and technology/engineering practice category. These items were dually coded, meaning they were coded to both a content reporting category and a practice reporting category. The MCAS practice reporting categories are listed in the Table 3-19.

Table 3-19. STE Practices Assessed on MCAS

MCAS Practice Category	*Science and Engineering Practices*
Investigations and Questioning	Asking Questions and Defining Problems Planning and Carrying Out Investigations
Mathematics and Data	Analyzing and Interpreting Data Using Mathematics and Computational Thinking
Evidence, Reasoning, and Modeling	Developing and Using Models Constructing Explanations and Designing Solutions Engaging in Argument from Evidence Obtaining, Evaluating, and Communicating Information

Regarding the STE practices, each content standard includes a reference to one STE practice. For example, standard 5-ESS2-1 states:

> *Use a model to describe the cycling of water through a watershed through evaporation, precipitation, absorption, surface runoff, and condensation.*

Although only a single practice is referenced within each standard, different practices may be assessed with the associated content. In the example above, items assessing standard 5-ESS2-1 may assess not only the "developing and using models" practice; they may also assess any other practice, such as constructing explanations or analyzing and interpreting data.

Each released item that assessed a practice was coded to one of the three practice categories listed in Table 3-19. However, when reporting results by reporting category, there was a general "STE Practices" reporting category. Results were not reported out on the three practice categories listed, due to the limited number of items.

3.2.4.5 STE COGNITIVE LEVELS

Each item on the STE tests is assigned a cognitive level according to the cognitive demand of the item. Cognitive levels are not synonymous with difficulty. The cognitive skill describes each item based on the complexity of the mental processing a student must use to answer the item correctly. Only one cognitive skill is designated for each item. STE uses a modified revised Bloom's taxonomy to code items by cognitive level. Items generally fall into either the understanding or applying/analyzing cognitive skill level. Table 3-20 describes the cognitive skills used for the STE test items.

Table 3-20. STE Cognitive Skill Descriptions

Cognitive Skill	*Description*
Remembering	Identify or define a fundamental concept with little or no context • Identify or recall fundamental concepts or definitions. *Does the item require recall of basic facts or definitions?*
Understanding	Identify, describe, or explain concepts using typical classroom examples • Using a model, explain how people on Earth experience day and night. • Describe the role of weathering and erosion in the production and movement of soil. • Identify processes illustrated in common science models such as the water cycle and particle models of matter. • Complete a life cycle with the stages birth, growth, reproduction, and death. • Distinguish between common inherited characteristics and common characteristics that are a result of the environment. • Describe how magnets will behave in familiar set-ups. • Identify characteristic properties that can be used to classify a substance. *Does the item require the recognition or a description of a familiar concept?*
Applying / Analyzing	Describe, explain, or apply scientific concepts to a novel situation, or Critically analyze data, graphs, and models of scientific phenomena • Use climate data to describe or predict the expected weather for a particular region. • Draw conclusions by interpreting data tables, graphs, or models, such as maps of plate boundaries, food webs, or steps of the communication process. • Compare different composter designs and describe benefits and drawbacks of their design features. • Given the results, determine whether combining novel substances results in a chemical reaction or a mixture. • Given a novel situation, explain how energy can be transferred from place to place. • Analyze investigations and predict outcomes. Use evidence from an investigation to support a claim and provide reasoning. • Describe or explain a scientific concept by applying a model to novel situations (e.g., use fossil data in rock layers to describe how the area has changed over time). • Determine a testable question that can be asked based on given information. *Does the item require drawing conclusions based on novel information?* *Does the item require critical analysis of information to make conclusions?*
Evaluating/ Creating	Generate an explanation or conclusion that involves the synthesis of multiple scientific concepts or processes • Construct models, graphs, charts, drawings, or diagrams and generate explanations or conclusions based on the information • Propose solution(s) to a scientific or engineering problem based on given criteria and constraints and generate an explanation for the solution(s) *Does the item require the synthesis of different concepts or skills to generate a solution?*

3.2.4.6 STE REFERENCE MATERIALS

Rulers were provided to students in grades 5 and 8. Hand-held rulers were provided to students taking the paper version of the STE test. Students taking the computer-based STE tests had access to two separate computer-based rulers: a centimeter ruler and a 1/8-inch ruler; students were not permitted to use hand-held rulers on the computer-based tests.

Students were provided a computer-based five-function calculator in grade 5 and a computer-based scientific calculator in grade 8. Hand-held calculators were given to students taking the paper-based tests.

3.2.5 Item and Test Development Process

Table 3-21 provides a detailed view of the item and test development process, in chronological order.

Table 3-21. Overview of Item and Test Development Process

Development Step	*Detail of the Process*
Select reading passages (for ELA only)	Contractor's test developers find potential passages and present them to DESE for initial approval; DESE-approved passages go to Assessment Development Committees (ADCs) composed of experienced educators, and then to a Bias and Sensitivity Committee (BSC) for review and recommendations. ELA items are not developed until passages have been reviewed by an ADC and a BSC. With the ADC and BSC recommendations, DESE makes the final determination as to which passages will be developed and used on a future MCAS test.
Develop items	Contractor's test developers generate items and edit items from subcontractors that are aligned to Massachusetts standards and specifications.
DESE and educator review of items	1. Contractor sends draft items to DESE test developers for review. 2. DESE test developers review and edit items prior to presenting the items to ADCs. 3. ADCs review items and make recommendations. 4. BSC reviews items and makes recommendations. 5. DESE test developers edit & revise items based on recommendations from ADC & BSC.
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. Comments and suggested edits are provided to DESE staff for review.
Benchmark constructed-response items and essays	DESE and contractor test developers meet to determine appropriate benchmark papers for training of scorers of field-tested constructed-response items and essays. Scoring rubrics and notes are reviewed and edited during benchmarking meetings. During the scoring of field-tested items, the contractor contacts DESE test developers with any unforeseen issues.
Item statistics meeting	ADCs review field-test statistics and recommend items for the common-eligible status, for re-field-testing (with edits, for math and discrete STE items, since ELA is passage-based), or for rejection. BSC also reviews items and recommends items to become common-eligible or to be rejected.
Test construction	Before test construction, DESE provides target performance-level cut scores to contractor's test 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. Each common set of items is delivered with proposed cut scores, including test characteristic curves (TCCs) and test information functions (TIFs). DESE test developers and editorial staff review and edit proposed sets of items. Contractor and DESE test developers and editorial staff meet to review edits and changes to tests. Psychometricians are available to provide statistical information for changes to the common form.

continued

Development Step	*Detail of the Process*
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	Approximately 50% of common items in grades 3–8 and 100% of common items in grade 10 are released to the public, and the remaining items are returned to the common-eligible pools to be used on future MCAS tests. An item description (a statement specifying the content of the item) is released for each common item (both released and non-released).

3.2.5.1 ITEM DEVELOPMENT AND REVIEW

Initial DESE Item Review

As described in the table above, all passages, items, and scoring guides are reviewed by DESE test developers before presentation to the ADCs for review. Passage selection information can be found in section 3.2.2.3. The DESE test developers evaluate new items for the following characteristics:

- **Alignment**: Are the items aligned to the standards?
- **Content**: Is the content accurate? Does the item elicit a response that shows a depth of understanding of the subject?
- **Contexts**: Are contexts grade-level appropriate? Are they realistic? Are they interesting to students?
- **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? What are the distractor rationales?
- **Mechanics**: How well are the items written? Are they grammatically correct? Do they follow the conventions of item writing? Is the wording grade-level appropriate and accessible for all students?
- **Technology**: Are the items scoring correctly? Is the item making the best use of the technology? Is there another type of item that is more appropriate?

After DESE's initial review, DESE and the contractor's test developers discuss and revise the proposed item sets in preparation for ADC review.

Assessment Development Committee (ADC) and Bias & Sensitivity Committee (BSC) Reviews

ADCs and the BSCs are each composed of approximately 10–12 Massachusetts educators from across the state (see Appendix B for lists of names). Each ADC and Bias meeting is co-facilitated by DESE and Cognia's test developers. There is an ADC for each content area and grade (e.g., ELA grade 3), and there are two BSCs—one for grades 3–7 and one for grades 8 and 10. All ADC and BSC recommendations remain with each item. ADC and BSC members meet several times a year to review new passages and items, and to review data from field test items. Members review items using Pearson's online platform ABBI. Each participant enters his or her "vote" and recommendations, and the facilitators record the consensus of the committee. The Department takes the recommendations of the ADCs and the BSCs into consideration and makes the final decision to approve items to become field-test eligible.

ADC Passage Review (ELA Only)

ELA ADCs review passages before any corresponding items are written. Committee members consider all the elements noted in section 3.2.2.3. If a passage is well known or if the passage comes from a book that is widely taught, then the passage is likely to provide an unfair advantage to those students who are familiar with the work. Committee members vote to accept or reject each passage, and the facilitators record the consensus of the group.

For each passage recommended for acceptance, committee members provide suggestions for item development. They also provide recommendations for the presentation of the passage, including suggestions for the purpose-setting statement, words to be footnoted or redacted, and graphics, illustrations, or photographs to be included with the text.

ADC Item Review

Once DESE test developers have reviewed and edited new items and scoring guides, the items are reviewed by the ADCs. Committees review items for the characteristics noted above. Members vote to accept, accept with edits (members may include suggested edits), or reject each item. The meeting facilitators record the consensus/majority opinion of the group.

Bias and Sensitivity Committee Passage and Item Review

After passages and items have been approved by the ADCs, they are also reviewed by a separate Bias and Sensitivity Review Committee (BSC). The role of the committee is to identify whether a passage or item contains material that is likely to significantly favor or disadvantage one group of students for reasons that are not educationally relevant. The purpose of the committee's review is to ensure that the ability to answer an item correctly reflects a student's learning, not cultural opportunities or life experiences. Specifically, a passage or item should be flagged by the committee if it is insensitive or disrespectful to a student's ethnic, religious, or cultural background (including disability, socio-economic status, and regional differences). The BSC votes to accept, accept with edits (including suggested edits), or reject (including their reasoning) each passage or item. The meeting facilitators record the consensus of the group.

External Content Expert Item Review

When items are selected to be included on the field-test portion of the MCAS, they are submitted to expert reviewers for their feedback. The task of the expert reviewer is to consider the accuracy of the content of items. Each item is reviewed by two independent expert reviewers. All expert reviewers for MCAS hold a doctoral degree (either in the content they are reviewing or in the field of education) and are affiliated with institutions of higher education in either teaching or research positions. Each expert reviewer has been approved by the DESE. The External Content Experts recommend either accepting or rejecting the item, including their reasoning. Expert reviewers' comments remain with each item.

Editing of Recommended Items

DESE test developers review the recommendations of the ADC, BSC, and expert reviewers and determine whether to revise an item based on the suggested edits. The items are also reviewed and edited by DESE and Cognia editors to ensure adherence to style guidelines in *The Chicago Manual of Style*, *American Heritage Dictionary*, MCAS 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 descriptions of what is required for a student to attain a maximum score;
- be written at a reading level that allows students to demonstrate their knowledge of the subject matter being tested.

3.2.5.2 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 typically answered by a minimum of 1,500 students, resulting in enough responses to yield reliable performance data.

Scoring of Field-Tested Items

All field-tested items, except for constructed-response items and essays, are machine-scored. These items include multiple-choice, multiple-select, short-answer, and technology-enhanced items.

All field-tested constructed-response items and essays are hand-scored. To train scorers, DESE 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 or essay. As with machine-scored items, 1,500 student responses are sufficient to provide reliable results. See section 3.4 for additional information on scorers and scoring.

Data Review of Field-Tested Items

Data Review by DESE

DESE test developers review all item statistics prior to making them available for review by the ADCs and BSCs. An item displaying statistics that indicate it did not perform as expected is closely reviewed and if it is found to be flawed it is rejected from the pool of items. After ADC and BSC reviews of item statistics, DESE test developers make final decisions regarding any recommendations.

Data Review by ADCs

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

- item difficulty (or mean score for polytomous items),
- item discrimination,
- Differential Item Functioning (DIF),
- distribution of scores across answer options and score points,
- distribution of answer options and score points across quartiles, and
- 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 and discrete STE items only, since ELA items are passage-based)
- reject

Data Review by BSCs

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

- female compared with male,
- African American/Black compared with white,
- Hispanic or Latino/a compared with white,
- English learners (EL) and former EL compared with non-EL

3.2.5.3 ITEM SELECTION FOR OPERATIONAL TEST

Cognia's test developers propose a set of previously field-tested or common, non-released 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 DESE. In preparation for meeting with the DESE test developers, the contractor's test developers consider the following criteria in selecting 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 and per reporting category for each content area. A broad coverage of standards and cognitive skills is expected. The previous year's common test should also be considered and items should not duplicated.
- **Item difficulty and complexity**. Item statistics drawn from the data analysis of items are used to ensure similar levels of difficulty and complexity from year to year as well as high-quality psychometric characteristics. Items can be "reused" if they have not been released and not used the previous year. 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.
- **Item types**. A variety of item types, including approximately 20–30% technology-enhanced items, should populate the common slots.

Field-test items are also selected during form construction. Field-test items are drawn from the field-test eligible pools and should mirror the operational test, to the extent needed. If a standard or reporting category is lacking in the common eligible item pool, items should be chosen to fill this need. 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).
- **"Clueing" items**. Items are reviewed for any information that might "clue" or help the student answer another item.
- **Item types**. A variety of item types should populate the matrix slots.

3.2.5.4 OPERATIONAL TEST DRAFT REVIEW

The proposed operational test is posted for DESE to review. DESE test developers consider the proposed items, make recommendations for changes, and then meet with the Cognia's test developers to construct the final forms of the tests. After form construction meetings, the test forms enter several rounds of review by test developers and editors. Items are checked to ensure that requested changes were made after the test construction meetings, and to ensure that all items are scoring correctly. In addition, items are checked again for any grammatical or "fatal flaw" errors and these are corrected before the test forms are published.

3.2.5.5 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 program (IEP) or in a 504 plan. All MCAS 2019 operational tests and retests were available in the following special editions for students with disabilities:

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

- **Braille**—This form included 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 DESE 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 was available only for those students who are blind or have a visual disability. Students who used a screen reader were also given a separate hard-copy Braille edition test in order to have 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 and 10 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 hearing what is being read aloud).
- **American Sign Language (ASL)**—The grade 10 MCAS mathematics computer-based test is available to students who are deaf or hard-of-hearing in an American Sign Language edition, which contains only the common items found in the operational test.
- **Spanish-English**—This version of the Grade 10 mathematics test is intended for Spanish-speaking EL students who have been in the U.S. less than 3 years. Spanish-English tests are available in computer- and paper-based formats. Paper-based tests consist of English-Spanish facing pages (side-by-side); and computer-based tests consist of "stacked" Spanish text above English text. Students may respond either in Spanish or English. (Note: For all other MCAS test versions, students must respond in English.)

Appendix C details other accommodations that did not require a special edition test form and also lists accessibility features that were available to all students, such as screen magnification and highlighting. After testing was completed, DESE received a list with the number of students who participated in the 2019 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 grades 3–8 and 10 next-generation MCAS tests were administered during two overlapping periods in spring 2019 as shown in Table 3-22:

Table 3-22. Test Administration Schedule, ELA and Mathematics Grades 3–8 & 10, STE 5 & 8

Content Area	*Complete the Student Registration/ Personal Needs Profile Process*	*Receive Test Administration Materials*	*Test Administration Windows*	*Deadline to Complete the Principal's Certification of Proper Test Administration (PCPA), Update Students' Accommodations, and Mark CBT as Complete*	*Deadline for Return of Materials to Contractor (for PBT Only)*
Grades 3–8 ELA	January 28–February 8	March 18	April 1–May 3	May 6 to update students' accommodations and mark tests complete	May 7
Grades 3–8 Mathematics	January 28–February 8	March 18	April 2–May 24	May 28 to update students' accommodations and mark tests complete	May 29
Grades 5 & 8 STE	January 28–February 8	March 18	April 3–May 24	May 28 to update students' accommodations and mark tests complete and to complete the PCPA (one combined form for grades 3–8)	May 29
Grade 10 ELA	January 30–February 12	March 12	March 26–April 4[1]	April 5 to update students' accommodations and mark tests complete	April 8
Grade 10 Mathematics	January 30–February 12	May 7	May 21–May 30[2]	May 31 to update students' accommodations and mark tests complete and to complete the PCPA (one combined form for grade 10 ELA and Mathematics)	June 3

[1]Prescribed dates, for schools to test the maximum number of students who could participate concurrently: March 26–27; Administration dates if needed, for schools to test any remaining students who did not participate in the first set of dates due to technology/device limitations: March 28–29

[2]Prescribed dates, for schools to test the maximum number of students who could participate concurrently: May 21–22; Administration dates if needed, for schools to test any remaining students who did not participate in the first set of dates due to technology/device limitations: May 23–24

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 Manual*s. 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 delays in reporting of test results, the invalidation of test results, the removal of school personnel from future MCAS administrations, employment consequences, 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 Pearson proprietary web monitoring tool to perform web monitoring. The Pearson web monitoring system leverages technology tools and human expertise to identify, prioritize, and monitor sites where sensitive test information may be disclosed. The following strategies were used:

- 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 DESE and Cognia, as required;
- worked systematically through the steps necessary to have infringing content removed if a threat was verified; and
- 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 2019 Principal's Administration Manual* (PAM),the *Spring 2019 Test Administrator's Manual for Computer-Based Testing* (CBT TAM), and the *Spring 2019 Test Administrator's Manual for Paper-Based Testing* (PBT TAM).

3.3.3 Participation Requirements

In spring 2019, 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 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

- English learner (EL) students
- students who have been expelled but receive educational services from a district
- foreign exchange students who are coded as #11 under “Reason for Enrollment” in the Student Information Management System (SIMS)

It was the responsibility of the principal to ensure that all enrolled students participated in testing as mandated by state and federal laws. To certify that **all** students participated in testing as required, principals were required to complete the online Principal’s Certification of Proper Test Administration (PCPA) following each test administration. For a summary of participation rates, see the 2019 MCAS Participation Report on DESE’s School and District Profiles website: profiles.doe.mass.edu/mcas/participation.aspx?linkid=26&orgcode=00000000&fycode=2019&orgtypecode=0&.

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 students without disabilities, 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 2019 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 2019 testing can be found in Appendix D; for a list of test accommodations, see Appendix C. The overall number of accommodations has increased in the next-generation MCAS administration because of CBT-specific accommodations such as text-to-speech.);
- completing and ensuring the accuracy of information provided on the PCPA;
- monitoring DESE’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; and
- providing DESE 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

3.4.1 Preparation

3.4.1.1 PREPARATION OF STUDENT RESPONSE BOOKLETS

Scoring of the 2019 MCAS tests was conducted by both Cognia and Pearson. Table 3-23 shows the breakdown of how scoring work was divided between Cognia and Pearson.

Table 3-23. Breakdown of Scoring Work

Cognia	*Pearson*
ELA & Math grade 10 operational ELA & Math grades 3–8 & 10 field tests ELA & Math grades 3–8 operational preparation of expanded training materials for hand-off to Pearson STE grades 5, 8, and HS operational and field tests	ELA & Math grades 3-8 operational

For paper-based tests, Cognia scanned each MCAS student answer booklet. Images for field-test items were loaded into iScore, Cognia's secure scoring platform. Images for operational items were transferred via FTP site to Pearson for uploading into the ePEN scoring platform For computer-based tests, images were uploaded into the appropriate scoring platform so that all scoring was conducted in a similar manner, regardless of the method of test administration.

A set of quality-control procedures was 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 that the scanners were working as intended.

For computer-based test takers, DESE 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 DESE 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 a 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 scaled scores—against the tables provided by Pearson psychometrics.

3.4.1.2 PREPARATION FOR SCORING CONSTRUCTED-RESPONSE ITEMS

Scoring of responses to short-answer, constructed-response, and essay items began by first preparing the documents for scoring. Student identification information, demographic information, and school contact information was converted to alphanumeric format. Digitized student responses to constructed-response items were sorted into specific content areas, grade levels, and items before being scored.

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

- Cognia provided annotated anchor, practice, and qualification sets for all existing items to Pearson for review in advance of scoring. Content specialists at Pearson and Cognia consulted with each other to address any questions and ensure clarity of training materials.
- Cognia 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 re-benchmarked due to modifications, content specialists from Cognia, Pearson, and DESE collaborated on the establishment of final scoring decisions.
- Weekly meetings between the Cognia and Pearson scoring departments were held to address any issues and questions before and during scoring.

3.4.2 Benchmarking Meetings

Samples of student responses to field-test items were read, scored, and discussed by members of Cognia's Scoring Services and Content Development and Publishing (CDP) Departments and by DESE test developers. Each benchmarking meeting is content- and grade-specific (e.g., grade 6 ELA). All decisions were recorded and considered final upon DESE signoff.

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

- revise, as necessary, an item's scoring guide and/or scoring rubric;
- revise, as necessary, an item's scoring notes based on student responses—these, along with scoring guides, provide detailed information about how to score an item;
- assign final score points to a given set of student responses; and
- approve anchor and training sets of responses that are used to train scorers.

3.4.3 Machine-Scored Items

Student responses to selected-response and short-answer items were machine-scored by PearsonAccessNext (PAN) Scoring. Student responses with multiple marks (possible only on paper-based tests) and blank responses were assigned zero points.

3.4.4 Hand-Scored Items

Once responses to hand-scored items were sorted into item-specific groups, student responses were 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 hand-scored items, scoring leadership had access to the student's entire answer booklet. Details on the procedures used to hand-score student responses are provided later in this chapter.

3.4.4.1 SCORING LOCATION AND STAFF

Hand-scoring of responses occurred in various locations, as summarized in Table 3-24.

Table 3-24. Summary of Operational Scoring Locations and Scoring Shifts

Cognia Scoring Sites	*Content*	*Grade*	*Shift*	*Hours*
Dover, NH	Math	10	Day	8:00 a.m.– 4:30 p.m.
Longmont, CO	Math	10	Day	8:00 a.m.– 4:30 p.m.
	Math	10	Night	5:30 p.m.– 10:00 p.m.
	ELA	10	Day	8:00 a.m.– 4:30 p.m.
	ELA	10	Night	5:30 p.m.– 10:00 p.m.
Menands, NY	STE	5, 8, HS	Day	8:00 a.m.– 4:30 p.m.
	STE	5, 8, HS	Night	5:30 p.m.– 10:00 p.m.
Pearson Scoring Sites	*Content*	*Grade*	*Shift*	*Hours*
Charlotte, NC	ELA	3, 7 and one grade 8 Essay	Day	8:00 a.m.– 4:30 p.m.
	ELA	3 Constructed Response	Night	6:00 p.m.– 10:00 p.m.
Columbus, OH	ELA	4 & 5 Essay	Day	8:00 a.m.– 4:30 p.m.
	ELA	4 Constructed Response	Night	6:00 p.m.– 10:00 p.m.
Iowa City, IA	ELA	6	Day	8:00 a.m.– 4:30 p.m.
Mesa, AZ	Mathematics	3	Night	6:00 p.m.– 10:00 p.m.
	ELA	two grade 8 Essays	Day	8:00 a.m.– 4:30 p.m.
San Antonio, TX	Mathematics	6–8	Day	8:00 a.m.– 4:30 p.m.
	Mathematics	4	Night	6:00 p.m.– 10:00 p.m.
Virginia Beach, VA	Math	5	Day	8:00 a.m.– 4:30 p.m.

The following staff members were involved with scoring the 2019 MCAS responses:

- Cognia Staff
  - The *Scoring Director for Content and Quality* was located in Dover, New Hampshire and provided guidance, direction, and leadership to MCAS scoring.
  - The *Scoring Project Manager* was located in Dover, New Hampshire and was responsible for the communication and coordination of MCAS scoring between Cognia and Pearson.
  - *Scoring Content Specialists* facilitated all benchmarking meetings in order to ensure consistency of content area benchmarking and field test scoring across all grade levels at each scoring location. They also handled all aspects for scoring of grade 10 ELA, Math, and grades 5, 8, and HS STE. Scoring content specialists prepared training materials for all operational scoring of ELA & Math grades 3–8 prior to scoring by Pearson., They also fielded any questions between Pearson and Cognia to ensure a consistent scoring approach across the scoring groups and years.
  - *Scoring Supervisors* were responsible for the training and qualification of scorers and Scoring Team Leaders, and for ensuring quality targets for their assignment items.
  - *Scoring Team Leaders* provided support and direction to scorers on quality, accuracy, and timely scoring completion.

- Pearson Staff
    - The *Scoring Portfolio Manager* was located in Iowa City, Iowa and was responsible for the coordination, management, and oversight of the MCAS scoring for Pearson.
    - The *Scoring Project Manager* was located in San Antonio, Texas and oversaw communication and coordination of MCAS scoring between Pearson and Cognia.
    - *Scoring Content Specialists* 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 also coordinated communication with their counterparts at Cognia regarding the training materials.
    - *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.
    - *Automated Scoring Team Members* were located in Boulder, Colorado and were responsible for training and monitoring the scoring performance of the Intelligent Essay Assessor (IEA) on the subset of the ELA prompts selected for automated scoring.

3.4.4.2 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 successful completion of at least two years of college, although hiring preference was given to individuals with a four-year college degree. Those scoring high school students' responses must have at least a 4-year degree and must either have a degree related to the content they were working on OR have at least two classes related to the content and have prior experience in the content area.

Teachers, tutors, and administrators (e.g., principals, guidance counselors) currently under contract or employed by or in Massachusetts schools, and people under 18 years of age, were not eligible to score MCAS responses. Potential scorers were required to submit an application and documentation 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 3-25.

Table 3-25. Summary of Scorer Backgrounds across Scoring Shifts and Scoring Locations (Operational Scoring)

Cognia Education	*Scorers*		*Leadership*	
	Number	*Percent*	*Number*	*Percent*
Master's degree/doctorate	158	36.2	30	42.3
Bachelor's degree	255	58.4	41	57.7
Associate degree/more than 48 college credits	24	5.5	0	0
Less than 48 college credits	0	0	0	0
TOTAL	437		71	
Teaching Experience				
College instructor	40	9.2	9	12.7
Teaching certificate or experience	148	33.9	25	35.2
No teaching certificate of experience	249	57.1	37	52.1
Scoring Experience				
3+ years of experience	130	29.8	44	62.0
1–3 years of experience	110	25.2	27	38.0
No previous experience as scorer/first season	198	45.3	0	0
Pearson Education	*Scorers*		*Leadership*	
	Number	*Percent*	*Number*	*Percent*
Master's degree/doctorate	507	39.6	32	31.1
Bachelor's degree	774	60.4	82	68.9
Associate degree/more than 48 college credits	0	0	0	0
Less than 48 college credits	0	0	0	0
TOTAL	1281		119	
Teaching Experience				
College instructor	8	0.6	1	0.8
Teaching certificate or experience	737	57.6	74	62.2
No teaching certificate of experience	534	41.8	44	37.0
Scoring Experience				
3+ years of experience	188	14.7	45	37.8
1–3 years of experience	339	26.5	68	57.1
No previous experience as scorer/first season	754	58.9	6	5.0

3.4.4.3 SCORER TRAINING

Scoring content specialists had overall responsibility for ensuring that responses were scored consistently, fairly, and according to the approved scoring guidelines. Scoring materials were carefully compiled and checked for consistency, and accuracy. Student identification information, demographic information, and school contact information was not visible to scorers. The sequence and manner in which the materials were presented to scorers was standardized to ensure that all scorers had the same training environment and scoring experience, regardless of scoring location, content, grade level, or item scored.

Four training methods were used to train scorers of MCAS hand-scored items:

- live face-to-face group training
- audio/video conferencing

- live large-group training via headsets
- recorded interactive modules (used for individuals, small groups, or large groups)

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

Scorers started the training process by receiving an overview of 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 DESE during field-test benchmarking meetings and used without any additions or deletions.

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

- **Anchor sets** are DESE-approved sets consisting of two or three sample responses at each score point. Each response represents a typical response, rather than an unusual or uncommon one; is solid and has a true score, meaning that this response has a precise score. Anchor sets are used to exemplify each score point.
- **Practice sets may** 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 DESE-approved scoring standards.

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[2] (at least 7 out of the 10 were exact score matches and either zero or one discrepant) on either of two potential qualifying sets. For multi-trait ELA items, each scorer had to meet the 70% / 90% passing threshold for each individual trait.

3.4.4.4 LEADERSHIP TRAINING

Scoring content specialists also had overall responsibility for ensuring that scoring leadership (Cognia scoring supervisors and Pearson scoring directors) continued their history of scoring consistently, fairly, and according to the approved scoring guidelines. Once they had completed their item-specific training, scoring leadership was required to meet or surpass a qualification standard of at least 80% exact and 90% exact-plus-adjacent scoring accuracy. For multi-trait ELA items, scoring leadership had to meet the 80% and 90% passing threshold for each individual trait.

3.4.4.5 METHODOLOGY FOR SCORING HAND-SCORED POLYTOMOUS ITEMS

In 2019, two scoring methods were used for ELA essay items in grades 3–8. First, hand scoring by human scorers was conducted on all field test items administered in 2018 and used on the 2019 next-generation MCAS tests in

[2] "Adjacent agreement" means that a pair of scores (for the same response) are only off by one point. "Exact-plus-adjacent agreement" means that a pair of scores was either the same or off by only one point.

2019. Next, hand scoring of all operational items was conducted using the procedures described below. In grades 3–8, the 10% double-blind scoring for ten ELA essay items (described below in this section) was conducted via automated scoring, using Pearson's Intelligent Essay Assessor (IEA). The double-blind scoring on the other 3–8 ELA and mathematics items was done by human scorers. Information on how the IEA works and how it was used on the MCAS essay scoring is provided in section 3.4.4.7 below.

3.4.4.6 MONITORING OF SCORING QUALITY CONTROL

The 2019 MCAS tests included constructed response items and essays (in addition to selected-response and short-answer items) that were scored by hand. Hand-scored items included

- constructed-response items with assigned scores of 0–3 (ELA grades 3 and 4 only)
- constructed-response items with assigned scores of 0–3 (mathematics grade 3) and 0–4 (mathematics grades 4–8 and 10);
- constructed-response items with assigned scores of 0–2 and 0–3 (STE grades 5, 8, and HS)
- essays with assigned scores of 0–7 (ELA grades 3–5) and 0–8 (ELA grades 6–8).

For each of these hand-scored items, a scoring guide was created. For examples of item-specific scoring guides, see the MCAS Student Work/Scoring Guides webpage at www.doe.mass.edu/mcas/student/.

The final non-numeric scores assigned by Cognia and Pearson could be designated as:

- Blank: The written response form is completely blank.
- Unreadable: The response cannot be read because of poor penmanship, or spelling cannot be deciphered, or 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 the response states that the student is refusing to participate in testing.
- Direct Copy: Direct copy of text from the passage or item prompt.

Scorers at both Cognia and Pearson could also flag a response as a "Crisis" response, which would be sent to scoring leadership for immediate attention.

A response would 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 language or thoughts that may require professional intervention;
- 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 once by a single scorer) or double-blind scored (response was independently read and scored by two scorers).

Double-Blind Scoring

In double-blind scoring, scorers were not aware that double-blind scoring was taking place. For a double-blind response with adjacent scores (within one point of each other) the higher score was used. Any double-blind response with discrepant scores greater than one point was sent to the arbitration queue and read by scoring leadership, where the expert score resolved the scoring discrepancy.

Double-blind scoring with the IEA scoring platform was conducted on 10% of the responses for ten ELA essay items across grades 3-8. For the remaining items in grades 3–8, human scorers conducted double-blind scoring at a rate of 10%. For the grade 10 ELA essay items, human scorers conducted double-blind scoring at a 100% rate.

A description of how the IEA functions and how it was used is provided in section 3.4.4.7. Scoring agreement statistics provided in Tables 3-28 and 3-29 are based on comparing human scoring to the 10% double-blind scoring (IEA scoring or human scoring depending on the prompt).

Read-Behind Scoring

In addition to the 10% or 100% double-blind scoring, scoring leadership, at random points throughout the scoring shift, engaged in read-behind (back-read) scoring for each scorer assigned to their team. In this process, scoring leadership views responses recently scored by a particular scorer and assigns a score to that same response. Scoring leadership then compares scores and advises or counsels the scorer as necessary.

Table 3-26 illustrates how the rules were applied for instances when two read-behind scores were not an exact match or when two scorers conducting double-blind scoring gave scores that did not match.

Table 3-26. Read-Behind and Double-Blind Resolution Examples

Read-Behind Scoring[1]			
Scorer #1	Scorer #2	Scoring Leadership Resolution	Final
4	--	4	4
3	3	4	4
3	--	2	2
Double-Blind Scoring[2] *of 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

[1] *In all cases, the scoring leadership score is the final score of record.*
[2] *If double-blind scores are adjacent (only 1 point different), the higher score is used as the final score. If double-blind scores are neither exact nor adjacent, the resolution score is used as the final score.*

3.4.4.7 DOUBLE-BLIND SCORING WITH THE INTELLIGENT ESSAY ASSESSOR (IEA)

The Intelligent Essay Assessor (IEA) is used to score student responses to essay prompts.[3] Like human scorers, IEA evaluates the content and meaning of text, as well as grammar, style, and mechanics. IEA learns to score via a range of machine learning and natural language processing technologies. The engine is trained individually on each prompt and trait using hundreds or thousands of human-scored student responses.

[3] Additional information about IEA can be found in Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K (2013). Implementation and applications of the Intelligent Essay Assessor. *Handbook of Automated Essay Evaluation*, M. Shermis & J. Burstein, (Eds.). Pp. 68-88. Routledge, NY, NY.

IEA measures the content and quality of responses by determining the features that human scorers evaluate when scoring a response. Given a set of human-scored responses to a prompt, IEA computes hundreds of different metrics that characterize each response in numerical ways. Some examples of these metrics include:

- Number of grammar errors
- Types of grammar errors
- Variety of words
- Maturity of words
- Variety of sentence types
- Coherence of the response
- Similarity of the response to other responses and/or source materials

All of these different metrics are fed to machine learning algorithms that determine which of them best predict the scores assigned by human scorers.

One of the hallmarks of IEA is its ability to score constructed responses in content areas beyond just ELA using a unique implementation of Latent Semantic Analysis (LSA). LSA analyzes large bodies of relevant text to generate semantic similarity of words and passages. LSA can then "understand" the meaning of text in much the same way as a human scorer.

IEA's background knowledge of English is based on a collection of text of about 12 million words—roughly the amount of text a student will read over the course of their academic career. Because LSA operates over the semantic representation of texts, rather than at the individual word level, it can evaluate similarity even when texts have few or no words in common. For example, LSA finds the following two sentences to have a high semantic similarity:

- Surgery is often performed by a team of doctors.
- On many occasions, several physicians are involved in an operation.

After several years of study on MCAS prompts, IEA was used operationally this year as the second double-blind score. IEA was trained before the operational assessment was administered using responses collected during the field test and scored by trained human scorers. For each prompt, IEA was trained using approximately 1200 responses per prompt and then evaluated using approximately 600 responses. Table 3-27 includes the specific N counts for each prompt. The responses were randomly assigned to each set (training or evaluation). Performance on the evaluation set was measured using a variety of criteria comparing IEA with human scoring using the industry standard metrics shown in Table 3-28.

Table 3-27. N Counts by Prompt

Grade	*Prompt*	*Training Set Size*	*Evaluation Set Size*
3	EL293264	1208	590
4	EL710438990	1209	586
5	EL709062207	1200	595
5	EL709229186	1200	597
6	EL707351199	1213	586
6	EL712756190	1205	593
7	EL707935717	1209	589
7	EL714343909	1121	546
8	EL709184717	1198	587
8	EL709130565	1206	591

Table 3-28. Industry Standard Metrics for Evaluating Automated Scoring[4]

Measure	*Threshold*
Pearson R	≥ 0.70
Quadratic Weighted Kappa (QWK)	≥ 0.70
Kappa	≥ 0.40
Exact Agreement	≥ 65% (or better than human-human agreement)
Per score point agreement	≥ 50% (or better than human-human agreement)
Standardized Mean Difference (SMD)	Within |0.15|

Thirteen of the sixteen prompts met the required performance criteria and ten were approved by DESE to be scored by IEA as the double-blind score to monitor quality during the operational assessment. Scoring performance on the operational assessment is described in the next section.

A comparison of the performance of IEA to human scoring on exact agreement by score point is presented in Table 3-29. As indicated, IEA accuracy is similar to or slightly higher than the human scoring accuracy at all score points. In particular, IEA accuracy tends to be higher than human accuracy at the highest score point, as seen in the idea development agreement statistics for grades 3–6 and in the idea development scores for the second prompt at grade 8.

Table 3-29. Comparison of Human and IEA Agreement with Validity Papers—ELA

Grade	*UIN*	*Trait*	*Validity*	*Exact*	*Exact Agreement by Score Point*					
					0	*1*	*2*	*3*	*4*	*5*
3	EL293264	Idea Development	IEA	91%	100%	83%	93%	93%	100%	n/a
			Human	90%	96%	94%	82%	62%	88%	n/a
			N	92	15	35	14	15	13	n/a
		Conventions	IEA	89%	94%	94%	93%	79%	n/a	n/a
			Human	92%	95%	93%	81%	94%	n/a	n/a
			N	92	16	34	14	28	n/a	n/a
4	EL710438990	Idea Development	IEA	98%	100%	97%	98%	90%	100%	n/a
			Human	83%	86%	87%	83%	67%	78%	n/a
			N	161	44	33	46	21	17	n/a
		Conventions	IEA	91%	100%	91%	89%	82%	n/a	n/a
			Human	88%	93%	89%	82%	84%	n/a	n/a
			N	161	44	33	46	38	n/a	n/a
5	EL709062207	Idea Development	IEA	91%	93%	96%	88%	76%	100%	n/a
			Human	85%	82%	95%	78%	62%	52%	n/a
			N	214	27	102	48	33	4	n/a
		Conventions	IEA	91%	89%	93%	83%	95%	n/a	n/a
			Human	85%	87%	90%	79%	75%	n/a	n/a
			N	214	37	92	48	37	n/a	n/a
	EL709229186	Idea Development	IEA	91%	88%	93%	92%	82%	100%	n/a
			Human	87%	94%	90%	82%	66%	82%	n/a
			N	100	34	27	13	11	15	n/a
		Conventions	IEA	91%	100%	94%	69%	92%	n/a	n/a
			Human	89%	92%	90%	81%	93%	n/a	n/a
			N	100	10	51	13	26	n/a	n/a

continued

[4] Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practices, 31,* 2.

Grade	UIN	Trait	Validity	Exact	Exact Agreement by Score Point					
					0	1	2	3	4	5
6	EL707351199	Idea Development	IEA	89%	100%	81%	94%	85%	88%	100%
			Human	84%	94%	94%	82%	83%	64%	81%
			N	75	10	16	16	13	17	3
		Conventions	IEA	93%	93%	100%	77%	97%	n/a	n/a
			Human	86%	88%	86%	78%	90%	n/a	n/a
			N	75	14	14	13	34	n/a	n/a
	EL712756190	Idea Development	IEA	88%	80%	100%	95%	78%	100%	0%
			Human	87%	90%	92%	89%	77%	71%	84%
			N	81	10	26	19	18	5	3
		Conventions	IEA	95%	100%	93%	95%	96%	n/a	n/a
			Human	91%	94%	92%	86%	91%	n/a	n/a
			N	81	8	28	20	25	n/a	n/a
7	EL707935717	Idea Development	IEA	83%	90%	83%	100%	75%	80%	33%
			Human	88%	94%	92%	88%	86%	58%	44%
			N	64	20	18	6	12	5	3
		Conventions	IEA	95%	96%	82%	100%	100%	n/a	n/a
			Human	92%	96%	91%	85%	90%	n/a	n/a
			N	64	27	11	6	20	n/a	n/a
	EL714343909	Idea Development	IEA	80%	83%	76%	80%	85%	55%	100%
			Human	86%	95%	90%	88%	72%	65%	73%
			N	138	36	25	35	26	11	5
		Conventions	IEA	86%	79%	81%	82%	98%	n/a	n/a
			Human	84%	87%	78%	73%	92%	n/a	n/a
			N	138	42	21	34	41	n/a	n/a
8	EL709130565	Idea Development	IEA	89%	100%	100%	100%	93%	50%	57%
			Human	84%	97%	86%	76%	86%	51%	57%
			N	76	23	14	10	14	8	7
		Conventions	IEA	96%	100%	92%	100%	93%	n/a	n/a
			Human	91%	97%	85%	78%	93%	n/a	n/a
			N	76	24	13	10	29	n/a	n/a
	EL709184717	Idea Development	IEA	98%	100%	96%	100%	96%	100%	100%
			Human	82%	78%	94%	87%	81%	57%	60%
			N	114	9	25	24	27	21	8
		Conventions	IEA	95%	100%	95%	96%	93%	n/a	n/a
			Human	87%	91%	88%	85%	88%	n/a	n/a
			N	114	15	19	24	56	n/a	n/a

3.4.4.8 MONITORING OF SCORING QUALITY

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, some form of intervention occurred, 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 quality control methods listed and further defined below:

- daily recalibration set (Cognia),
- embedded responses (Cognia),
- validity responses (Pearson),

- read-behind scoring (RBs)/back-reading,
- double-blind scoring (DBs), and
- compilation reports (summary of scoring agreement statistics).

Daily recalibration sets (Cognia) were administered at the very beginning of a scoring shift and each set consisted of five responses representing various scores. If scorers had an exact score match on at least four of the five responses, and were at least adjacent on the fifth response, they were allowed to begin scoring operational responses. Scorers who had discrepant scores, or only two or three exact score matches, were retrained and, if approved by leadership, were allowed to return to scoring with extra monitoring. Scorers who had zero or one out of the five exact were typically reassigned to another item or released for the day.

Embedded responses (Cognia) were approved by the scoring content specialist and loaded into iScore for blind distribution to scorers at random points during the scoring of their first 200 operational responses. Embedded responses comprised 5% of responses scored by a scorer during this period. Scorers who fell below the 70% exact and 90% exact-plus-adjacent accuracy standard were provided counseling and additional read-behind monitoring.

Validity responses (Pearson) were used to monitor the scorer's accuracy of scoring. These responses were approved by scoring leadership and distributed to scorers based on a percentage of their total number of responses scored. For the first two days, validity responses routed to scorers comprised 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.

Alert messages were issued to scorers who did not meet minimum validity metrics after 10 validity responses. If after an additional five validity responses, the scorer had not improved, ePEN automatically blocked that scorer, and launched a 10-response targeted calibration set. 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 item for which the calibration set was administered. If the scorer passed the targeted calibration, ePEN was unblocked and the scorer regained admission to operational responses. 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.

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 scorer and at least five times for each evening shift and partial-day shift scorer. 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. 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. Responses given discrepant scores by two independent scorers were read and scored by scoring leadership.

Compilation reports were generated at both Cognia and Pearson. Compilation reports displayed all the statistics for each scorer, including the percentage of exact, adjacent, and discrepant scores on the RBs as well as the percentage of exact, adjacent, and discrepant scores on recalibration sets (Cognia) or validity sets (Pearson). As

scoring leadership conducted RBs, the scorers' overall percentages on the compilation report were automatically calculated and updated. If the compilation report at the end of the scoring shift listed any individuals who were still below the 70% exact and 90% exact-plus-adjacent standard, their scores for that day were voided. Responses with voided scores were returned to the scoring queue for other scorers to score.

3.4.4.9 INTERRATER CONSISTENCY

Interrater consistency statistics are evaluated 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. Ten percent of constructed-response items in grades 3–8 were randomly selected and scored independently by two different scorers. As described in the previous section, for ten of those prompts, IEA was the second scorer. Results of the double-blind scoring were used 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 3-30. Results in the table are organized by content area and grade. The table shows the number of score categories (number of possible scores for an item type), the number of included scores, the exact agreement percentage, the adjacent agreement percentage, the 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. Linearly weighted kappa is also included in Table 3-30 as a measure of scorer consistency by accounting for chance agreement. It is defined as:

$$\kappa = \frac{O - E}{1 - E},$$

where

$$O = \sum_{i=1}^{n}\sum_{j=1}^{n}\left[1 - \frac{|i-j|}{n-1}\right]a_{ij} \text{ and } E = \sum_{i=1}^{n}\sum_{j=1}^{n}\left[1 - \frac{|i-j|}{n-1}\right]p_i q_j,$$

with a_{ij} being the proportion of that scorer 1 gives score i and scorer 2 gives score j, p_i being the proportion of that scorer 1 gives score i and q_i being the proportion of that scorer 2 gives score j. O and E are observed agreement and chance agreement, respectively.

Table 3-30. Summary of Interrater Consistency Statistics Organized across Items by Content Area and Grade

Content Area	Grade	Number of		Percentage		Correlation	% Third Scores[1]	LW Kappa
		Score Categories	Included Scores	Exact	Adjacent			
ELA	3	4	19,844	73.8	25.8	0.81	0.81	0.683
		5	13,218	74.4	24.7	0.82	1.10	0.683
	4	4	20,456	72.6	26.7	0.81	1.33	0.690
		5	13,573	67.6	31.1	0.79	1.84	0.648
	5	4	21,153	70.0	29.1	0.80	1.96	0.685
		5	21,153	68.0	30.4	0.84	1.96	0.696
	6	4	21,132	71.1	27.9	0.84	2.41	0.726
		6	21,132	67.1	31.1	0.86	2.41	0.723
	7	4	20,762	72.1	27.5	0.85	1.10	0.728
		6	20,762	67.7	31.6	0.88	1.10	0.722
	8	4	20,604	74.2	25.2	0.85	2.15	0.742
		6	20,604	66.2	31.9	0.87	2.15	0.726
	10	4	139,148	74.9	24.2	0.81	2.05	0.705
		6	139,148	63.6	34.8	0.84	2.05	0.694
Mathematics	3	4	26,941	91.2	7.8	0.95	0.97	0.917
	4	5	27,881	86.3	12.6	0.94	1.18	0.886
	5	5	28,676	86.1	13.0	0.96	0.94	0.903
	6	5	28,284	87.4	11.9	0.96	0.75	0.914
	7	5	27,888	88.1	11.2	0.97	0.68	0.916
	8	5	20,840	88.3	11.7	0.97	0.02	0.924
	10	5	281,573	86.5	12.9	0.96	0.59	0.904
STE	5	3	7,277	76.7	23.0	0.69	0.32	0.628
		4	29,091	72.3	26.0	0.84	1.69	0.698
	8	3	7,251	70.2	28.3	0.66	1.46	0.579
		4	21,223	81.6	17.7	0.88	0.65	0.820

[1]For ELA items, percentages of exact, adjacent, and third score do not sum exactly to 100%. This is because resolutions are done by item and it is entirely possible that only one trait (either idea development or conventions) on a writing item has a non-adjacent score. For instance, if the idea development score for an item were non-adjacent, the item would also receive a third score for conventions, even if it initially received a non-adjacent score for conventions.

Table 3-30 summarizes the interrater consistency across score categories for the double-blind scored responses. To evaluate the interrater consistency at each score, Table 3-31 summarizes the proportion of exact agreement by score points at the test level. Item-level results are also included in Appendix F. The proportion of exact agreement at each score point is calculated as the proportion of responses where the double-blind scores are the same as the initial score at each score point. As noted in section 3.4.4.6, the double-blind scores for ten of the grades 3–8 essay responses are generated by IEA.

Table 3-31. Summary of Proportion of Exact Agreement by Score Points

Content Area	Grade	Number of		Score Points					
		Score Categories	Included Scores	0	1	2	3	4	5
ELA	3	4	19,844	70.1	75.3	66.9	59.1		
	3	5	13,218	73.2	80.5	65.5	52.6	48.3	
	4	4	20,456	79.1	71.5	69.7	64.0		
	4	5	13,573	67.8	67.9	67.6	53.6	51.6	
	5	4	21,153	69.0	73.1	62.1	78.8		
	5	5	21,153	78.4	62.8	67.8	57.5	53.6	
	6	4	21,132	69.3	70.1	64.3	83.0		
	6	6	21,132	74.6	75.3	69.0	57.6	48.1	65.0
	7	4	20,762	68.3	62.3	66.9	85.9		
	7	6	20,762	69.0	68.1	70.9	58.6	50.3	54.1
	8	4	20,604	73.4	62.3	69.7	85.2		
	8	6	20,604	77.2	70.3	70.0	56.8	52.3	55.7
	10	4	139,148	60.8	59.4	65.1	86.0		
	10	6	139,148	58.0	64.8	68.1	62.8	62.4	29.2
Mathematics	3	4	26,941	95.8	91.7	77.9	93.2		
	4	5	27,881	88.4	76.3	84.2	81.3	89.6	
	5	5	28,676	89.9	78.8	84.2	78.6	92.6	
	6	5	28,284	93.0	81.1	77.6	82.8	94.6	
	7	5	27,888	95.4	88.1	85.4	86.5	89.4	
	8	5	20,840	97.5	92.0	82.6	76.6	92.8	
	10	5	281,573	95.2	81.6	77.8	77.5	76.3	
STE	5	3	7,277	70.7	75.3	79.2			
		4	29,091	84.8	69.1	62.3	62.8		
	8	3	7,251	76.8	68.2	58.8			
		4	21,223	90.3	83.2	76.3	73.6		

As described in section 3.4.4.8, validity responses were used to monitor the scorer's accuracy of scoring. Table 3-32 provides a summary of "validity" statistics. These statistics denote accuracy in scoring; they provide an average of the human and IEA agreement with the validity responses (which provide the true scores for each essay).

Table 3-32. Summary of Validity Statistics

Content Area	*Grade*	*Number of Score Categories*	*Number of Included Validity Responses*	*Exact Agreement (%)*	*Agreement by Score Point (%)*					
					0	*1*	*2*	*3*	*4*	*5*
ELA	3	3	174	93.6	100.0	90.4	100.0			
		4	15,659	90.3	91.2	94.2	81.9	85.2		
		5	3,685	90.0	95.6	93.8	82.3	61.9	88.0	
	4	3	231	89.9	87.9	90.6	100.0			
		4	11,886	86.6	92.9	88.7	83.7	81.9		
		5	7,326	85.3	86.2	91.7	83.6	67.3	78.5	
	5	4	11,619	87.7	89.2	90.2	83.4	86.4		
		5	11,215	86.0	92.3	93.3	83.4	68.8	74.4	
	6	4	12,042	87.7	90.1	89.3	78.9	89.5		
		5	186	89.8	75.0	96.1	88.2	91.9	71.0	
		6	11,859	86.2	94.7	93.7	82.7	77.7	62.0	80.5
	7	4	11,535	90.0	93.6	89.7	80.7	90.6		
		5	4,027	92.8	93.8	95.4	89.8	77.1	61.0	
		6	7,519	87.6	94.4	91.2	88.5	78.4	63.2	64.6
	8	4	11,228	87.6	92.7	85.5	79.6	89.3		
		5	3,860	90.2	95.8	92.2	73.0	67.4	50.1	
		6	7,371	83.1	93.8	90.7	82.9	82.5	55.5	58.2
Mathematics	3	4	8,416	96.1	97.8	95.5	91.9	98.1		
	4	5	8,462	93.4	97.2	86.4	92.4	93.1	97.8	
	5	5	8,557	94.6	97.6	91.4	93.9	91.6	96.7	
	6	5	8,603	96.2	97.9	93.2	94.6	96.1	98.3	
	7	5	8,183	93.3	98.5	93.9	89.8	93.7	92.7	
	8	4	1,779	97.3	99.8	91.1	97.5	100.0		
		5	6,700	94.1	98.3	95.3	91.0	88.7	96.5	

Table 3-33 provides a breakout of IEA's performance on the validity responses in comparison with human scorer performance for the ten prompts that IEA scored.

Table 3-33. Comparison of Human and IEA Agreement with Validity Responses—ELA

Grade	*UIN*	*Trait*	*Validity*	*Exact*	*Exact Agreement by Score Point*					
					0	*1*	*2*	*3*	*4*	*5*
3	EL293264	Idea Development	IEA	91%	100%	83%	93%	93%	100%	n/a
			Human	90%	96%	94%	82%	62%	88%	n/a
			N	92	15	35	14	15	13	n/a
		Conventions	IEA	89%	94%	94%	93%	79%	n/a	n/a
			Human	92%	95%	93%	81%	94%	n/a	n/a
			N	92	16	34	14	28	n/a	n/a
4	EL710438990	Idea Development	IEA	98%	100%	97%	98%	90%	100%	n/a
			Human	83%	86%	87%	83%	67%	78%	n/a
			N	161	44	33	46	21	17	n/a
		Conventions	IEA	91%	100%	91%	89%	82%	n/a	n/a
			Human	88%	93%	89%	82%	84%	n/a	n/a
			N	161	44	33	46	38	n/a	n/a
5	EL709062207	Idea Development	IEA	91%	93%	96%	88%	76%	100%	n/a
			Human	85%	82%	95%	78%	62%	52%	n/a
			N	214	27	102	48	33	4	n/a
		Conventions	IEA	91%	89%	93%	83%	95%	n/a	n/a
			Human	85%	87%	90%	79%	75%	n/a	n/a
			N	214	37	92	48	37	n/a	n/a
	EL709229186	Idea Development	IEA	91%	88%	93%	92%	82%	100%	n/a
			Human	87%	94%	90%	82%	66%	82%	n/a
			N	100	34	27	13	11	15	n/a
		Conventions	IEA	91%	100%	94%	69%	92%	n/a	n/a
			Human	89%	92%	90%	81%	93%	n/a	n/a
			N	100	10	51	13	26	n/a	n/a
6	EL707351199	Idea Development	IEA	89%	100%	81%	94%	85%	88%	100%
			Human	84%	94%	94%	82%	83%	64%	81%
			N	75	10	16	16	13	17	3
		Conventions	IEA	93%	93%	100%	77%	97%	n/a	n/a
			Human	86%	88%	86%	78%	90%	n/a	n/a
			N	75	14	14	13	34	n/a	n/a
	EL712756190	Idea Development	IEA	88%	80%	100%	95%	78%	100%	0%
			Human	87%	90%	92%	89%	77%	71%	84%
			N	81	10	26	19	18	5	3
		Conventions	IEA	95%	100%	93%	95%	96%	n/a	n/a
			Human	91%	94%	92%	86%	91%	n/a	n/a
			N	81	8	28	20	25	n/a	n/a

continued

Grade	UIN	Trait	Validity	Exact	Exact Agreement by Score Point					
					0	1	2	3	4	5
7	EL707935717	Idea Development	IEA	83%	90%	83%	100%	75%	80%	33%
			Human	88%	94%	92%	88%	86%	58%	44%
			N	64	20	18	6	12	5	3
		Conventions	IEA	95%	96%	82%	100%	100%	n/a	n/a
			Human	92%	96%	91%	85%	90%	n/a	n/a
			N	64	27	11	6	20	n/a	n/a
	EL714343909	Idea Development	IEA	80%	83%	76%	80%	85%	55%	100%
			Human	86%	95%	90%	88%	72%	65%	73%
			N	138	36	25	35	26	11	5
	EL714343909	Conventions	IEA	86%	79%	81%	82%	98%	n/a	n/a
			Human	84%	87%	78%	73%	92%	n/a	n/a
			N	138	42	21	34	41	n/a	n/a
8	EL709130565	Idea Development	IEA	89%	100%	100%	100%	93%	50%	57%
			Human	84%	97%	86%	76%	86%	51%	57%
			N	76	23	14	10	14	8	7
		Conventions	IEA	96%	100%	92%	100%	93%	n/a	n/a
			Human	91%	97%	85%	78%	93%	n/a	n/a
			N	76	24	13	10	29	n/a	n/a
	EL709184717	Idea Development	IEA	98%	100%	96%	100%	96%	100%	100%
			Human	82%	78%	94%	87%	81%	57%	60%
			N	114	9	25	24	27	21	8
		Conventions	IEA	95%	100%	95%	96%	93%	n/a	n/a
			Human	87%	91%	88%	85%	88%	n/a	n/a
			N	114	15	19	24	56	n/a	n/a

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 have been conducted to ensure that 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 MCAS assessments in spring 2019. 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, have also been performed for field-test items; the statistics are used during the item review process and during form assembly for future administrations.)

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 full credit for the item (i.e., all of the item points).

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 0.40 and 0.70. 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.

It is desirable for an item to be one on which higher-ability students perform better than lower-ability students. The correlation between student performance on a single item and total test score is a commonly used measure of this item characteristic. 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 2019 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 0.20 to 0.60.

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 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 3-34. 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. Also note that the numbers of items in Table 3-34 may differ from those found in other reports from the 2019 MCAS because they include both technology-enhanced items used only on the CBT and "paper cousin" items used only on the PBT. For more about paper cousins, see section 3.2.1.3.

Table 3-34. Summary of Item Difficulty and Discrimination Statistics by Content Area and Grade

Content Area	*Grade*	*Item Type*	*Number of Items*	*p*-Value		*Discrimination*	
				Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
ELA	3	ALL	29	0.64	0.17	0.41	0.11
		CR	1	0.40		0.41	
		ES	4	0.40	0.09	0.61	0.07
		SR	24	0.69	0.13	0.37	0.08
	4	ALL	29	0.65	0.14	0.45	0.12
		CR	1	0.59		0.57	
		ES	4	0.40	0.05	0.65	0.02
		SR	24	0.70	0.10	0.41	0.10
	5	ALL	31	0.67	0.15	0.43	0.14
		ES	6	0.45	0.08	0.66	0.05
		SR	25	0.72	0.11	0.37	0.07
	6	ALL	29	0.59	0.12	0.47	0.15
		ES	6	0.47	0.12	0.74	0.02
		SR	23	0.63	0.10	0.40	0.07
	7	ALL	30	0.61	0.13	0.46	0.16
		ES	6	0.47	0.12	0.73	0.02
		SR	24	0.65	0.11	0.39	0.10
	8	ALL	30	0.62	0.13	0.47	0.15
		ES	6	0.53	0.13	0.74	0.02
		SR	24	0.65	0.12	0.40	0.06
	10	ALL	36	0.78	0.11	0.46	0.12
		ES	4	0.66	0.14	0.72	0.02
		SR	32	0.79	0.09	0.43	0.08
Mathematics	3	ALL	92	0.62	0.17	0.49	0.08
		CR	9	0.42	0.11	0.64	0.04
		SA	21	0.58	0.16	0.51	0.05
		SR	62	0.66	0.15	0.46	0.07
	4	ALL	92	0.63	0.16	0.48	0.10
		CR	8	0.53	0.09	0.65	0.02
		SA	15	0.63	0.17	0.49	0.09
		SR	69	0.64	0.16	0.46	0.09
	5	ALL	54	0.56	0.17	0.47	0.13
		CR	4	0.53	0.07	0.69	0.02
		SA	13	0.51	0.14	0.52	0.05
		SR	37	0.57	0.19	0.42	0.12
	6	ALL	51	0.57	0.17	0.50	0.11
		CR	4	0.50	0.11	0.73	0.04
		SA	5	0.64	0.17	0.50	0.09
		SR	42	0.57	0.17	0.47	0.08

continued

Content Area	*Grade*	*Item Type*	*Number of Items*	*p*-Value		*Discrimination*	
				Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
Mathematics	7	ALL	53	0.44	0.16	0.52	0.13
		CR	5	0.49	0.12	0.75	0.03
		SA	13	0.39	0.18	0.53	0.08
		SR	35	0.45	0.15	0.48	0.12
	8	ALL	52	0.57	0.16	0.50	0.12
		CR	5	0.50	0.09	0.77	0.04
		SA	11	0.53	0.19	0.51	0.08
		SR	36	0.59	0.15	0.46	0.09
	10	ALL	108	0.61	0.13	0.53	0.12
		CR	8	0.48	0.05	0.76	0.03
		SA	16	0.59	0.16	0.56	0.09
		SR	84	0.63	0.13	0.50	0.11
STE	5	ALL	51	0.61	0.13	0.39	0.10
		ES	7	0.47	0.15	0.51	0.08
		SR	2	0.66	0.06	0.33	0.03
	8	ALL	42	0.63	0.11	0.37	0.09
		ES	56	0.57	0.16	0.41	0.11
		SR	8	0.47	0.16	0.56	0.06

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, and therefore 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 classical statistics are provided in Appendix G. On these MCAS items, the item difficulty and discrimination indices are within generally acceptable and expected ranges. Very few items were answered correctly at near-chance or near-perfect rates. Similarly, the positive discrimination indices indicate that students who performed well on individual items tended to perform well overall. There are a small number of items with discrimination indices below 0.20, but none were negative. While it is acceptable to include items with low discrimination values or with very high or very low item difficulty values when their content is needed to ensure that the content specifications are appropriately covered, there were very few such cases on the 2019 MCAS. Item-level score point distributions are provided for constructed-response items in Appendix H; for each item, the percentage of students who received each score point is presented.

3.5.2 DIF

The *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004) explicitly states that subgroup differences in performance 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 the 2019 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 2019 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. DIF statistics were calculated for all subgroups with at least 75 students.

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 indicate actual item bias. Consequently, all items with either high or low 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) suggested that index values between -0.05 and 0.05 denote either a negligible amount of DIF or the absence of DIF. The majority of 2019 MCAS items fell within this range. Dorans and Holland further stated that items with values between -0.10 and -0.05 and between 0.05 and 0.10 (i.e., "low" DIF) should be inspected to ensure that no possible effect is overlooked, and that items with values outside the -0.10 to 0.10 range (i.e., "high" DIF) are more unusual and should be examined very carefully before being used operationally.

For the 2019 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 compared with female,
- not LEP/FLEP compared with LEP/FLEP,[5]
- not economically disadvantaged compared with economically disadvantaged,
- white compared with African American/Black,
- white compared with Hispanic or Latino, and
- without disabilities compared to with disabilities.

After the 2019 Spring administration, DIF analyses were conducted again as a post-hoc quality check based on the operational data. The tables in Appendix I present the number of items classified as either "low" or "high" DIF, in total and by group favored. Very few items exhibited high DIF in the operational data, which suggested that the bias and sensitivity review that occurred after the field testing effectively ruled out large DIF for the MCAS 2019 Spring

[5] LEP=Limited English Proficiency / FLEP=Former Limited English Proficiency.

tests. Note that the numbers of items in Appendix I include both technology-enhanced items used on the CBT and "paper cousin" items used on the PBT (see section 3.2.1.3. for more about paper cousins).

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 2019 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 MCAS test forms used during the spring 2019 administrations. The majority of students took the test on a computer, and the paper-based test forms were treated as accommodated forms. A total of 16 computer-based test forms were analyzed; the results for these analyses are reported in sections 3.5.3.1 and 3.5.3.2 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 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 from the nonclustered items. The DIMTEST statistic follows a standard normal distribution under the null hypothesis of unidimensionality.

DETECT is an effect-size measure of multidimensionality. As with DIMTEST, the data are first randomly divided into a training sample and a cross-validation sample (these samples are drawn independently of those used with DIMTEST). The training sample is used to find a set of mutually exclusive and collectively exhaustive clusters of items that best fit a systematic pattern of positive conditional covariances for pairs of items from the same cluster and negative conditional covariances for pairs composed of items from different clusters. Next, the clusters from the training sample are used with the cross-validation sample data to average the conditional covariances: within-cluster conditional covariances are summed; from this sum, the between-cluster conditional covariances are subtracted. This difference is divided by the total number of item pairs, and this average is multiplied by 100 to yield an index of the average violation of local independence for an item pair. DETECT values less than 0.2 indicate very weak multidimensionality (or near unidimensionality); values of 0.2 to 0.4, weak to moderate multidimensionality; values of 0.4 to 1.0, moderate to strong multidimensionality; and values greater than 1.0, very strong multidimensionality (Roussos & Ozbek, 2006).

DIMTEST and DETECT were applied to the operational items of the MCAS tests administered during spring 2019. For all computer-based forms, there were over 56,000 student examinees per grade in all subjects[6]. 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, and 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 0.05 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 all of the data sets, 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 the 2017 to 2019 tests. Table 3-35 displays the multidimensionality effect-size estimates from DETECT.

Table 3-35. Multidimensionality Effect Sizes by Grade, and Content Area

Content Area	*Grade*	*Effect Size*		
		2017	*2018*	*2019*
ELA	3	0.25	0.17	0.27
	4	0.30	0.35	0.29
	5	0.35	0.28	0.34
	6	0.38	0.26	0.42
	7	0.34	0.34	0.49
	8	0.38	0.35	0.47
	10	0.20	0.24	0.26
	Average	0.33	0.29	0.36
Mathematics	3	0.20	0.17	0.20
	4	0.19	0.22	0.10
	5	0.19	0.15	0.15
	6	0.21	0.13	0.21
	7	0.13	0.14	0.15
	8	0.11	0.15	0.13
	10	0.12	0.09	0.09
	Average	0.17	0.16	0.15
STE	5	0.08	0.11	0.08
	8	0.08	0.13	0.08

[6] Students taking the computer-based accommodated forms, including text-to-speech, human reader, and screen reader forms, are not included in the dimensionality analysis because they are likely to have a different dimensionality structure from the non-accommodated sample, and they are not included in the IRT and equating analysis.

The DETECT values indicate weak or very weak multidimensionality for all the 2019 mathematics and next-generation STE test forms, which are consistent with previous years' results. The 2019 ELA test forms in both modes show weak to moderate multidimensionality (with the DETECT effect size indicating stronger multidimensionality). The effect size magnitudes for ELA in higher grades become slightly higher than previous years' values; this increase could be due to various causes, such as students' differential growth on different types of items or changes in item discrimination or difficulty. Despite the higher effect size magnitudes, the cluster patterns are exactly the same between 2019 and 2018: for each test in ELA, the essay and constructed-response items tend to form a different cluster from the selected-response items, especially for tests in grades 6 to 8 (see below).

The way in which DETECT divided the tests into clusters was 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. Specifically, for the ELA test forms, every grade had one set of clusters dominated by selected-response items and another set of clusters dominated by writing prompt items. On the mathematics and next-generation STE test forms, there was less clear evidence of consistent separation of selected-response and constructed-response items.

In summary, for the 2019 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 previous years, with ELA tending to display more separation than mathematics.

3.6 MCAS IRT Linking and Scaling

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

- evaluations of the calibration processes (e.g., checking the number of cycles required for convergence for reasonableness);
- checking item parameters and their standard errors for reasonableness;
- examination of test characteristic curves (TCCs) and test information functions (TIFs) for reasonableness;
- evaluation of model fit;
- evaluation of equating items (e.g., delta analyses, *b*-*b* analyses, *beta* analyses);
- examination of *a*-plots and *b*-plots for reasonableness; and
- evaluation of the scaling results (e.g., comparing look-up tables to the previous year's).

Section 3.6.3 summarizes the equating procedure and results to place the 2019 next-generation MCAS tests on the same scale as the previous year. An equating report, which provided complete documentation of the quality-control procedures and results, was reviewed by the DESE and approved prior to production of the *Spring 2019 MCAS Tests Parent/Guardian Report*s (Cognia Psychometrics and Research Department, *2018-2019 MCAS Equating Report*, unpublished manuscript).

In addition to performing year-to-year equating, a special mode comparability analysis was conducted for one district where there was a testing mode difference for the majority of schools between Spring 2018 and 2019: All schools took the test on the paper-based forms in Spring 2019, whereas a majority of schools took the computer-based forms in Spring 2018. The mode comparability study was conducted to minimize the bias introduced by the testing mode so as to achieve valid growth measures between years. The mode comparability study consists of two steps. In the first step, matching was performed to evaluate the size of mode effect. In the second step, adjustment

was conducted to minimize the mode effect. Methods for this analysis were evaluated by the MCAS Technical Advisory Committee and by the Massachusetts DESE. Section 3.6.4 describes the mode comparability study.

3.6.1 IRT

All MCAS items are 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(θ)*] 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(θ)* (Hambleton & van der Linden, 1997; Hambleton & Swaminathan, 1985). The process of determining the mathematical relationship between *θ* and *P(θ)* 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(θ)*. 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 2019 next-generation MCAS 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 represents the scored response on an item,
i indexes the items,
j indexes students,
α represents item discrimination,
b represents item difficulty,
c is the pseudo guessing parameter,
θ is the student proficiency, and
D is a normalizing constant equal to 1.701.

For the 2PL model, this equation 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

IRT calibration was conducted using flexMIRT 3.03 (Cai, 2012). IRT calibration was conducted for the computer-based tests in all grades. Because paper test forms are treated as accommodated forms, item parameters for computer-based items were applied to their paper counterparts. The tables in Appendix J give the IRT item parameters and associated standard errors of all operational scoring items on the 2019 MCAS tests. Appendix K 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 as follows:

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

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.

The number of cycles required for convergence for each grade and content area during the IRT analysis can be found in Table 3-36. The number of cycles required for convergence fell within acceptable ranges (less than 150) for all tests.

Table 3-36. Number of Cycles Required for Convergence

Content Area	*Grade*	*Computer-based Initial Cycles*
ELA	Grade 3	39
	Grade 4	42
	Grade 5	43
	Grade 6	42
	Grade 7	38
	Grade 8	43
	Grade 10	54
Mathematics	Grade 3	60
	Grade 4	69
	Grade 5	48
	Grade 6	57
	Grade 7	86
	Grade 8	51
	Grade 10	78
STE	Grade 5	38
	Grade 8	36

3.6.3 Equating

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

The 2019 administration of the next-generation MCAS used a raw score-to-theta equating procedure in which test forms were equated to the theta scale established on the reference form (i.e., the form used in the most recent standard setting). The groups of students who take equating items on the MCAS tests are never strictly equivalent to the groups who took the tests in the reference years. IRT is particularly useful for equating scenarios that involve nonequivalent groups (Allen & Yen, 1979). Equating for the MCAS uses the anchor test–nonequivalent groups design described by Petersen, Kolen, and Hoover (1989). In this equating design, no assumption is made about the equivalence of the examinee groups taking different test forms (i.e., naturally occurring groups are assumed). Comparability is instead evaluated by using a set of anchor items (also called equating items), assuming they perform in the same way in both groups and can, thus, accurately measure the differences in the two groups.

Item parameter estimates for 2019 next-generation computer-based test forms were placed on the 2019 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 equating items from both tests should have the same item parameters. Thus, prior to implementing this method, three evaluations of the equating items were conducted to check for parameter drift, as follows.

- Delta method: compares two years' delta values (the percent correct transformed into a scale "with an effective range of 6 [very easy item] to 20 [very difficult item]"[7]) for equating items, and flags an item if its standardized distance to the principal axis line is at or above 3 in absolute value.
- *b*-*b* method: compares current year's freely estimated IRT difficulty parameters with the previous year's values for equating items, and flags an item if its standardized distance to the principal axis line is at or above 3 in absolute value.
- IRT curve-based beta method: a measure of the weighted average difference between the item response function (IRF) curves between two years for each equating item (Jiang, Roussos & Yu, 2017; Wang & Roussos, 2018). The current year's IRF is calculated based on transformed item parameters using the SL constants estimated with all equating items. The difference index is denoted as β, its estimate is denoted as $\hat{\beta}$, and the following threshold is used to categorize an item into negligible, moderate, or large drift:
  - $|\hat{\beta}| < 0.05$, negligible drift
  - $0.05 \leq |\hat{\beta}| < 0.1$, moderate drift
  - $|\hat{\beta}| \geq 0.1$, large drift

Items that were flagged as a result of these evaluations are listed in Table 3-37. Detailed results from each drift analysis, along with Delta and *b*-plots are presented in Appendix J.

Following the statistical evaluation, each of these flagged items went through a content review process to further investigate whether there are construct-irrelevant or relevant factors that may have resulted in the item parameter drift. Anything pertaining to the content being measured is considered a construct relevant factor, such as any instructional shift in certain content areas. A list of content irrelevant factors follows:

[7] Walker, M. E. (2014, May 13). Enhancing the Equating of Item Difficulty Metrics: Estimation of Reference Distribution. ETS Research Report Series. P. 1. Retrieved 1.10.20 from: https://onlinelibrary.wiley.com/doi/full/10.1002/ets2.12006

- changes to item administration mode
- word/graphic changes to any part of the item
- change to option order
- change in position (e.g., beginning of test vs end of test)
- whether an item experiences "clueing" in one administration but not in the other
- whether there are test security risks associated with the flagged items
- any other difference that may affect the testing experience

An item is removed from the equating set if a construct irrelevant reason is identified in the content review. If a content relevant reason is identified, an item is kept as an equating item. If the content review does not find any reason, an item is removed if it is flagged by any of these three criteria: (1) standardized distance in the delta plot $\geq$ 3, (2) *b-b* standardized distance in the *b-b* plot $\geq$ 3, and (3) $|\hat{\beta}| \geq 0.1$.

The equating items that were flagged from the delta and *bb* analyses are presented in Table 3-37.

Table 3-37. Year-to-year Equating Items Watch List*

Content Area	*Grade*	*Item ID*	*Statistical Reason*	*Content Reason*	*Action*
ELA	3	IA00458A	beta	No reason identified	Retained
	4	IA00226	beta	No reason identified	Retained
	5	IA00497	b-b	No reason identified	Removed
Mathematics	3	IA00811	beta	No reason identified	Retained
		IA00852	beta	No reason identified	Retained
		IA00924	beta	No reason identified	Retained
		IA00994	beta	No reason identified	Retained
	4	IA00912	beta & b-b	No reason identified	Removed
	5	IA01027	beta	No reason identified	Retained
	6	IA00975	beta	No reason identified	Retained
		IA00782	beta	No reason identified	Retained
	7	IA01017	beta	No reason identified	Retained
		IA01100	beta	No reason identified	Retained
	8	IA01042	beta & b-b	No reason identified	Removed
		IA00897	beta	No reason identified	Retained

** Because this was the first year of the next-generation tests for grade 10 ELA and Math and grades 5 and 8 STE, no equating was conducted for these tests and thus no items were flagged.*

The equating items that successfully survived these evaluation procedures were then employed in the SL method, and the linking relationship obtained from the SL method was used to transform the item parameters for all items in the 2019 next-generation computer-based administration onto the target scale. The transformed item parameters were then used to build the raw score to theta look-up tables for the 2019 tests. The SL constants are presented in Table 3-38.

Table 3-38. Stocking and Lord Constants

Content Area	*Grade*	*Slope*	*Intercept*
ELA	3	1.04	0.26
	4	1.01	0.15
	5	1.06	0.14
	6	1.27	0.12
	7	1.11	-0.02
	8	1.14	0.07
	10	0.98	0.15
Mathematics	3	0.94	0.18
	4	0.96	0.11
	5	1.04	0.18
	6	1.08	0.09
	7	1.04	0.05
	8	1.04	0.26
	10	1.01	0.15

3.6.4 Mode Comparability and Adjustment

As mentioned in section 1.4.2 and in the introduction of section 3.6, there was a major testing mode shift for one school district in Spring 2019. All schools in that district took the test in paper form in Spring 2019. Given that there is a common concern of construct-irrelevant variance due to different testing modes, a mode comparability study was conducted to evaluate and adjust the testing mode bias so as to achieve more accurate year-to-year comparison for that district.

The rationale behind mode comparability evaluation is to compare two equivalent groups' performance on the two testing modes. Specifically, the comparison is made between the paper-tested district and a sample from the computer-tested population that has ability equivalent to the paper-tested district. Given the preexisting ability difference between the computer- 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 testing mode (computer vs. paper) on a number of covariates, including the prior year scaled score[8] and demographic variables (race, gender, EL status, economically disadvantaged status, special education status, years in Massachusetts).

After propensity scores were estimated, nearest neighbor matching with a caliper size of 0.02 was conducted on propensity scores. Given that the paper-tested district consisted of only approximately 1000 students, students from the computer group were selected to match those of the paper group. Specifically, for each student in the paper group, a student from the computer group with the closest propensity score was selected. The resulting matched sample was denoted as the paper-equivalent group.

As a prior score was not available for 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' scores for that school in year 1.

[8] For the scaled scores in the 2017–18 paper tests, mode-adjusted scaled scores were used.

2. Find the grade 3 students' scores 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.

To reduce the sampling error in generating the paper-equivalent group, the matching process was replicated 10 times. In other words, a total of 10 paper-equivalent groups were generated. To ensure matching effectiveness for each replication, the standardized difference between two matched groups on each matching variable was controlled to be smaller than 0.1 in their absolute values (Austin & Mamdani, 2006).

For each pair of matched samples, mode effect size was calculated as the standardized difference between matched samples on students' scaled scores. A nonparametric permutation test was further conducted to evaluate the statistical significance of the differences. Table 3-39 shows the average scale score in each group and the standardized difference between groups before and after matching. The results in Table 3-39 are summarized across replications. The results showed ELA tests in all grades except grade 10 demonstrated small to moderate mode effect favoring paper forms; ELA tended to have a larger effect size than mathematics or STE, and this pattern was similar to that observed in previous years.

Table 3-39. Summary of Mode Effect Before and After Adjustment

Content Area	*Grade*	*Adjustment*	*Average Scale Score*		*Effect Size*		*% of Significant Rejections*[1]
			Paper	*Computer*	*Mean*[1]	*SD*[1]	
ELA	3	Before	499.3	496.2	0.14	0.03	1
		After	496.1	496.2	0.00	0.03	0
	4	Before	498.2	495.6	0.13	0.02	1
		After	495.2	495.6	-0.02	0.02	0
	5	Before	501.4	495.3	0.30	0.02	1
		After	495.1	495.3	-0.01	0.02	0
	6	Before	498.4	492.2	0.27	0.02	1
		After	492.0	492.2	-0.01	0.02	0
	7	Before	500.1	490.6	0.43	0.02	1
		After	490.5	490.6	-0.01	0.02	0
	8	Before	502.7	492.6	0.44	0.03	1
		After	492.6	492.6	0.00	0.04	0
	10	Before	490.2	489.9	0.01	0.01	0
		After	489.3	489.9	-0.02	0.01	0
Mathematics	3	Before	494.0	491.0	0.14	0.03	1
		After	490.8	491.0	-0.01	0.02	0
	4	Before	493.5	491.8	0.08	0.03	0.8
		After	491.6	491.8	-0.01	0.03	0
	5	Before	495.1	492.8	0.13	0.04	0.9
		After	493.0	492.8	0.01	0.04	0
	6	Before	493.1	493.7	-0.03	0.03	0.1
		After	493.9	493.7	0.01	0.03	0
	7	Before	492.1	487.1	0.24	0.02	1
		After	487.6	487.1	0.02	0.02	0

continued

Content Area	Grade	Adjustment	Average Scale Score		Effect Size		% of Significant Rejections[1]
			Paper	Computer	Mean[1]	SD[1]	
Mathematics	8	Before	492.2	490.9	0.06	0.01	0.1
		After	491.2	490.9	0.01	0.01	0
	10	Before	491.6	489.5	0.09	0.02	0.7
		After	490.0	489.5	0.02	0.02	0
STE	5	Before	496.4	494.5	0.08	0.03	0.4
		After	494.3	494.5	-0.01	0.03	0
	8	Before	492.1	495.1	-0.13	0.02	0.9
		After	495.0	495.1	0.00	0.02	0

[1] *Mean and standard deviation of effect size as well as the proportion of significant rejections in permutation tests are calculated across replications.*

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 computer equivalent score was calculated for each θ on the paper scale. Computer equivalent scores were used as the adjusted paper scores. Given a total of 10 pairs of matched samples, the equipercentile linking was conducted for each of them, and this resulted in 10 sets of adjusted paper scores. Then, kernel smoothing (Nadaraya,1964; Watson, 1964) was implemented to smooth the adjusted scores by taking weighted averages. The smoothed scores were used in the final adjusted lookup table. The adjusted look-up tables are presented in Appendix L. Mode effect analysis was conducted again—this time between the computer scores and the adjusted paper scores, as summarized in Table 3-39. The nonsignificant mode difference in each test suggested the mode adjustment was effective.

3.6.5 Achievement Standards

Cutpoints for the next-generation MCAS tests were set via standard setting in 2017 for grades 3–8 ELA and mathematics tests, and in 2019 for grade 10 ELA and mathematics tests and grade 5 and 8 STE tests (see Appendix M for the 2019 standard-setting report and the *2017 Next-Generation MCAS and MCAS-Alt Technical Report* for the 2017 standard-setting report). The standard setting establishes the theta cutpoints used for reporting each year. These theta cuts are presented in Table 3-40. 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.

Table 3-40. Cut Scores on the Theta Metric and Reporting Scale by Content Area and Grade

Content Area	Grade	Theta			Scale Score				
		Cut 1	Cut 2	Cut 3	Min	Cut 1	Cut 2	Cut 3	Max
ELA	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
	10	-1.728	-0.299	1.130	440	470	500	530	560

continued

Content Area	Grade	Theta			Scale Score				
		Cut 1	Cut 2	Cut 3	Min	Cut 1	Cut 2	Cut 3	Max
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
	10	-1.721	-0.317	1.087	440	470	500	530	560
STE	5	-1.621	-0.112	1.398	440	470	500	530	560
	8	-1.499	-0.020	1.459	440	470	500	530	560

3.6.6 Reported Scale Scores

Because the θ scale used in IRT calibrations is not understood by most stakeholders, reporting scales were developed for the 2019 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 scaled scores preserve the interval scale properties (except for the truncated scaled 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 2019 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 2019 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. Scale scores are calculated using the following 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 3-41 shows the slope and intercept terms used to calculate the scaled scores for each grade and content area. Note that the values in Table 3-41 will not change unless the standards are reset.

Appendix N contains raw-score-to-scale-score look-up tables for computer-based testing forms.[9] The tables show the scaled score equivalent of each raw score for the 2018 next-generation MCAS tests. Additionally, Appendix N contains scaled score distribution graphs for each grade and content area for computer-based testing forms. These distributions were calculated using the data matrix files that were used in the IRT calibrations.

Table 3-41. Scale 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
	10	20.995	506.274
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
	10	21.373	506.775
STE	5	19.875	502.220
	8	20.287	500.409

3.7 MCAS Reliability

Although an individual item's performance is an important factor in evaluating an assessment, a complete evaluation must also address the way items grouped in a set function together and complement one another. Tests that function well provide a dependable assessment of a student's level of ability. Just like the measurement of physical properties, such as temperature, any measurement tool contains some amount of measurement error, which leads to different results if the measurements were taken multiple times. The quality of items, as the tools to measure the latent ability, determines the degree to which a given student's score can be higher or lower than his or her true ability on a test.

There are a number of ways to estimate an assessment's reliability. The approach that was implemented to assess the reliability of the 2019 next-generation MCAS tests is the α coefficient of Cronbach (1951). This approach is

[9] The raw-score-to-scale-score lookup tables and the scaled score distribution graphs for paper testing forms after mode adjustment are in Appendix L, "Mode Adjustment Lookup Tables."

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 3-42 presents descriptive statistics, Cronbach's α coefficient, and raw score SEMs for each content area and grade. Statistics are based on operational items only. The reliability estimates range from 0.86 to 0.94, which are generally in acceptable ranges.

Table 3-42. Raw Score Descriptive Statistics, Cronbach's Alpha, and SEMs by Content Area and Grade—Computer-based

Content Area	*Grade*	*Number of Students*	*Raw Score*			*Alpha*	*SEM*
			Maximum	*Mean*	*Standard Deviation*		
ELA	3	66,906	44	25	7.59	0.86	2.84
	4	68,811	44	26	8.29	0.89	2.80
	5	71,020	48	28	8.94	0.88	3.09
	6	71,289	51	28	10.47	0.90	3.26
	7	70,150	51	28	10.22	0.90	3.29
	8	69,820	51	29	10.33	0.90	3.24
	10	69,411	51	38	9.23	0.90	2.88
Mathematics	3	66,993	48	28	11.19	0.93	3.03
	4	68,793	54	33	11.78	0.92	3.25
	5	71,017	54	30	12.28	0.91	3.59
	6	71,261	54	30	12.95	0.93	3.50
	7	70,125	54	25	13.23	0.93	3.47
	8	69,799	54	30	12.93	0.93	3.39
	10	69,477	60	35	14.57	0.94	3.57
STE	5	70,997	54	31	10.05	0.90	3.25
	8	69,605	54	29	10.30	0.90	3.24

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 2019 next-generation MCAS tests. Appendix O 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 0.79 to 0.94 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 O 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 the 2019 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 O. The reliability coefficients for the reporting subcategories range from 0.45 to 0.87, with a median of 0.71 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 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 2019 next-generation MCAS tests, students were classified into one of four achievement levels: *Not Meeting Expectations*, *Partially Meeting Expectations*, *Meeting Expectations*, or *Exceeding Expectations*. Appendix P shows achievement-level distributions by content area and grade for the 2019 next-generation MCAS tests.

Cognia 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 2019 next-generation MCAS tests because it is easily adaptable to all types of testing formats, including mixed formats.

The DAC estimates reported in Tables 3-43 and 3-44 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 2019 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.

Cognia also measured consistency on the 2019 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. Results of the DAC analyses are provided in Table 3-43 and 3-44. 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 (0.79–0.85), consistency (0.70–0.79), and kappa (0.54–0.67) indicate that the vast majority of students were classified accurately and consistently with respect to measurement error and chance.

In addition to overall accuracy and consistency indices, accuracy and consistency values conditional on achievement level are also given. For the calculation of these conditional indices, the denominator is the proportion of students associated with a given achievement level. For example, from Table 3-43, the conditional accuracy value is 0.77 for *Not Meeting Expectations* for the grade 3 ELA computer-based form. This figure indicates that among the students whose true scores placed them in this classification, 77% would be expected to be in this classification when categorized according to their observed scores. Similarly, a consistency value of 0.70 indicates that 70% 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 about 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, since it summarizes the percentage of students who are correctly classified either above or below the particular cutpoint. Table 3-44 provides 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 Table 3-44 range from 0.88–0.94 and 0.83–0.91, respectively. The false positive and false negative decision rates at the *Partially Meeting Expectations/Meeting Expectations* threshold range from 3% to 6%. 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 3-43. Summary of Decision Accuracy and Consistency Results by Content Area and Grade—Overall and Conditional on Achievement Level

Content Area	*Grade*	*Overall*	*Kappa*	*Conditional on Achievement Level*			
				Not Meeting Expectations	*Partially Meeting Expectations*	*Meeting Expectations*	*Exceeding Expectations*
ELA	3	0.79 (0.70)	0.54	0.77 (0.56)	0.80 (0.74)	0.78 (0.71)	0.79 (0.63)
	4	0.81 (0.74)	0.59	0.79 (0.62)	0.84 (0.78)	0.79 (0.73)	0.81 (0.66)
	5	0.81 (0.74)	0.58	0.78 (0.59)	0.83 (0.77)	0.81 (0.75)	0.79 (0.61)
	6	0.80 (0.72)	0.59	0.82 (0.68)	0.81 (0.74)	0.78 (0.71)	0.83 (0.71)
	7	0.80 (0.72)	0.58	0.84 (0.72)	0.82 (0.76)	0.79 (0.73)	0.70 (0.52)
	8	0.80 (0.71)	0.57	0.84 (0.72)	0.80 (0.73)	0.79 (0.72)	0.76 (0.61)
	10	0.81 (0.73)	0.59	0.80 (0.64)	0.84 (0.78)	0.77 (0.70)	0.84 (0.72)
Mathematics	3	0.84 (0.78)	0.66	0.83 (0.70)	0.85 (0.80)	0.84 (0.79)	0.84 (0.72)
	4	0.84 (0.77)	0.64	0.83 (0.70)	0.85 (0.79)	0.84 (0.79)	0.79 (0.64)
	5	0.85 (0.78)	0.64	0.78 (0.61)	0.84 (0.79)	0.86 (0.82)	0.79 (0.61)
	6	0.85 (0.79)	0.67	0.83 (0.69)	0.84 (0.79)	0.86 (0.82)	0.83 (0.71)
	7	0.82 (0.75)	0.63	0.68 (0.52)	0.81 (0.76)	0.85 (0.80)	0.85 (0.75)
	8	0.84 (0.77)	0.65	0.79 (0.65)	0.85 (0.80)	0.84 (0.79)	0.82 (0.70)
	10	0.85 (0.78)	0.67	0.76 (0.61)	0.84 (0.79)	0.87 (0.82)	0.83 (0.73)
STE	5	0.82 (0.75)	0.61	0.80 (0.65)	0.83 (0.77)	0.82 (0.76)	0.82 (0.68)
	8	0.82 (0.74)	0.60	0.79 (0.64)	0.83 (0.77)	0.81 (0.76)	0.78 (0.63)

Table 3-44. Summary of Decision Accuracy and Consistency Results by Content Area and Grade—Conditional on Cutpoint

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*	
			Pos	*Neg*		*Pos*	*Neg*		*Pos*	*Neg*
ELA	3	0.97 (0.96)	0.01	0.02	0.88 (0.83)	0.06	0.06	0.94 (0.91)	0.04	0.02
	4	0.97 (0.96)	0.01	0.02	0.90 (0.86)	0.05	0.05	0.95 (0.92)	0.04	0.02
	5	0.97 (0.96)	0.01	0.02	0.89 (0.85)	0.05	0.05	0.95 (0.93)	0.03	0.01
	6	0.96 (0.94)	0.01	0.03	0.90 (0.86)	0.05	0.05	0.94 (0.92)	0.04	0.02
	7	0.96 (0.94)	0.02	0.03	0.90 (0.87)	0.05	0.05	0.94 (0.91)	0.04	0.02
	8	0.96 (0.94)	0.02	0.03	0.91 (0.87)	0.05	0.05	0.93 (0.90)	0.04	0.03
	10	0.98 (0.97)	0.01	0.01	0.91 (0.87)	0.05	0.05	0.92 (0.89)	0.05	0.03
Mathematics	3	0.97 (0.96)	0.01	0.02	0.92 (0.89)	0.04	0.04	0.95 (0.93)	0.03	0.02
	4	0.97 (0.96)	0.01	0.02	0.92 (0.88)	0.04	0.04	0.95 (0.93)	0.03	0.02
	5	0.97 (0.96)	0.01	0.02	0.91 (0.87)	0.04	0.05	0.96 (0.95)	0.02	0.01
	6	0.97 (0.96)	0.01	0.02	0.92 (0.89)	0.04	0.04	0.96 (0.94)	0.03	0.02
	7	0.94 (0.92)	0.02	0.04	0.93 (0.89)	0.04	0.04	0.96 (0.94)	0.02	0.02
	8	0.96 (0.94)	0.01	0.02	0.92 (0.89)	0.04	0.04	0.95 (0.94)	0.03	0.02
	10	0.97 (0.96)	0.01	0.02	0.94 (0.91)	0.03	0.03	0.94 (0.92)	0.03	0.02
STE	5	0.97 (0.96)	0.01	0.02	0.90 (0.86)	0.05	0.05	0.95 (0.93)	0.03	0.02
	8	0.96 (0.94)	0.01	0.03	0.91 (0.87)	0.05	0.05	0.95 (0.93)	0.03	0.02

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

As with other methods of evaluating reliability, DAC statistics that are calculated based on groups with smaller variability can be expected to be lower than those calculated based on groups with larger variability. For this reason, the values presented in Tables 3-43 and 3-44 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.* (This language is different than that used for the legacy tests.) New in 2019, grade 10 ELA and mathematics and grades 5 and 8 STE were reported using next-generation

standards and achievement levels. Students were given a separate achievement-level classification in each content area.

*Parent/Guardian Report*s and student results labels are the only printed reports; they were mailed to districts for distribution to parents/guardians and schools. See section 3.8.1 for additional details of the *Parent/Guardian Report.*

The DESE 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 *High School Parent/Guardian Report* used to report ELA and mathematics results was redesigned to be in line with the 3-8 next-generation reports. The next-generation *Parent/Guardian Report* is available online to schools via PearsonAccess Next (PAN). The *Parent/Guardian Report* was generated for each student eligible to take the MCAS tests. It is a stand-alone single page (11" x 17") color report that is folded. Two full-color copies of each student's report were printed: one for the parent/guardian and one for the school's records. A sample report is provided in Appendix Q.

The report is designed to present parents/guardians with a detailed summary of their child's MCAS performance and to enable comparisons with other students at the school, district, and state levels. The DESE 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 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, range of scores, and growth percentile for students who met "attemptedness" for the content area. A student meets attemptedness if they have attempted at least one common item in each session. Otherwise, it provides the Not Tested reason applicable to the content area. At the High School level, it also provides a note that informs parents/guardians and students as to whether the student has met, has not met, or previously met the graduation requirement for each content area.

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 (described as the low end of the Meeting Expectation achievement level category) 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 DESE's website. Students who tested only in ELA and mathematics received a report with a back page that provides information about the aMAzing Educators program.

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 non-calculator session of the mathematics test.

3.8.2 Student Results Label

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's label was shipped with the *Parent/Guardian Report*s.

3.8.3 Analysis and Reporting Business Requirements

To ensure that MCAS results are processed and reported accurately, the documents detailing analysis and reporting business requirements and data processing specifications are updated to reflect any changes/additions necessary for reporting each year. The processing and analysis and reporting business requirements are observed in the analyses of the MCAS test data and in reporting results. These requirements also guide data analysts in identifying which students will be excluded from school-, district-, and state-level summary computations. A copy of the *Analysis and Reporting Business Requirements* document for the 2019 next-generation MCAS administration is included in Appendix R.

3.8.4 Quality Assurance

Quality-assurance measures are implemented throughout the process of analysis and reporting at Cognia. 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 DESE, reports were run using demonstration data to test the application of the analysis and reporting business requirements. The populated reports were then approved by the DESE.

Another type of quality-assurance measure used at Cognia 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 analysis and reporting business

requirements 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, and
- 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 DESE 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 establishment of intended uses and interpretations of test results—and conformance 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, these sections provide 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 validity is rooted in the item development process, including how the test blueprints and test items align to the curriculum and standards. All items are developed, edited, administered, reviewed, and scored to represent the expectations from the state curriculum frameworks. This process is described further 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.

A 2017 content alignment study on the next-generation MCAS tests, conducted by Boston College researchers under the leadership of Michael Russell (See Appendix S), found a high degree of content alignment. For mathematics, over 90% of the domains assessed across the grade level tests showed high levels of alignment. For ELA, alignment was also found to be strong across grade levels and domains. When both the items and essay scoring criteria were considered, over 95% of the alignment considerations were deemed adequate. Only two domains, Grade 7 and Grade 8 Reading Informational Text, were identified as candidates for improved alignment. In addition, analyses of the level of agreement among panel members' ratings showed high levels of agreement for the vast majority of ratings following the consensus process. While the study found a few select opportunities to improve alignment, the results from the analyses provide evidence of strong alignment across the vast majority of the tests examined.

3.9.2 Response Process Validity Evidence

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 DESE ensures that evidence of response process validity is collected and reported for all new MCAS item types used in the next-generation assessments.

DESE conducted a 2019 study to determine the readiness of grade 10 students and educators in Massachusetts schools to respond to the next-generation MCAS items. Two standalone field tests were administered to students in every high school in the state. Data from these standalone field tests were then analyzed to determine the following:

- The psychometric properties of the test items and the field tests
- The response time students took to successfully respond to the test

Student response time data was used to filter out the results of students who did not spend sufficient time on their answers. The data from the remaining motivated students was used to examine item discrimination and ensure that new scoring rubrics were keyed correctly. Next-generation test forms were then developed from these sampled results.

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.

Furthermore, to evaluate whether different reporting categories constitute statistically different dimensions, item-level confirmatory factor analysis (CFA) was conducted to assess the internal structure of the MCAS ELA and Mathematics assessments in grade 10 from the School Year 18-19. The CFA model for each test was specified such that the number of factors equaled the number of reporting categories and each item loaded onto the factor that corresponded to the reporting category to which the given item contributed. The results showed very high correlations between different factors, suggesting that there is very little unique variance among the given set of reporting categories. In other words, different reporting categories are essentially measuring the same thing. These results are highly consistent with the unidimensionality results from the DIMTEST and DETECT analyses, as well as the previous CFA analyses conducted on MCAS ELA and Mathematics assessments in grade 3-8 from the School Year 17-18. The full CFA report is included in Appendix T. Although the CFA analysis suggested unidimensionality among different reporting categories, the high and positive factor loadings do suggest the items

provide good measurement for each reporting category. Unidimensionality, meaning items from one reporting category correlate highly to other reporting categories, can be evidence that students have learned different content areas within each subject in an integrated fashion.

3.9.4 Validity Evidence in Relationship to Other Variables

DESE continues 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. In 2019, DESE conducted concurrent validity studies. The first compared student results on the next-generation MCAS tests to course grades and course taking in middle and high school. Specifically, the relationships among MCAS results and student course grades in the respective subjects (in ELA and Mathematics) showed that MCAS results were more strongly associated with course grades than other covariates tested, including course level, economic disadvantage, being on an IEP, or being an English learner. In Mathematics in grades 8 and 10, MCAS achievement levels in math were significantly associated with taking advanced mathematics courses. Convergent validity evidence was also reported between MCAS test portions and subjects. These analyses are shown in Appendix U.

In 2019, DESE conducted a study examining predictive validity of grade 8 MCAS results on grade 9 course-taking patterns and GPAs. Results from this study will be published as a white paper on the DESE website at www.doe.mass.edu/mcas/tech/.

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

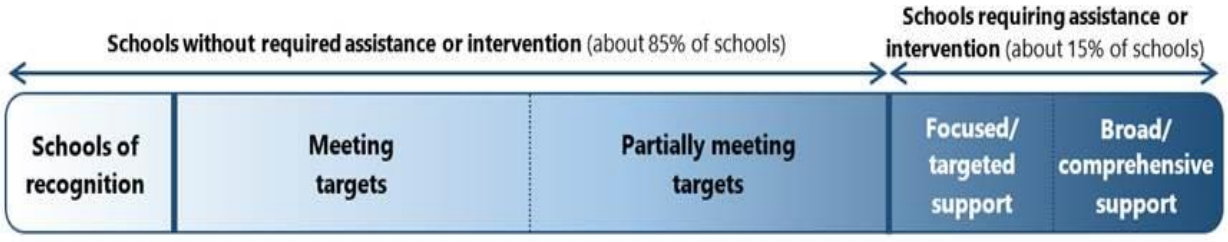
The DESE takes many steps to support the intended uses of MCAS data. (The intended uses are listed in section 2.3 of this report.) This section will examine some of the reporting systems and policies designed to address each use.

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

In 2018, DESE updated its accountability plan to conform to state and federal requirements. Measures of student achievement and growth are prominently featured alongside other indicators in the new school and district accountability system. Each school's performance on all measures is compared to its targets and to the performance of other schools in the state. The system includes incentives designed to focus schools on their lowest-performing students from prior years.

In the system, schools are placed into categories that describe their performance relative to state goals. As shown in Figure 3-1, the categories reflect how much assistance or intervention each school requires under the system. School and district accountability report cards are publicly available at www.doe.mass.edu/accountability/report-cards/.

Figure 3-1. School Categories in Massachusetts Accountability System



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 Report*s. The current versions of these reports (see the samples provided in Appendix Q) 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 DESE website.

The DESE'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 DESE 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 2019, *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. Example of School Results by Standards Report—Mathematics, Grade 7

All Students Students (161)

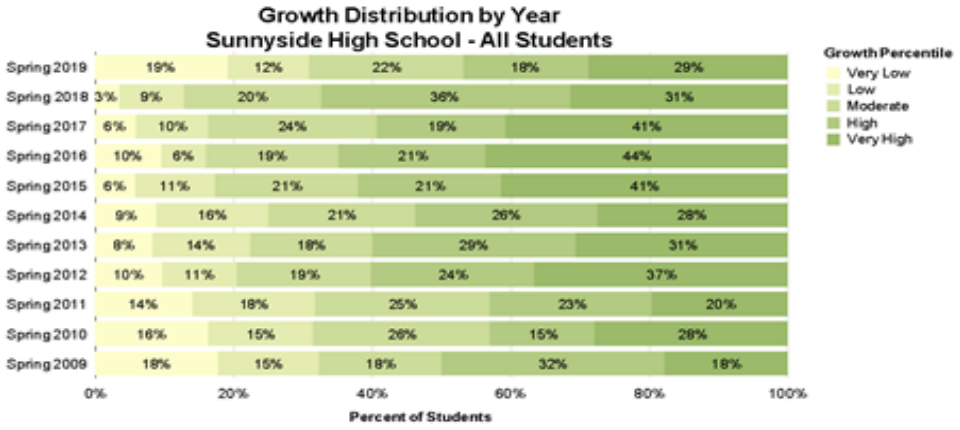
Standards: MA 2017 Standards **Show results with <10 students :** No

	Possible Points	School % Possible Points	District % Possible Points	State % Possible Points	School/ State Diff
Mathematics					
All items	54	48%	48%	47%	1
Question Type					
Constructed Response	16	48%	49%	48%	1
Short Answer	14	41%	42%	39%	2
Selected Response	24	52%	51%	51%	1
Domain / Cluster					
Expressions and Equations	**14**	**47%**	**48%**	**47%**	**-1**
Solve real-life and mathematical problems using numerical and algebraic expressions and equations.	10	54%	54%	52%	2
Use properties of operations to generate equivalent expressions.	4	28%	31%	36%	-8
Geometry	**8**	**42%**	**43%**	**44%**	**-2**
Draw	2	39%	44%	47%	-9
Solve real-life and mathematical problems involving angle measure	6	43%	43%	43%	0
Ratios and Proportional Relationships	**11**	**55%**	**54%**	**53%**	**2**
Analyze proportional relationships and use them to solve real-world and mathematical problems.	11	55%	54%	53%	2
Statistics and Probability	**11**	**36%**	**36%**	**37%**	**0**
Draw informal comparative inferences about two populations.	3	29%	30%	32%	-2
Investigate chance processes and develop	6	36%	35%	36%	0
Use random sampling to draw inferences about a population.	2	48%	45%	47%	2
The Number System	**10**	**62%**	**59%**	**54%**	**8**
Apply and extend previous understandings of operations with fractions to add	10	62%	59%	54%	8

NOTE: MCAS results are suppressed for group counts of less than 10.

School results only include students enrolled in the school since Oct.1.

Figure 3-3. Example of Growth Distribution Report—ELA, Grade 10



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	Mean SGP	# Students (SGP)	% Proficient or Higher	% Meeting or Exceeding Expectations	# Students (Ach. Level)
Spring 2019	18	11	21	17	27	58.5	55.0	94		65	95
Spring 2018	3	8	17	31	27		66.6	86	99		88
Spring 2017	5	9	21	16	35	70.5		86	100		86
Spring 2016	9	6	18	20	41	71.5		94	99		94
Spring 2015	5	10	18	18	36	73.0		87	99		87
Spring 2014	7	13	17	21	22	65.0		80	95		83
Spring 2013	7	12	15	25	26	67.0		85	100		86
Spring 2012	9	10	18	22	34	70.0		93	98		93
Spring 2011	15	19	27	25	21	53.0		107	85		109
Spring 2010	14	13	22	13	24	51.0		86	74		87
Spring 2009	11	9	11	20	11	60.5		62	88		69
Spring 2008									81		43

Aggregate student growth percentile (SGP) is not calculated if the number of students with SGP is less than 20.

The assessment data in Edwin Analytics are also available on the DESE 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 DESE'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 DESE'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 MCAS-Alt 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 assessments. 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 cognitive 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 assessments 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 cognitive 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 cognitive 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 cognitive 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 initial implementation in 2001.

4.1.3 Format

The MCAS-Alt consists of a portfolio containing a structured set of "evidence" collected during instructional activities in each subject to be assessed 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, all of 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.

DESE's *Resource Guide to the Massachusetts Curriculum Frameworks for Students with Disabilities* (the *Resource Guide*) describes the content to be assessed by the 2019 MCAS-Alt, and contains the 2017 English language arts (ELA) standards, the 2017 mathematics standards, and the 2016 science and technology/engineering (STE) standards.

The *Resource Guide* also provides strategies for adapting and using the state's learning standards to instruct and assess students taking the MCAS-Alt. The fall 2018 *Resource Guide* is intended to ensure that all students receive instruction in the Massachusetts Curriculum Frameworks in ELA, mathematics, and STE at levels that are challenging and attainable for each student. For the MCAS-Alt, students are expected to achieve the same standards as their peers without disabilities. 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 MCAS-Alt 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. In the MCAS-Alt, the range and level of complexity of the standards being assessed have been modified, yet without altering the essential components or meaning of the standards. The MCAS-Alt content areas and strands/domains required for the assessment of students in each grade level are listed in Table 4-1.

Table 4-1. 2019 MCAS-Alt: Requirements

Grade	*ELA Strands Required*	*Mathematics Domains 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	For any three of the four STE disciplines:* select one core idea in each discipline; assess six entry points within each core idea.
6	▪ Language ▪ Reading ▪ Writing	▪ Ratios and Proportional Relationships ▪ The Number System	
7	▪ Language ▪ Reading ▪ Writing	▪ Ratios and Proportional Relationships ▪ Geometry	

continued

Grade	*ELA Strands Required*	*Mathematics Domains Required*	*STE Strands Required*
8	▪ Language ▪ Reading ▪ Writing	▪ Expressions and Equations ▪ Geometry	For any three of the four STE disciplines:* select one core idea in each discipline; assess six entry points within each core idea.
10	▪ Language ▪ Reading ▪ Writing	Any three of the five mathematics *conceptual categories*: ▪ Functions ▪ Geometry ▪ Statistics and Probability ▪ Number and Quantity ▪ Algebra	Select three standards in one of the following disciplines: ▪ Biology ▪ Chemistry ▪ Introductory Physics or ▪ Technology/Engineering

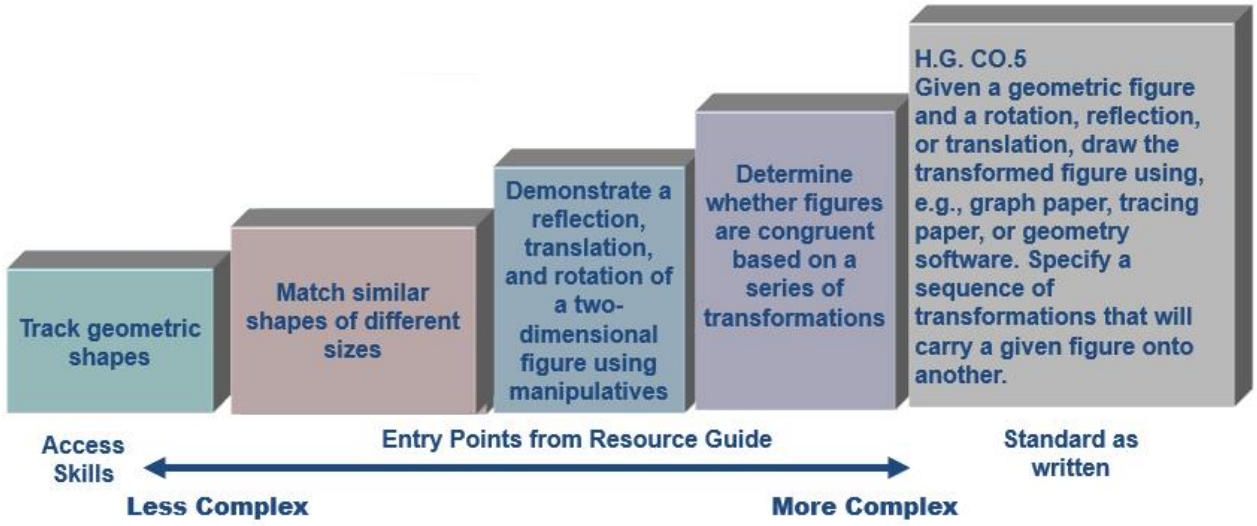
** Earth and Space Science, Life Science, Physical Sciences, Technology/Engineering*

4.2.1.1 ACCESS TO THE GRADE-LEVEL CURRICULUM

The Fall 2018 *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 cognitive disabilities can access the *essence* (i.e., key concepts, ideas, and core knowledge) of each learning standard by addressing one of several entry points listed in the *Resource Guide*. Entry points are academic 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. 2019 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 assessments for ELA–Language, ELA–Reading, Mathematics, and high school STE consist of a collection of primary evidence, supporting documentation, and other required information.

Primary Evidence

For the content areas listed above, the portfolios must include three or more pieces of primary evidence in each strand being assessed.

One of the three pieces of evidence 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 and next-generation STE for grades 5 and 8.

The MCAS-Alt assessment for ELA–Writing consists of one "baseline" writing sample (not included in the student's score), plus three final writing samples in any of three writing types, generated using the student's primary mode of communication (included in the final score).

The MCAS-Alt assessment for STE in grades 5 and 8 consists of primary evidence in three STE disciplines. Each discipline includes evidence of six entry points within one core idea.

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* (see Appendix V), 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** and alignment with learning standards in the Massachusetts curriculum frameworks in the content area being assessed
- **accuracy** of the student's responses to questions 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** demonstrating 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. Rubric dimensions and possible scores are as follows

- 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" would signify 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 to yield a strand subscore; those subscores are combined, as shown in Appendix W, to yield an overall score in the content area. Students taking alternate assessments based on alternate academic achievement standards (AA-AAAS) receive scores of either *Progressing*, *Emerging*, or *Awareness.*

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 2019 are provided in section 4.3.3.3.

The participation guidelines section of the *2019 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 2019 MCAS-Alt grade-level and competency portfolios were evaluated, see section 4.4, MCAS-Alt Scoring.

Competency Portfolios

All high school students, including students with disabilities, are required to meet the Competency Determination (CD) standard to be eligible to earn a high school diploma. Students must attain a score of either *Proficient* (legacy MCAS tests/retests) or *Meeting Expectations* (next-generation MCAS tests/retests) or higher on the MCAS ELA and mathematics tests (or a score of *Needs Improvement* [legacy] or *Partially Meeting Expectations* [next-generation], 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 those 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 have met the appropriate standard of performance in that subject. Because students with significant cognitive disabilities comprise the majority of students taking alternate assessments, the proportion of students who achieve scores of *Needs Improvement/Partially Meeting Expectations* 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* or higher on the assessment.

4.2.2 Test Development

4.2.2.1 RATIONALE

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

The Elementary and Secondary Education Act (ESEA) requires states to include an alternate assessment option for students with significant cognitive disabilities. This requirement ensures that students with significant cognitive 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 Department, 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 achievement.

4.3 MCAS-Alt Test Administration

4.3.1 Evidence Collection

Evidence for English Language Arts (except ELA–Writing), Mathematics, and Grades 5 and 8 Science and Technology/Engineering

Each portfolio strand 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 from 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 to assess the student. The data chart 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 evidence 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 on the data chart.

Evidence for 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:

- Opinion (grades 3–5)/Argument (grades 6–8 and 10)
- Informative/Explanatory text
- Narrative, including Poetry

In addition to three writing samples, one *baseline sample* must be submitted, and may include an outline, completed graphic organizer, or a draft of a writing assignment. The baseline sample should provide information to guide additional instruction in writing in that text type.

Evidence for Next-Generation STE Strands (Grade 5 and 8)

The format described below is intended to encourage the teaching of a unit of science instruction based on a core idea. Teachers are directed to follow these steps:

Step 1: Select three (3) of the following STE disciplines:

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

Step 2: Select a core idea within the chosen discipline that is both relevant and that engages and challenges the student.

Step 3: Select at least six (6) different entry points within one core idea. At least three (3) different science practices must be addressed within the six selected entry points. This step encourages teachers to design related activities that address a theme or unit of study.

Step 4: List the following information on each STE Activity Summary Sheet:

- student's name
- core idea
- entry point addressed during the activity
- numbered science practice for that entry point
- accuracy and independence for each task or response, and the summary percent
- date
- description of the activity

Step 5: Select <u>three</u> work samples to include in the portfolio strand that clearly show the final product of instruction. Each sample should represent a different science practice. (Note: Photographs and/or videos may be submitted as primary evidence if they are labeled and clearly show the final product of instruction.)

4.3.2 Construction of Portfolios

The student's MCAS-Alt portfolio must include all elements listed below. Required forms may either be photocopied from those found in the *2019 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 for a particular outcome.
- **Work sample description** attached to each piece of primary evidence, providing required labeling information. (If work sample description labels 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 4.3.1).
- **STE Summary Sheet (Next-Gen STE only)** completed by the teacher submitting the portfolio (as detailed in section 4.3.1).

The contents listed above, plus all evidence and other documentation, are placed inside a three-ring plastic binder 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 Department's Student Information Management System (SIMS). Students with significant cognitive 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 complexity or severity of his or her cognitive 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 to consider for students taking 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 in order for the student to participate?
- Does the student require an alternate assessment? (Alternate assessments are intended for a very small number of students with significant disabilities who are unable to take standard MCAS tests, even with accommodations.)

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. Participation Guidelines

OPTION 1

Characteristics of Student's Instructional Program and Local Assessment	Recommended Participation in MCAS
If the student is 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,	*Then* the student should take the **standard MCAS test**, either under routine conditions or with accommodations and accessibility features that are generally consistent with the instructional accommodation(s) used in the student's educational program (according to accessibility and accommodations policies available at: 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 has a significant cognitive disability and is 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,	*Then* 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 a) working on learning standards at or near grade-level expectations; ***and*** is b) **sometimes able** to take a paper-and-pencil test, either without accommodations or with one or more accommodation(s); ***but*** c) has a complex and significant disability that does not allow the student to fully demonstrate knowledge and skills on a test of this format and duration, (Examples of complex and significant disabilities for which the student may require an alternate assessment are provided below.)	*Then* the student should take the **standard MCAS test**, if possible, with necessary accommodations that are generally consistent with the student's instructional accommodation(s) and the Department's accessibility and accommodations policies. Accommodations must be 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.

Although 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 complexity and severity of a student's disability present significant barriers or challenges to standardized testing, even with the use of accommodations, but the student is working at or near grade-level expectations, 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 (high school) portfolio option.

According to the criteria outlined in Option 3 above, the following are examples of unique circumstances that could warrant the use of either the MCAS-Alt grade-level portfolio or the MCAS-Alt competency portfolio.

- A student with a severe emotional, behavioral, or attentional 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 7,453 students, or 1.4 percent of students who took standard MCAS assessments, participated in the 2019 MCAS-Alt in one or more subjects in grades 3–10. In ELA, 6,969 students took the MCAS-Alt (1.4 percent); in mathematics, 7,021 students took the MCAS-Alt (1.4 percent); and in STE, 2,850 students took the MCAS-Alt (1.3 percent). This represents an overall decrease of 0.1 percent from 2018, or 148 students. In ELA, 300 fewer students took the MCAS-Alt; in mathematics, 360 fewer students took the MCAS-Alt; and in STE, the number remained the same as in 2018.

Additional information about MCAS-Alt participation rates is provided in the 2019 MCAS-Alt State Summary, including the comparative rate of participation in each MCAS assessment format (i.e., routinely tested, tested with accommodations, or alternately assessed), available at: www.doe.mass.edu/mcas/alt/results.html.

4.3.4 Educator Training

During October 2018, a total of 2,395 educators and administrators received training on conducting the 2019 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, or an overview for school and district administrators.

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, and
- collecting data on student performance and progress based on measurable outcomes.

Topics for the update session included the following:

- a summary of the statewide 2018 MCAS-Alt results,
- changes to the MCAS-Alt requirements for 2019,
- avoiding mistakes that lead to an achievement level of *Incomplete*,
- new requirements for next-generation Science and Technology/Engineering for grades 5 and 8,
- competency and grade-level portfolio requirements, and
- improving the process of selecting challenging entry points for assessment.

Topics for the administrator's overview 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, and
- changes regarding the percentage of students who may be assessed through an alternate assessment based on alternate academic achievement standards.

In January–March 2019, a total of 1,890 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).

4.3.5 Support for Educators

A total of 78 MCAS-Alt training specialists were trained by DESE to provide assistance and support for teachers conducting the MCAS-Alt in their districts, as well as to assist DESE at eight Department-sponsored portfolio review training sessions in January–March 2019. In addition, DESE 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 Cognia 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 to train Service Center staff in relevant areas such as web support, enrollment inquiries, and discrepancy follow-up and resolution procedures.

4.4 MCAS-Alt 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, considering 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 criteria outlined in the *2019 Guidelines for Scoring MCAS-Alt Portfolios,* available in Appendix V. 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 a statewide advisory committee.

4.4.1 Scoring Logistics

MCAS-Alt assessments were scored in Dover, New Hampshire, during April and May 2019. DESE and Cognia 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 in section 4.4. Scores were entered into a computer-based scoring system designed by Cognia and DESE, and scores were frequently monitored for accuracy and completeness.

Security was maintained at the scoring site by restricting access to unscored portfolios to DESE and Cognia 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 were 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 DESE and Cognia staff plus four teacher consultants, met daily over the course of scoring in 2019 and periodically throughout the 2018–2019 school year to accomplish the following:

- nominate prospective scorers and scoring leaders for the 2019 scoring institute;
- select sample portfolio strands to use to train, calibrate, and qualify scorers in 2019; and
- discuss recurring issues and themes to be addressed during the following fall educator training sessions.

All sample strands were scored using the 2019 scoring guidelines, noting any scoring problems that arose during the review. Concerns were resolved by using the *2019 Educator's Manual for MCAS-Alt* and by following additional scoring rules agreed upon by the PLT and subsequently addressed in the final 2019 scoring guidelines.

Of the portfolios reviewed the previous year, several sample strands were set aside as possible exemplars to train, qualify, 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 and other agencies, Cognia 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 DESE and Cognia. Twelve 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 OF SCORERS

Before scoring actual student portfolios, each potential scorer was required to take a qualifying assessment consisting of eight sample portfolio strands that contained a total of 213 score points. The threshold percentage for qualification on the 213 available score points was 85% (181 correct out of 213).

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 DESE using the same methods and criteria used to qualify scorers, except that they were required to achieve a score of 90% correct or higher on the qualifying test.

4.4.3 Scoring Methodology

4.4.3.1 ENGLISH LANGUAGE ARTS (EXCEPT ELA–WRITING), MATHEMATICS, AND LEGACY SCIENCE AND TECHNOLOGY/ENGINEERING

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, DESE staff members were available to provide clarification.

Scorers were randomly assigned a portfolio by their TL. Scorers were required to ensure that the required strands for each grade were submitted, and then to determine if each submitted strand was complete. 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 2019 MCAS-Alt score distributions for all scoring dimensions are provided in Appendix H.

During spring 2019, scorers used an automated, customized scoring program called *AltScore* to score MCAS-Alt assessments. 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 to 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. 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

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, and
- 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 *2019 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 had already 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

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.

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 listed on 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. 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 reflects 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 listed 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 percentage of independence for:

- all data points during the final 1/3 time frame listed on 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 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.

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 had already 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. 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. 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. 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 pre-score each of their student's three final writing samples using the state-provided rubric in Appendix W, according to the appropriate text type:

- Opinions/Arguments
- Informative/Explanatory texts
- Narrative
- Poetry

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

Writing samples were to be produced as independently as possible by the student. If teachers provided text for the student or applied their own revisions to the student's work, this must have been reflected in the score, particularly in the rubric area of Independence. Teachers were expected to explain how edits and revisions were made and indicate the student's contribution to the creation of the sample.

Writing samples were produced using the student's primary mode of communication; for example, dictated to a scribe, with the scribe assuming the use of capital letters and basic punctuation. Teachers were permitted to submit a student's constructed-response to reading comprehension questions or other topics as the basis for their writing samples, even if those responses were already included in the evidence compiled for another portfolio strand.

4.4.3.3 NEXT-GENERATION SCIENCE AND TECHNOLOGY/ENGINEERING (GRADES 5 AND 8)

The requirements for STE included teachers selecting any three (3) of the following STE disciplines:

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

For each discipline submitted, the scorer confirmed the following using the *AltScore* program:

1. Is the student's name, valid date, % of accuracy, and % independence listed on at least 6 STE Summary Sheets?
2. Do at least three STE Summary Sheets have primary evidence attached?
3. Do the three pieces of primary evidence reflect three different science practices?
4. Do the activities on the six STE Summary Sheets reflect the same core idea?

After verifying the above, the scorer used the complexity, accuracy, independence, and self-evaluation values provided on the six STE Summary Sheets to calculate the Level of Complexity (Table 4-2), Demonstration of Skills and Concepts (Table 4-3), Independence (Table 4-4), Self-Evaluation (Table 4-5), and Generalized Performance (Table 4-6).

4.4.3.4 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-blind scoring and resolution (i.e., read-behind) scoring.

4.4.3.5 DOUBLE-BLIND SCORING

In double-blind scoring, two scorers independently score a response, without knowing either the identity of the other scorer or the score that was assigned. Neither scorer knows which response will be (or already has been) scored by another randomly selected scorer. 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) each day was double-scored for each scorer.

Scorers were required to maintain a scoring accuracy rate of at least 80% exact agreement with the TL's score. The TL retrained any scorer whose interrater consistency fell below 80% agreement. The TL reviewed discrepant scores (those that differed by two or more points from the TL's score) with the responsible scorers and determined when they might resume scoring.

Table 4-10 in section 4.7.3 shows the percentages of interrater agreement for the 2019 MCAS-Alt.

4.4.3.6 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 rescoring 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.

4.4.3.7 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.

4.4.4 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 *2019 Educator's Manual for MCAS-Alt*. Section 4.2.1.4 of this report also discusses grade-level and competency portfolios.

4.4.4.1 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 *2019 Educator's Manual for MCAS-Alt*.

Each grade-level portfolio was evaluated in each subject by a panel of content area experts (i.e., veteran educators who also served on the Department's Assessment Development Committees) to determine whether it achieved a score of *Partially Meeting Expectations* or higher. To receive an achievement level at or above *Partially Meeting Expectations*, 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* on the standard MCAS test in that content area.

4.4.4.2 COMPETENCY PORTFOLIOS IN HIGH SCHOOL

Students in high school who required an alternate assessment due to a complex and significant but not necessarily cognitive disability, and 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 (i.e., veteran educators who also served on the Department's Assessment Development Committees) to determine whether it met *Needs Improvement/Partially Meeting Expectations* (or higher) achievement-level requirements. To receive an achievement level of *Needs Improvement/Partially Meeting Expectations* 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
- provided evidence of knowledge and skills at a level comparable with a student who received an achievement level of *Needs Improvement/Partially Meeting Expectations* 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 2019 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-17 through G-23 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-4 through H-10 of Appendix H. Note that the Self-Evaluation and Generalized Performance dimension scores are also included in Appendix H.

4.5.1 Difficulty

Based on 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 consisted 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 0.40 to 0.70 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. It is desirable for an item to be one on which higher-ability students perform better than lower-ability students or one that 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 -1.00 to 1.00. 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. 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	0.78	0.20	0.46	0.06
	4	9	0.78	0.19	0.44	0.07
	5	9	0.78	0.19	0.43	0.07
	6	9	0.79	0.19	0.40	0.08
	7	9	0.79	0.19	0.39	0.08
	8	9	0.79	0.19	0.43	0.08
	10	9	0.79	0.19	0.36	0.08
Mathematics	3	15	0.84	0.19	0.62	0.07
	4	15	0.84	0.19	0.61	0.11
	5	12	0.84	0.19	0.60	0.07
	6	15	0.84	0.19	0.58	0.15
	7	15	0.84	0.19	0.57	0.10
	8	15	0.91	0.13	0.42	0.36
	10	15	0.84	0.18	0.32	0.10
STE	5	12	0.81	0.18	0.40	0.14
	8	12	0.82	0.17	0.44	0.13
Biology	HS	12	0.84	0.19	0.32	0.06
Chemistry	HS	9	0.85	0.19	0.49	0.19
Introductory Physics	HS	9	0.80	0.16	0.61	0.06
Technology/Engineering	HS	9	0.83	0.18	0.31	0.18

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. 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	0.22	0.19	0.22	0.07	0.12	0.02
	4	3	0.19	0.26	0.18	0.05	0.07	0.03
	5	3	0.14	0.24	0.16	0.11	0.06	0.03
	6	3	0.15	0.17	0.18	0.02	0.12	0.04
	7	3	0.12	0.17	0.15	0.04	0.12	0.07
	8	3	0.17	0.21	0.18	0.02	0.12	0.08
	10	3	0.13	0.18	0.10	0.05	0.08	0.11
Mathematics	3	2	0.20	0.18	0.19	0.01	0.04	0.02
	4	2	0.19	0.26	0.12	0.04	0.05	0.05
	5	2	0.18	0.17	0.12	0.02	0.02	0.04
	6	2	0.20	0.04	0.14	0.00	0.04	0.00
	7	2	0.07	0.14	0.13	0.01	0.05	0.01
	8	2	0.15	0.12	0.09	0.08	0.01	0.03
	10	5	0.18	0.08	0.07	0.08	0.09	0.05
STE	5	4	0.23	0.17	-0.01	0.13	0.04	0.08
	8	4	0.22	0.07	0.01	0.07	0.15	0.01
Biology	HS	4	0.11	0.12	0.07	0.06	0.07	0.03
Chemistry	HS	3	0.51	0.13	-0.09	0.50	0.30	0.03
Introductory Physics	HS	3	0.23	0.21	0.04	0.05	0.06	0.13
Technology/ Engineering	HS	3	0.15	-0.07	0.01	0.25	0.01	0.06

** 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), with the exception of one—the correlation in Chemistry. 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 MCAS-Alt 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 DESE 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 DESE'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 MCAS-Alt 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 2019 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 2019 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. 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	939	39	28.97	3.42	0.68	1.94
	4	968	39	29.22	3.29	0.67	1.89
	5	1,056	39	29.12	3.40	0.65	2.01
	6	908	39	28.97	3.70	0.65	2.18
	7	954	39	29.01	3.52	0.64	2.10
	8	898	39	29.05	3.74	0.66	2.18
	10	842	39	28.94	3.53	0.60	2.23
Mathematics	3	892	26	21.26	1.42	0.66	0.83
	4	908	26	21.27	1.44	0.66	0.85
	5	984	26	21.27	1.37	0.63	0.83
	6	862	26	21.35	1.31	0.62	0.81
	7	885	26	21.35	1.29	0.58	0.84
	8	816	26	21.40	1.25	0.57	0.82
	10	826	39	30.83	3.31	0.79	1.50
STE	5	949	39	30.02	3.49	0.80	1.55
	8	809	39	30.53	3.28	0.78	1.55
Biology	HS	626	39	31.22	2.94	0.69	1.65
Chemistry	HS	38	39	32.21	1.83	0.77	0.87
Introductory Physics	HS	59	39	30.56	3.59	0.77	1.71
Technology/ Engineering	HS	85	39	31.22	2.59	0.58	1.68

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 test is a reliable assessment. However, the interpretation of reliability estimate coefficient should take into account the characteristics of the testing sample (such as the variability within the sample) and the test (such as the test length). For MCAS-Alt, due to the special population and the short test length, the range of the α coefficient in the 2019 assessments is reasonable.

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 2019 MCAS-Alt. Tables O-4 through O-10 in Appendix O 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 O 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 percent 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, exact agreement percentage, adjacent agreement percentage, the correlation between the first two sets of scores, and the percentage of responses that required a third score. This information is also provided at the item level in Tables F-4 through F-10 of Appendix F.

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

Content Area	Grade	Number of			Percentage		Correlation	% Third Scores
		Items	Score Categories	Included Scores	Exact	Adjacent		
ELA	3	6	4	802	99.00	1.00	0.99	1.87
		3	5	437	98.86	0.69	0.90	3.89
	4	6	4	1,098	97.63	2.00	0.98	4.19
		3	5	596	98.49	1.51	0.87	4.03
	5	6	4	1,130	98.23	1.68	0.99	2.74
		3	5	639	97.65	1.10	0.68	5.01
	6	6	4	816	98.28	1.35	0.98	2.70
		3	5	480	98.54	0.21	0.72	4.58
	7	6	4	2,346	97.70	2.17	0.98	4.31
		3	5	1,386	98.41	1.15	0.83	4.83
	8	6	4	772	98.70	1.04	0.98	2.07
		3	5	439	97.95	0.68	0.64	3.19
	10	6	4	1,172	97.95	1.88	0.98	4.44
		3	5	679	98.82	0.29	0.73	5.01
Mathematics	3	4	4	526	98.48	1.52	0.95	1.90
		2	5	292	97.95	1.71	0.80	2.05
	4	4	4	694	99.28	0.72	0.98	1.15
		2	5	387	98.97	1.03	0.92	1.03
	5	4	4	724	98.90	1.10	0.97	1.66
		2	5	422	98.58	1.18	0.88	1.42
	6	4	4	566	99.47	0.53	0.99	0.88
		2	5	316	99.68	0.00	0.89	0.63
	7	4	4	1,586	98.93	1.01	0.97	2.40
		2	5	936	98.93	0.43	0.83	1.92
	8	4	4	524	99.05	0.95	0.97	1.15
		2	5	302	98.34	0.33	0.72	1.99
	10	10	4	1,178	98.56	1.44	0.95	2.55
		5	5	686	98.98	0.58	0.83	1.31
STE	5	8	4	996	98.80	1.20	0.98	1.61
		4	5	591	98.48	1.52	0.93	1.52
	8	8	4	728	99.18	0.82	0.99	1.10
		4	5	411	100.00	0.00	1.00	0.00
Biology	HS	6	4	770	98.96	1.04	0.96	1.30
		3	5	450	98.67	0.44	0.71	1.78
Chemistry	HS	6	4	58	100.00	0.00	1.00	0.00
		3	5	33	100.00	0.00		0.00
Introductory Physics	HS	6	4	72	100.00	0.00	1.00	0.00
		3	5	36	100.00	0.00		0.00
Technology/ Engineering	HS	6	4	112	99.11	0.89	0.98	1.79
		3	5	64	100.00	0.00	1.00	0.00

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 according to the same learning standards applied 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. While MCAS has recently transitioned to next-generation achievement levels in grades 3–8, the description of each alternate and grade-level academic achievement level (shown in Table 4-11) 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 P.

Table 4-11. 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. Strand Achievement-Level Look-Up

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

Level of Complexity	*Demonstration of Skills*	*Independence*	*Achievement Level*
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 MCAS-Alt Reporting of Results

4.9.1 Primary Reports

Cognia 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 in each strand) 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 X.

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. The report was redesigned again in 2017 to parallel the layout and format of the next-generation MCAS *Parent/Guardian Report* based on next-generation MCAS tests.

4.9.1.3 ANALYSIS AND REPORTING BUSINESS REQUIREMENTS

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

4.9.2 Quality Assurance

Quality-assurance measures are implemented throughout the entire process of analysis and reporting at Cognia. 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 Reporting Services 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 analysis and reporting business requirements 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 DESE for review and signoff.

4.10 MCAS-Alt Validity

One purpose of the *2019 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 establishment of intended uses and interpretations of test results and conformance to these uses are of paramount importance in relation 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, the sections 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 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. Annual state summaries of the participation and achievement of students on the MCAS-Alt are available at www.doe.mass.edu/mcas/alt/results.html.

In an effort to ensure that all students were provided access to the Massachusetts curriculum frameworks, federal and state laws and DESE policy require that all students in grades 3–8 and 10 are assessed each year on their academic achievement and that all students are included 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.

Aside from legal requirements, 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 learning "daily living skills" is critical for those students to function as independently as possible, academic skills are

extremely important for *all* students and are the primary focus of teaching and learning in the state's public schools. 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 a competency portfolio 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 X 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 knowledge, skills, abilities, and achievement of students with significant disabilities based on 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.

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

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York, 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. (2012). flexMIRT: Flexible multilevel item factor analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group, LLC.

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.

Haisfield, L. & Yao, E. (2018, April). *Industry standards for an emerging technology: Automated scoring*. Presented at the National Council on Measurement in Education, New York, NY.

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.

Jiang, J., Roussos, L., & Yu, L. (2017, April). *An iterative procedure to detect item parameter drift in equating items*. Paper presented at the National Council on Measurement in Education Conference, San Antonio, TX.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC.

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.

Lochbaum, K.E., Flanagan, K., Walker, M., Way, D., & Zurkowski, J. (2016, June). *Continuous Flow Scoring of Prose Constructed Response: A Hybrid of Automated and Human Scoring*. Presented at the National Conference on Student Assessment (NCSA), Philadelphia, PA. Available from https://ccsso.confex.com/ccsso/2016/webprogram/Session4615.html

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.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Applications, 9*(1), 141-142.

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.

Wang, X**.,** Roussos, L. (2018, April). *A Simple Parametric Procedure for Detecting Drift in Anchor Items.* Paper presented at the National Council on Measurement in Education Conference, New York, NY.

Watson, G.S. (1964). Smooth regression analysis. *Sankhy: The Indian Journal of Statistics*, *26*(4), 359-372.

Yao, E., Wood, S., Lottridge, S., Rupp, A., Wendler, C., & Lochbaum, K. (2019, March). *Industry Themes for Implementing Automated Scoring*. Presented at Association of Test Publishers Innovations in Testing, Orlando, FL.

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

APPENDICES